Syntactic Behaviour and Semantic Kinship of Selected Danish Verbs

Anna Braasch
University of Copenhagen

The paper discusses relationships between the syntactic behaviour and meaning of selected verbs, with the focus on exploiting observable syntactic similarities for uncovering of semantic kinship. The investigation is inspired by the demand in language technology for large-scale lexicons that combine morphological, syntactic and semantic descriptions of the lemmas. The development of such a lexical resource is rather demanding, therefore, an enhancement of existing resources with additional information types is a worthwhile task. The computational lexicon for Danish SprogTeknologisk Ordbase (STO) comprises a comprehensive syntactic layer which is assumed to be suitable for enhancement with semantic information. The theoretical background for the current approach is the consensus on obvious relationships between a syntactic behaviour and a particular sense of lemmas, as a surface complementation structure reflects the underlying semantic argument structure. The idea is to test the feasibility of deriving semantic information systematically from the syntactic structures encoded in syntactic patterns.

In the pilot project, a sub-set of trivalent verbs that share syntactic constructions are extracted from STO; the material consists of 216 verbs subcategorising for a direct object and a prepositional object covered by eight syntactic patterns. The examination takes a syntactically based grouping of these verbs as its starting point and focuses on defining lexical classes in terms of shared prevalent meaning components. These components form the basis of the semantic label assignment to the particular groups. The material provides 20 basic semantic groups, such as force, urge, judge, consider, remove, cheat, etc. that can be refined into sub-groups along further semantic features or generalized into classes-e.g. communicate-persuade, cause-change-of , according to different degrees of granularity required. The present classifications of the verbs are also examined in relation to Levin's English verb classes (1993).

Our findings suggest that it is feasible-though within recognized limits-to exploit systematically the formalised syntactic descriptions in meaning group prediction.

1. Introduction

This paper briefly presents some aspects of the relationship between the syntactic behaviour and meaning of selected verb groups and discusses the feasibility of exploiting observable syntactic similarities for uncovering semantic kinship. In this paper, semantic kinship is understood as the relationship that exists between verbs that share similar meanings or central meaning components; syntactic similarity is seen as the shared sense of the verbs that is realized in identical or very similar syntactic constructions.

It is a well-known fact that both the elaboration of a comprehensive dictionary for humans and the development of a lexical resource for computers are highly demanding tasks. During the last decade, various types of lexical resources for computational use have been developed for several languages, many of them focusing either on morphology/syntax or on semantics, but without a link between the resources of a given language. On the contrary, comprehensive dictionaries for humans usually contain morphological, syntactic and semantic description of the lemmas.

The current investigation for Danish is highly inspired by the demand in language technology for large-scale lexical resources that combine morphological, syntactic and semantic descriptions on the one hand, and on the other recognising the fact that semantic networks, like WordNet (Fellbaum 1998), usually do not provide syntactic information at all, or only to a very limited degree. Wordnets contain semantic information in terms of synonym sets (synsets) and

semantic relations. In the wordnet community, there is currently, however, an increasing interest for supplying the semantic description with syntax and morphology; this is the case e.g. for Dutch in the newly initiated CORNETTO project (Vossen et al. 2007). This state of things indicates that there is a need for the development of effective means to combine syntactic and semantic description methods and also existing resources containing these information types.

The paper is organized as follows. Section 2 and 3 present the point of departure and background for this study. Section 4 gives a detailed account of the syntactic description in the lexical resource used for the investigation. Section 5 focuses on the investigation itself, discussing the approach to the semantics-syntax relation and the method of material selection to be examined and presenting selected findings. Finally, section 6 concludes on the problems encountered in the material and the advantages and drawbacks of the method employed.

2. Pilot investigation studies for Danish

The recent development of DanNet, a Danish wordnet comprising approx. 40,000 synsets, and its upcoming extension up to 100,000 synsets within the framework of the national CLARIN project (*http://www.cst.dk*) will constitute a comprehensive computational resource for Danish semantics. In order to enhance its future utility in practical language technology applications, such as systems for advanced information retrieval, disambiguation systems, an enrichment of DanNet with morphological and syntactic information is foreseen. To this end, two pilot investigations are carried out; one examining the compatibility between *SprogTeknologisk Ordbase* (henceforth STO), a large computational lexicon for Danish, containing morphological and syntactic information and DanNet, aiming at an estimate of the feasibility of merging these two resources. This investigation was mainly concerned with nouns (Pedersen et al. 2008).

This paper reports on the other pilot project with a focus on selected verb groups sharing syntactic properties. The idea is to derive underlying semantic information systematically from observed syntactic surface structures; the method is highly inspired by Levin's classification of English verbs (Levin 1993). The main question addressed here concerns the evaluation of the feasibility of exploiting syntactic patterns and specific prepositions of verbs to deduce shallow semantic information associated with their meaning.

