Exploitation of syntactic patterns for sense group identification

Anna Braasch
Center for Sprogteknologi
University of Copenhagen
Njalsgade 80
DK-2300 Copenhagen S
e-mail: anna@cst.dk

Abstract
The complementation structure of a word reflects its semantic arguments and indicates a particular sense of that word. Evidently, consistent and detailed syntactic descriptions of words provides a firm basis for their semantic analysis.
The large-scale Danish lexical database, STO, is worked out for computational use in natural language processing. It contains very rich formalised information on the syntactic properties of 45,000 lemmas, whereof only a subset is provided with sematic information. Encoding of semantic information is rather time-consuming, therefore it is worth investigating whether the encoded syntactic descriptions can serve as a basis for a kind of shallow semantics. This paper describes an approach to exploiting syntactic information in STO for identification of sense groups. Systematic semantic relationships are captured within sets of lemmas which show a similar syntactic behaviour, viz. verbs governing zero and four complements, respectively.
Our ultimate objective is to evaluate the feasibility of a prospective semi-automatic coding of senses.

1 Introduction

The large-scale Danish computational lexicon STO has been ready for both commercial and research purposes since early 2005, and it is currently in use in various natural language processing (NLP) projects.[1] The lexicon proved also to be a useful source of linguistic knowledge to linguists and teachers/learners of Danish because of its consistent and very detailed morphological and syntactic description of words. The development project and the resource itself have been presented at previous EURALEX congresses from various lexicographic and computational points of views (Braasch & Olsen 2000; Braasch & Sandford Pedersen 2002). This paper discusses an exploitation scope of the syntactic data from the syntax-semantics association's point of view.

[1] Further information on the homepages http://cst/dk/sto and http://www.elda.org

2 Method of syntactic description in STO

2.1 The "lexicalist approach" to the grammar

The total number of entries exceeds 81,300 of which 45,000 lemmas are provided with syntactic description, and 8,500 of these (mainly nouns, approx. 10,000 senses) are encoded with semantic information (Braasch & Sandford Pedersen 2002). Particular importance has been attached to a systematic and uniform treatment of all lemmas, which is documented in detail in the STO Linguistic Specifications (2005). The encoding principles follow the so-called lexicalist approach, which means that a *lexical entry* includes much information that traditionally is expressed in *grammar* rules. This approach has proved to be advantageous in NLP applications because looking up information in the lexicon is faster than computing by means of grammar rules. The boundary between lexicon and grammar in NLP moves along a sliding scale as long as the basic requirement is met. That is, all relevant information must be present either in the form of general grammar rules or included in the individual lexicon entries. Recent trends in general lexicography have much in common with the lexicalist approach. Many present-day dictionaries contain patterns, construction codes or other formalised information on the usual syntactic contexts of the headword. Hunston & Francis (2000) reports on a lexical grammar approach (what they call *pattern grammar*) in relation to the Collins COBUILD English Dictionary project. Also the six-volume Dictionary of Contemporary Danish (DDO) enumerates the syntactic constructions of the headword in a semi-formalised way.

2.2 Syntactic description

The syntactic description in STO is highly inspired by the valence theory (Helbig & Schenkel 1980) and its adaptation to NLP purposes (Somers 1987) comprising fine-grained information on the combinatorial properties of lemmas. All information is encoded on the basis of occurrences of the lemma in a corpus of 35 M tokens in order to capture its various syntactic contexts. Preference is given to more frequent syntactic patterns when an exhaustive encoding was not feasible because of the overwhelming number of syntactic constructions of a lemma (Braasch & Sandford Pedersen 2002).

The present investigation focuses on verbs because syntax is most prevalent for verbs: they subcategorize for a variety of complements, which naturally correspond to semantic arguments. That is why the syntactic descriptions of verbs (viz. more than 8,500) make up a material that lends itself to systematic investigation from the semantic point of view.

First, the basic, inherent properties of the lemma itself are described according to its syntactic category membership, e.g. reflexivity, use of auxiliary, and in the case of a phrasal verb also the particle in question. Second, the particular syntactic features of each selected construction of the verb are described as follows. The valency pattern contains the number of governed complements (i.e. *arity*, between 0 and 4). For each complement is given its syntactic function (e.g. subject, object) and its syntactic construction potential, the latter is expressed in syntactic categories (e.g. noun phrase, prepositional phrase, clause). Further details specified are the preposition of each governed prepositional phrase, the control type in case of infinite clauses (subject or object control), and whether the governed complement is mandatory or optional, similarly to what is mentioned in Hanks (2004:87) as "subvalency features". Each par-

ticular set of these features is expressed in a formalised *syntactic description* or a pattern. Basically, different syntactic descriptions of a lemma are encoded in distinct *syntactic units*; each noun, verb and adjective has at least one syntactic unit in STO.