3. Background

The Danish Lexicon for Language Technology Applications (*S*prog*T*eknologisk *O*rdbase, henceforth STO) is the largest computational lexicon for Danish currently available (*http://cst.dk/sto/index.html*). It is based on the PAROLE lexicon model (Navarretta 1997 and Braasch 2002) and contains detailed and formalised morphological and syntactic descriptions of more than 81,000 and 45,000 lemmas, respectively (Braasch and Olsen 2004). The STO resource has been used since 2003 in various applications, although without a semantic description level. In the SIMPLE project, however, a small semantic lexicon is developed for experimental purposes, and a small part of the STO lemmas is interlinked with very rich semantic descriptions (cf. Pedersen and Paggio 2004). The description method is based on the qualia structures in Pustejovsky's very complex model for the Generative Lexicon (Pustejovsky 1995).

In the Danish language technology user community there is a growing demand for supplying the STO lexicon with more simple semantic information to be used in practical applications, which is the direct cause of the investigation presented. This paper specifically addresses a possible exploitation of shared syntactic descriptions of verbs for uncovering their common semantic features. A reasonable semantic grouping of these verbs is seen as useful means that can support the sense description and encoding of semantic information.

4. The nature of the material examined

4.1. *Syntactic description in STO—main principles*

The object of this investigation consists of selected groups of Danish verbs encoded in STO showing obvious similarities in their syntactic constructions. The organizing principle is based on the observable syntactic features of a lemma in its immediate context(s). The description itself applies a complex valence-based system comprising both the syntactic category and the function of the complements subcategorized for, and it comprises also the preposition that introduces a prepositional complement. If a complement, such as a prepositional object can be realised alternatively by different syntactic categories e.g. by a noun phrase or that-clause, then the description covers both constructions, provided that the same preposition introduces the syntactic alternatives in question.

In this approach, governed prepositions e.g. *for* ("for") versus *til* ("to") cf. examples in (1) differentiate syntactic descriptions in the same way as any other encoded feature such as it is the case for *anse* + *for* ("consider"+ "as/ to be") and *anslå* + *til* ("estimate" + "at").

(1a) Mange *anser* ham *for* at være en gangster ("Many consider him to be a gangster").

(1b) Kommunen *anslår* tabet *til* at være på syv milliarder (Litt.: "The municipality estimates the loss to be at seven billions").

The semantics of prepositions is very complex, especially when used to introduce free adjuncts (also called modifiers), e.g. in case of temporal or modal adverbials or describing instrument or material used. The Danish Dictionary (DDO, 2003-2005) lists 20 main senses of the preposition *for* ("for") and 24 main senses of *til* ("to"), this indicates that a disambiguation of these highly polysemantic prepositions is also needed when they appear in a valence bound structure.

The selection of governed prepositions is semantically defined—or restricted—by that argument which is realized by the complement in question. This is one of the surface realisation features that guide the present study. Recent trends in research into syntax-semantics relationship also investigate the observable role of prepositions for a proper identification of semantic arguments. Kipper et al. (2004) concludes that "The significance of prepositions and their relation with verbs is of the outmost importance for a robust verb lexicon, not only as a syntactic restrictor, but also as a predictor of semantic content."

Two additional basic features are relevant for a syntactic description of a construction. First, whether an element in a given context of a word is governed by that word or not (viz. to be regarded as a valence bound complement of a construction or not); second, whether a complement is mandatory or not (viz. optional) in a construction. In these regards, the general valence theory is in STO adapted to the needs of language technology applications as discussed in Braasch (2006). The most relevant difference is the inclusion of frequently occurring, prototypical but weakly bound elements, the so-called "middles", being a category between complements and adjuncts, as long they are central to the particular meaning of the word (for a detailed discussion see Somers 1987: 27 ff).

4.2. *Syntactic description of verbs*

For verbs, this means that their syntactic description comprises a *combination* of relevant complementation properties and other features specific to the word class, such as reflexivity, control and raising; all are formalized in *attribute/value-pairs*. An attribute is the name of a given linguistic property, e.g. *syntactic function*, and a given value describes this property of the complement in question, e.g. *object.* A unique combination of relevant attribute/value-pairs makes up a *syntactic pattern* describing a particular syntactic behaviour. This means, that each lemma has at least one syntactic pattern e.g. *bortforklare* ("explain away"), but it may also have several patterns e.g. *tage* ("take"). The *syntactic unit* is a unique combination of a lemma and one of its syntactic patterns, reflecting a particular syntactic behaviour (and most often also a particular sense) of the lemma. According to this principle, the verb *bortforklare* has a single syntactic unit—and a single sense only, whereas *tage* has several syntactic units and it is also a highly polysemantic verb.

It is worth noting that the syntactic description in STO also comprises adjectival, adverbial and sentential complements and particles as well, besides noun phrase and prepositional phrase complements.

The descriptions of lemmas are hand-crafted and based on corpus evidence; though STO is rather comprehensive-size its coverage is not fully exhaustive, mainly because of the particular type (first of all newspaper texts) and delimited size (20 M tokens) of the primary corpus. Despite this fact, it is assumed that material with the above features encoded may very well provide the springboard for the identification of *common semantic denominators* or sense labels. A further fact supports this assumption, viz. the circumstance that some basic observations concerning sense disambiguation have been taken into consideration in the process of distinguishing syntactic units in such a way that different surface realisations which structurally in principle could be kept together, are split into separate syntactic descriptions if they indicate separate senses (e.g. an object realised by a noun phrase and with a that-clause, resp.).

4.3. Encoded verbs—an overview

The total number of verb lemmas provided with a syntactic description in STO is approx. 5,660 giving 8,558 verb readings (syntactic units), thus one verb is in average described by ~1.55 syntactic unit. This indicates, that a large number of the verbs covered in the resource is provided with a single syntactic description only, because verbs with multiple particles (*på* "on, onto,...", *af* "of, from...", *til* ("to, onto,...") and/or complex semantics, e.g. *tage* ("take"), *læse* ("read, study") have obviously several syntactic units.

Table 1 provides an overview of these 8,558 verb readings with regard to the syntactic constructions in which they can appear; that is their distribution onto *valence types* (zero- to tetravalent).

VALENCE TYPE (arity)	SYNTACTIC UNITS (verb readings)	SYNTACTIC PATTERNS (construction types)	SYNT. UNITS per SYNT.PATTERN (average distribution)
Dv0 (zerovalent)	47	4	~11.8
Dv1 (monovalent)	1,164	95	~12.2
Dv2 (divalent)	5,722	479	~12.5
Dv3 (trivalent)	1,573	304	~5.17
Dv4 (tetravalent)	52	6	~8.66
Total	8,558	888	Total average ~9.65

Table 1. Distribution of syntactic descriptions of approx. 8,558 verb readings

The following details for each valence type are given in columns 2, 3 and 4. In column 2, for each valence type the total number of syntactic patterns is stated. Column 3 shows the total number of syntactic units (verb readings) for the valence type in question. The 4th column is a kind of summarisation of the distribution: it provides the average number of syntactic units for each syntactic pattern of the valence type in question. These figures are used as a level of reference when defining the "population" of a syntactic pattern in terms of being large, small or (close to) average. The zerovalent type is the least comprehensive one represented by 47 verb readings, though the average distribution of verb readings (syntactic units) onto patterns is not significantly lower than the highest average distribution figure, viz. ~ 12.5 for divalent verbs (see Table 1).

Not surprisingly, verb constructions with a subject and an object governed are by far the most frequent ones in Danish. They give rise to the largest number of syntactic patterns, the figure of average syntactic unit/pattern distribution (~12.5), which is significantly higher than the average for all verbs covered in STO (~9.65); on the other hand it is only slightly higher than the figures for the two less complex (mono- and zerovalent) construction types.

In the present examination, the population/distribution has a central role in the selection of the material, and some relatively complex, trivalent syntactic patterns are taken as a starting point.

As shown in Table 1, the *average* distribution of syntactic units onto trivalent patterns is rather low, viz. ~5.17 units per pattern, but the *individual* figures for verbs sharing the selected patterns are significantly higher, ranging from 7 to 56. This means, that many verb readings share the syntactic behaviour described by the pattern in question, therefore this material lends itself very well to examining with respect to shared lexical semantic properties.

Similar selectional conditions were applied in an earlier examination of syntax-semantics relationship (Braasch 2006), though with opposite intentions: the simplest possible (viz. the zerovalent) syntactic patterns and all their instantiations in STO comprising mainly weather and sound emission verbs, and the most complex, (tetravalent) patterns were selected describing transport verbs and verbs related to transfer/transmission.

5. Approach and related work

The theoretical background for the current approach is the consensus on the obvious relationship between syntactic behaviour and a particular sense of each lemma saying that surface complementation structure reflects the underlying structure of semantic arguments. Levin (1993: 1) states that "[…] verb behaviour can be used effectively to probe for linguistically pertinent aspects of meaning". Further, "[…] verbs that fall into classes according to shared behaviour would be expected to show shared meaning components" (op. cit. p. 5).

Levin pinpoints the ability of native speakers concerning the proper production and understanding of various expressions containing a given verb and possible combinations of arguments and adjuncts. This conception forms the basis of her theoretical perspective on English verb alternations. Though syntactic alternations in Danish differ at the more detailed levels in several respects from their English counterpart, Levin's work on English verb classes and alternations became a kind of frame of reference in research into the syntax and lexical semantics relations of verbs. At this point it should be mentioned, that the two languages have also a number of general alternation types in common, e.g. the ergative, reciprocal and dative alternations. However, these will not be discussed further in this paper. The basic criteria for identification of verb classes in Levin (1993: 19) are referred in the following way: "What is important is the existence of core sets of verbs with specific sets of properties that can provide the basis for the later identification of meaning components." The point of departure chosen here falls also in line with these important observations, which means that formalized syntactic descriptions of STO presumably provide a suitable basis for an identification of semantically defined kinships.

Although the approach presented here starts from the "wrong side" compared to some comprehensive analyses of the semantics-syntax relationship of English verbs (e.g. Tang Dang et al. 1998 and Korhonen et al. 2003), it might also be employed in acquisition and prediction of verb senses. In fact, Dorr and Jones (1995) suggest similar acquisition techniques for word senses. Interestingly, various research articles related mainly to VerbNet and PropBank discuss approaches to semantics starting from the syntactic angle, such as Kingsbury and Kipper (2003) and Kipper et al. (2006), looking for syntactic regularities as a basis for clustering of verbs with similar meaning and usage and the conclude: "Clustering of syntactic patterns can quickly and automatically provide a first approximation of the groupings into which meaning-bearing items fall." (Kingsbury and Kipper 2003: 76). The work mentioned differs from the study presented here in several respects, the most important one being the very basis of this research. The two resources for English, i.e. the VerbNet, being a broad-coverage lexical resource with explicit syntax and semantics and the PropBank with predicate-argument structures encoded, provide a much more elaborate and comprehensive material than what is extracted from STO for the purposes of this study. In spite of this fact, the aim and approach described here have much in common with the subjects of projects related to English verbs, and therefore their results have an encouraging influence on the present investigations too.

For Danish, investigations with a related object have been carried out formerly, though from a different point of view e.g. in the Odense Valency Dictionary project in the nineties (see e.g. Schøsler and Van Durme 1996). This project employed a constructivist method, the so-called

pronominal approach to examining the elements of a construction which were selected by the valence kernel. A further example of relevant research into the semantics of motion verbs is described in Pedersen (1997).

5.1. *Selection method*

With the aim to define a sub-set of verbs suitable for a detailed examination concentrating on possible semantic similarities, a selection procedure of three main steps was designed.

First, for each of the five (zero- to tetravalent) valence types, all syntactic patterns are extracted from the STO database, the overview of numeric figures is shown in Table 1. In the second step, after a closer inspection, an interesting and representative subset of *syntactic patterns* (construction types) is selected: eight subtypes of trivalent constructions comprising also some ditransitive types (also called "double object constructions" in Levin 1993: 274) which show a *considerable degree of formal similarity*. Third, all *syntactic units* (i.e. verb readings) sharing one of the selected syntactic patterns are extracted form the STO lexicon, and lists of verbs with corpus examples attached are generated appropriately for investigation of syntax-semantics relationship.

After these steps the material to be examined consists of a subset of trivalent verbs comprising 216 syntactic units (viz. ~14.5 % of all trivalent verbs). They are described with 8 (of the 314 trivalent) syntactic patterns, which gives an average distribution of ~27.0 units per chosen syntactic pattern. This figure is substantially higher than the total average distribution value (viz. ~9.65 cf. Table 1), and more than five times as high as the average for trivalent verbs (which is ~5.17). In the first run this is taken as evidence of the fact that each of the selected patterns describes a comparatively large number of verbs that may share meaning-related properties as well. For this reason, the selected material seemed appropriate to illustrate the question of exploitation feasibility.

The examination started from the unfolding and systematizing of attribute/value pairs that differentiate the syntactic patterns in question. In this respect, the relevant attributes describe the prepositional object of the construction. For each attribute, some disjunctive values are possible as shown in Table 2. Further, each unique combination of these attribute/value-pairs is unfold, listed and instantiated by a number of verbs (see Table 3).

Attribute name	Values
Syntactic function	POBJ
Governed preposition (introducer)	*for* ("for") or *til* ("to")
Syntactic category	NP or NP and *at*- ("that"-)infinitive or infinitive clause
Control (for that-infinitive and that-clause)	no control or object control
Syntactic realisation is mandatory	yes or no

Table 2. Differentiating features of the selected syntactic patterns

Table 3 explicates the main properties described by the eight trivalent patterns selected, i.e. the governed preposition and the syntactic category of the complement. (The information on whether the prepositional object is mandatory or optional in the construction is not fully consistently encoded in the material studied therefore it is given a secondary importance at this point.) For each pattern, a prototypical example and the number of syntactic units are provided as well.

FEATURES of the POBJ in the syntactic pattern			Examples*	Syntact. Units
Prep.	Complement type	Optional		
til	NP	yes	genopstille ngn (til posten) "renominate sby (for a post)"	56
	NP	no	udråbe ngn til vinder "proclaim sby a winner"	31
	NP/ that-clause with obj. control	yes	besnakke ngn (til {et lån/at låne…}) "talk sby round" "talk sby into a loan/ into lending"	20
	NP/ that-clause with obj. control	no	beordre ngn til {mobilisering/at mobilisere} "command sby {to mobilization/mobilize}"	30
for	NP	yes	snyde ngn (for penge) "cheat sby (out of some money)"	35
	NP	no	tiltale ngn for drab "indict sby for murder"	21
	NP/ that-clause with obj. control	yes	kritisere ngn (for {sjuskethed/for at være sjusket}) "criticize sby (for {sloppiness/for being sloppy})"	16
	NP/ that-clause with obj. control	no	anse ngn for {en god læge/at være en god læge} "consider sby {a good doctor/to be a good doctor}"	7
TOTAL				216

Table 3. Overview of the features of the prepositional object in the selected syntactic patterns

*Note on the use of brackets in the examples: not mandatory, viz. optional complements are bracketed with ( ); alternative syntactic realizations of a complement are separated by a slash and put in { }.

5.2. *Semantic investigation—basic steps*

The investigation of the semantic properties and recording of semantic kinship of the verbs focuses on three central points: first, identification of the *prevalent meaning component* of each verb under consideration of the corpus example illustrating the construction in question; second, extraction of *meaning components shared* by a group of verbs; third, *generalizations to be captured* with respect to the observable relationship between subcategorized complements (syntactic surface realisation) and semantic arguments of verbs. Finally, the findings are summarized in overview tables. In a further step, these tables can form the basis for development of encoding templates or frames that capture systematically the properties of the semantic groups or classes.

Each verb in the eight lists is provided with a *provisory semantic label* that was chosen intuitively, identifying the prevalent meaning component(s) of the verb. In the first run, terms like "genus proximum" or "immediate superordinate" are avoided deliberately, because these terms are used in hierarchical organisation of concepts and therefore too specific for the present purpose of coarse-grained analyses. Instead, the labels assigned here are key (or more general) verbs or verbal expressions in English that comprehend the shared meaning components of verbs; their choice is motivated by linguistic intuitions. This assignment method is in accordance with commonly accepted views, as e.g. stated in (Fellbaum 1998: 70): "Dividing the verb lexicon into semantic domains initially on a purely intuitive basis might lead one to discover relations that organise verbs and concepts." The initially assigned labels have been refined stepwise along a few components, e.g. the JUDGE/CONSIDER semantic label was assigned in the first run both to verbs like *dømme* ("judge, convict") but also to verbs like *kritisere* ("criticize"), *rose* ("prize"), *anse* ("consider"), although also COMMUNICATE is a prevalent meaning component of the three last mentioned verbs equally to JUDGE/CONSIDER. Various additional components belong closely to the meaning of the verbs, too, such as the reason for the judgement, the judgement type and the attitude of communication, the last mentioned can be negative, positive or neutral. Therefore, in the second run the label is

converted into COMMUNICATE_JUDGEMENT. All further labels were in the same way systematized in order to eliminate differences in granularity, overlaps and assignment errors, etc. The revised labels function thereafter as a kind of *common semantic denominators* for verbs that share meaning, viz. they have a prevalent meaning component in common. These denominators suggest groups with a certain semantic kinship; the material provided in the first run in total 20 tentative groups. Some of these groups can then be put together in the generalisation process on the basis of the extent of shared prevalent semantic properties, as will be outlined below for the COMMUNICATE_PERSUADE class.

5.3. *Shared properties and semantic grouping*

The selectional restriction on the first complement (viz. the subject) of all investigated verbs, is *human*, only a human being is able to perform intentional actions such as *persuade*, *estimate*, *assign*, etc. Therefore, this shared property is included by default in the generalisation.

Table 4 provides a list of illustrative examples of semantic labels that are shared by the verb group members, all of which share syntactic patterns with the prepositional object introduced by *til* ("to"); the governed prepositional objects are provided with examples. Remembering that the preposition *til* is highly polysemantic, it is obvious that the grouping of verbs is also influenced by the particular sense with which the preposition contributes to the semantics of the construction. The basic meaning component of *til* is "towards" expressing senses like in the direction of, leading to, in relation to, with a view to, etc. which correspond to semantic roles like GOAL, DESTINATION, RECIPIENT, etc.

Only a small number of the verbs below (translated into English) are classified by Levin, such as e.g. *elect* (29.1 Appoint verbs), *estimate* and *value* (54.4. Price verbs), *report* (37.4 Verbs of instrument of Communication). A few other verbs are represented in Levin's classification, though in a different sense, e.g. *ask.* There is an obvious reason for this: the majority of the verbs in the Table 4 do not have alternative syntactic constructions in English, and the basis of Levin's classification is the system of alternations. Further, complements expressed by *til* ("to") + a noun phrase or *til* ("to") + *at* ("that") infinitive/sentential complement realizing arguments like goal/purpose, measure, etc. are by Levin treated and as an exception only, being regarded as oblique complements, these types correspond often to above mentioned middles in terms of Somers (1997: 27).

SEMANTIC GROUP LABEL	VERB examples	OBJ. NP SEM. ROLE examples	POBJ examples	POBJ : Selection features	POBJ. SEM. ROLE
FORCE	*tvinge* "compel" beordre "order"	PATIENT (human)	denne handel "deal" *at lukke dørene* "to close the doors"	human activity	TOPIC
URGE/ REQUEST	overtale "persuade" *formane* "admonish"	PATIENT (human)	*en rejse* "a journey" at melde sig "to volunteer"	human activity	TOPIC
ASK/ADVISE INSPIRE	*råde* "advise" *motivere* "motivate"	PATIENT (human)	*modstand* "resistance" at læse "to read"	human activity	TOPIC
ESTIMATE	værdiansætte "value" *anslå* "estimate"	PATIENT (concrete)	6 mill. Euro "6 m Euros" at vare 3 måneder "to last 3 months"	scalar	MEASURE
LIMIT	*normere* "set_a_norm" *begrænse* "restrict"	PATIENT (inanimate)	*8 timer* "8 hours" at arbejde 8 timer "to work 8 hours"	scalar	MEASURE
ELECT	*udtage* "choose" *udpege* "select"	PATIENT (human, organisation)	*kampen* "the match" *at spille mod Polen* "to play against Poland"	human activity	GOAL

QUALIFY	uddanne "train" *opdrage* "bring_up"	PATIENT (human)	et job "a job" at hjælpe andre 'to help others"	human activity	GOAL
RELATE	henregne "reckon" tilordne "assign"	PATIENT/ THEME (entity)	en ny kategori "a new class" bestemte gener "particular genes"	abstract; class/type	GOAL [target]
ADAPT	*tilpasse* "adjust" *akklimatisere* "acclimatize"	PATIENT/ THEME (entity)	*publikum* "the audience" *nye vejrforhold* "new climate conditions"	state	GOAL [target]
GIVE_ KNOWLEDGE	*angive* "report" *indtelefonere* "report_by_phone"	THEME (semiotic)	*politiet* "the police" *avisen* "the newspaper"	human/ organisation	RECIPIENT
FIX	fastgøre "fasten" *tøjre* "hitch"	PATIENT (concrete)	*et træ* "a tree"	concrete	LOCATION

Table 4. Trivalent syntactic patterns of verbs with the preposition *til* (selection)

The overview in Table 4 indicates similarities wrt. semantic properties of verbs not only between the members of one single group but also at a more general level, between some of the established groups too where the selection features of the prepositional object are identical, e.g. between the ESTMATE / LIMIT, and ELECT / QUALIFY groups respectively. Broken separating lines between two rows of the table indicate such a semantic kinship of these groups. Moreover, for the first three groups, the generalisations captured could be formulated as follows: a formulation capturing these are verbs describing *communication* of subject's *intention* directed to a *human object* with the goal to *persuade* the object to act in a certain way (the central semantic components are in italics.) Accordingly, these three groups can be accumulated in a *class* under the generalised semantic label COMMUNICATE_PERSUADE.

The last two groups in Table 4 (rows separated by double lines) are slightly different from the rest of the groups in that the prepositional object is a realisation of benefactive (GIVE_KNOWLEDGE verbs) or directional (FIX verbs) semantic roles; though the GIVE_KNOWLEDGE group still shares the core meaning component of communication (but not the persuasive one) with the first three groups. This illustrates that this classification type accepts certain intersections, which means that the groups, as being defined here, are not totally mutually exclusive.

5.4. *Process of refinement and classification*

The preliminary grouping can in the above outlined manner be refined into classes and subclasses along various semantic features and according to degrees of granularity required. As shown in Table 4, the semantic features are registered in terms of thematic roles of arguments with focus on the first (direct) and the second (prepositional) objects; the syntactic realisation of the prepositional object is provided as well. Also selectional restrictions on arguments (e.g. human, concrete, abstract, event, scalar, quantity…) are taken into consideration.

In order to establish/define appropriate lexical classes, the generalizations to be captured are singled out. This process is performed partly by semi-automatic sorting of verb readings into groups sharing the semantic features assigned, followed by human adjustments. The outcome of the overall process is captured in classification sheets comprising the set of members, their syntactic complementation and semantic argument properties. In this way, the verbs originally selected on the basis of their syntactic patterns are classified in a semantically motivated manner. In the final step, the verb readings and their present classification are compared to the English verb classification system provided in (Levin 1993). The Danish verbs are translated into English in consideration of the difficulty of giving preference to one possible translation over other(s); the choice depends on the focus on a particular semantic component. Semantic generalizations captured are developed in summarizing tables that describe the classes identified

in the examination. Refinement of the analysis for an illustrative sample of an extract of trivalent verbs is shown in Table 5; it captures double object constructions with focus on the prepositional object consisting of preposition *for* ("for") + a noun phrase and its semantic properties. The description of the direct object is considerably simplified here because of space limitations and serves as an indication only. The table also contains references to the relevant Levin verb classes, where possible.

In such a refinement process, obviously common semantic features are captured as generalisations, in the present case for a set of verb groups which describe a certain kind of intentionally caused change of a particular *feature* of the direct object, such as location, possession or state. The groups labelled REMOVE, CHEAT, CLEAR, SELL, OFFER and DEBET in Table 5 are comprised under the common label CAUSE_CHANGE_of_<*FEATURE*> in a generalised semantic class.

SEMANTIC CLASS: CHANGE_of	VERB examples	OBJ: NP SEM. ROLE examples	POBJ: for + NP examples	POBJ : Selection features	POBJ: SEM. ROLE	LEVIN'S CLASS
LOCATION (REMOVE)	*tømme* "empty", *rydde* "remove" tappe "tap"	LOCATION *bus* "coach" *tønde* "barrel"	pakker "parcels", passagerer "passengers", *vand* "water"	concrete	THEME (located)	10: Verbs of Removing
POSSESSION (CLEAR/ REMOVE)	*rippe* "strip" *lænse_2* "drain"	PATIENT (possessor) *familien* "family" *firmaet* "company"	*arven* "estate" aktiebeholdningen "holding"	valuables	THEME (possessed)	10.1+3 Remove, Clear
POSSESSION (CHEAT/ FOOL)	afpresse "extort" *narre* "defraud" *snyde* "swindle"	PATIENT (possessor) chefen "boss" *firmaet* "company"	*penge*"money" *stort beløb* "large sum"; *gave* "gift"	value (grant, payment)	THEME (possessed)	10.5+6 Steal, Cheat Poss. Deprivation
POSSESSION (OFFER)	*skænke* "pour out"; *ofre* "devote"	PATIENT *en øl* "beer"; *karriere* "career"	*gæsterne* "guests" *familien* "family"; *sagen* "cause, business"	animate; abstract	BENE-FICIARY	13.3 Future having (Ch_Possession)
POSSESSION (SELL)	sælge "sell" videresælge "resell"	PATIENT *huset* "house"	500 kroner "500 DKK" en formue "a fortune"	payment	THEME	13.1 Give (Ch_Possession)
POSSESSION (DEBET/ CHARGE)	*debitere* "debit" *bone* "bill"	PATIENT *kunde* "customer"	*reparation* "repair"; 400 kroner "400 crowns"	goods/ services; payment	THEME	54.5 Bill (Measure)
STATE (CURE)	*kurere* "cure" *behandle* "treat"	PATIENT *barnet* "child"	mæslinger "chicken pox"	disease/ disorder	THEME	10.6 (Steal)

Table 5. Overview of CAUSE_CHANGE verbs with location, possession or state <*FEATURE*>s

6. Findings and perspectives

The primary interpretation of the findings suggests that it is possible to establish groups of verbs based on their semantic kinship and predict senses on the basis of the formalised description of their syntactic behaviour. The insight emerging from this investigation is useful for the preparation of further work as it indicates a potential for semantic classification; more

specifically, semantic classes can, to a certain point, be induced from the syntactic descriptions of STO. A resource developed by combining the two information types is of practical use e.g. in shallow semantic annotation tasks.

The semi-automatic procedures employed in selection and sorting, etc., proved their usefulness in processing and organising the material extracted from STO, although within certain limitations some of which being of computational type while others of a linguistic nature as exemplified below.

The vast majority of the selected verbs with a shared syntactic pattern are classifiable on the basis of their semantic kinship, even if some "problem verbs" with syntactic similarity and semantic "otherness" are left over. Such single verb readings, e.g. *hjælpe* "help" in a construction like "help him to *get a job*" could not be classified straightforwardly in any of the established main classes; they are therefore stored in a separate list for further, individual treatment. Another type for further manual treatment is made up by verb readings with one syntactic realisation of two different argument structures, e.g. *sælge* "sell", (a) sell something *for an amount* (b) sell something *for somebody*, where the prepositional may realize two different semantic roles, viz. in (a) THEME and in (b) CAUSER (that is the role of a referent which instigates the event rather than doing it). The distinction between these two semantic varieties is not reflected in Table 5, but it is a case that shall be accounted for together with other, similar verb behaviours.

In some other cases, it is difficult to distinguish adverbial complements from formally identical prepositional objects without human inspection, e.g. in case of the verbs of fixing such as *fastgøre* "fasten", *tøjre* "hitch" with a governed directional preposition *til* ("to") + noun phrase (these verbs belong to class 22.4 Tape verbs in Levin's classification). It is noting worth, that the verb describes the end state of CAUSED_CHANGE _of_LOCATION of the direct object and not the way this end state is reached (cf. Levin 1993: 161 ff.), though the preposition governed in Danish is clearly directional. As shown in Table 3, these verbs have particular semantic properties that differentiate them from the other verb groups, thus the semi-automatic sorting has to be followed up by manual treatment.

Finally, it is important to recognize the limits of the method applied. The approach outlined in Section 5 is probably less well-suited for transitive verb (simple divalent constructions). In this case the output from processes of sorting syntactic constructions is less informative as regards possible semantic grouping of verbs because of the degree of structural overlaps between syntactic constructions of verbs with different semantic features. Divalent constructions have a less specific syntactic structure than verbs in the material investigated here, because they comprise only one complement, a direct object, besides the subject (i.e. there is no prepositional complement). As mentioned in Section 4.1., prepositions have a significant importance for the semantics of a verb, and they are also clearly identified in STO when they occur in a syntactic construction of a verb. The absence of this semantically significant category in divalent constructions weakens the prediction power of the semi-automatically generated groups therefore the process must be supervised closely.

During the analysis process, some encoding deficiencies were detected in the STO material, e.g. the syntactic constructions of a verb are not encoded exhaustively especially in case of verbs having several particle constructions, only a few alternation types are covered, etc. As a benefit of the process, shortcomings that became visible during the investigation can be corrected systematically in a follow-up process.

7. Summing up

The outcome of the examination seems encouraging: the existing syntactic resource (i.e. the STO lexicon) can be enriched with semantic information being systematically derived or induced from the syntactic descriptions itself. Although several sub-processes can be performed semi-automatically, there will still be a need for substantial manual lexicographic work. An

analysis of the general predictive power of a broader selection of syntactic patterns has to follow, where both strengths and weak points of the method will be considered systematically.

Acknowledgements

Thanks to Bolette Sandford Pedersen for providing valuable comments in our discussions of the topic and to Geoffrey Williams for providing helpful corrections to language and style.

References

Braasch, A. (2002). "Current developments of STO—the Danish Lexicon Project for NLP and HLT Applications". In *Third International Conference on Language Resources and Evaluation, Proceedings.* Las Palmas de Gran Canaria. Vol. III. 986-993.

Braasch, A.; Olsen, S. (2004). "STO: A Danish Lexicon Resource—Ready for Applications". In *Fourth International Conference on Language Resources and Evaluation, Proceedings*. Lisboa. Vol. IV. 1079-1083.

Braasch, A. (2006). "Exploitation of Syntactic Patterns for Sense Group Identification". In *Proceedings, XII EURALEX International Congress.* Alessandria. 133-140.

Dorr, B. J.; Jones, D. (1995). "Automatic extraction of Semantic Classes from Syntactic Information in Online Resources". *Technical Report CS-TR-3481*. Maryland: University of Maryland. URL: *http://lit.csci.unt.edu/~wordnet/.*

Fellbaum, C. (1998). "A Semantic Network of English Verbs". In Fellbaum, C. (ed.) *WordNet. An Electronic Lexical Database*. Cambridge, London: The MIT Press.

Hjorth, E.; Kristensen, K. (eds.) (2003-2005). *Den Danske Ordbog* [DDO]. Det Danske Sprog- og Litteraturselskab. Gyldendal. København.

Kingsbury, P.; Kipper, K. (2003). "Deriving Verb-Meaning Clusters from Syntactic Structure". In *Proceedings of the HLT-NAACL 2003 workshop on Text meaning.* 70 - 77.

Kipper, K.; Snyder, B; Palmer, M. (2004). "Using prepositions to extend a verb lexicon". *Computational Lexical Semantics, Workshop in conjunction with HLT/NAACL 2004*. Boston.

Kipper, K. et al. (2006). "Extending VerbNet with Novel Verb Classes". In Proceedings of 5th international conference on Language Resources and Evaluation. Genova, Italy. URL: *http://www.cl.cam.ac.uk/~alk23/lrec06.pdf*

Korhonen, A.; Krymolowski, Y.; Marx, Z. (2003). "Clustering Polysemic Subcategorization Frame Distributions Semantically". In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. Sapporo, Japan. 64-71.

Levin, B. (1993). *English Verb Classes and Alternations. A preliminary Investigation*. Chicago, London: The University of Chicago Press.

Navarretta, C. (1997) "Encoding Danish Verbs in the PAROLE model". In *Proceedings from RANLP ´97*. Tzigov Chark, Bulgaria. 359-363.

Palmer, M. (2006). A Class-Based Verb Lexicon. [on-line]. *http://verbs.colorado.edu/palmer/projects/verbnet.html*

Pedersen, B. (1997). "Danish Motion Verbs: Syntactic Alternations and the Hypothesis of Semantic Determination". In Nikkanne, U. (ed.) *Nordic Journal of Linguistics* 18. 63-89.

Pedersen, B. et al. (2008, in press). "Merging a Syntactic Resource with a WordNet—a feasibility study of merge between STO and DanNet". In *Sixth International Conference on Language Resources and Evaluation, Proceedings,* Marrakech.

Pedersen B.; Paggio P. (2004). "The Danish SIMPLE Lexicon and its Application in Content-based Querying". In *Nordic Journal of Linguistics* 27 (1). 97-127.

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

Schøsler, L.; Van Durme, K. (1996). *The Odense Valency Dictionary: An introduction*. Odense: Odense University.

Somers, H. (1987). *Valency and case in computational linguistics*. Edinburgh: Edinburgh University Press.

Trang Dang, H. et al. (1998). "Investigating regular sense extensions based on intersective Levin classes". In *Coling/ACL-98, 36th Association of Computational Linguistics Conference*, *Proceedings*. Montreal. 293-300.

Vossen, P. et al. (2007). "The Cornetto Database: Architecture and User-Scenarios". In *DIR 2007*. Leuven, Belgium. 89-96.