3 The relationship between sense and syntax

In spite of the fact that there is no strict 1:1 relationship of syntactic and semantic readings, it is possible to observe several systematic connections or dependencies between sets of semantically related lemmas and the resemblance of their syntactic properties. Sinclair (1991:65) states: "It seems that there is a strong tendency for sense and syntax to be associated." Obviously, this fact has already been recognized by many linguists, but in the age of corpus linguistics the association between a syntactic structure and a given sense of a lexical item can be investigated systematically and in every detail as e.g. presented in Kilgarriff & Rundell (2002). Also Moon (1987:89 ff.) mentions the essential role of formal features as criteria for sense distinctions and says that "Syntax is one of the clearest criteria of all.". Further, she states on the basis of some illustrative examples that "In all these cases, other factors reinforce the meaning distinction shown by syntactic distinction, but syntax remains a fairly clear guide." This is an observation with undoubted validity and relevance for the investigation described below.[2]

3.1 Approaching sense groups from a syntactic angle

The investigations on possible semantic relationships between verbs that share syntactic description(s) in STO are based on the above observations saying that a syntactic complementation structure of a word reflects its semantic argument structure. The matter can be approached from different angles, e.g. producing a list of all the words provided with a particular syntactic description, or starting from the lemma side in order to study the disambiguating power of its syntactic patterns, etc. In the analysis presented here we adopted the following method:

- Extracting a list of all *syntactic descriptions* of verbs encoded in the lexicon, organised according to various parameters, such as complexity, phrase structure type, contained particle and/or preposition, population, etc. and selecting a particular syntactic description for investigation,
- Extracting the list of *lemmas* (here verbs) sharing the selected syntactic description.

Subsequently, the above can be combined for investigation of possible relationships between syntactic construction types and systematic polysemy by extracting a list of lemmas sharing more than one single syntactic description.

3.2 Illustrative examples of syntax-semantics relationships observed

In the following, observations on two highly different syntactic structure types of lexical

[2] Kohl et al., (WordNet 1998) approaches the subject from the opposite side in an interesting attempt to enhance the semantic treatment of verbs in the Princeton WordNet with syntactic frames. Fellbaum points to the observation "... the kind of semantic similarity on which the WordNet is built differs from that which is regularly reflected in syntax, although there is a significant overlap." (WordNet 1998:12).

units are presented. First, verbs with the simple, zerovalent syntactic pattern are dealt with in detail, in order to illustrate the method of semantic analysis and sense grouping. Second, verbs with the more complex, tetravalent pattern are examined and an overview of their basic sense groups is given. The zero- and the tetravalent pattern types respectively are only shared by a rather limited number of verbs, so these types are well suited to illustrating our approach. It should be noted that STO does not contain all Danish verbs, thus the lists below may appear incomplete.

3.2.1 Zerovalent verbs: Syntactic description and sense groups

Zerovalent verbs are intransitive verbs lacking an agent. They take the pronoun *det* "it" as formal subject (which is not included in the semantic arity number), e.g. *Det regner* (lit: "It rains" i.e. "It's raining"). These verb descriptions are identified by the basic code *Dv0.* Some zerovalent verbs are phrasal verbs, which is reflected in their syntactic descriptions. The basic code is extended accordingly, e.g. *Dv0x-ned*, where *x* refers to the phrasal structure of the verb, and *ned* ("down") is the particle itself. Altogether, 47 verbs are registered as zerovalent, and going through the list of verbs extracted and their corpus examples, the following observations on their structures appear:

- 37 are simple verbs (without particle)
- 10 are particle verbs (with one single occurrence of *af* ("off"), three occurrences of *op* ("up") and six occurrences of *ned* ("down") the two last mentioned being directional particles).

Grouping the verbs together from the semantic point of view, the following three semantic fields are identified:

- Weather/climate verbs: 37 (with/without particle, cf. Figure 1.)
- Sound emission verbs (all without particle): 6 (e.g. *bippe* "beep", *plaske_2* "splash", etc.)
- Miscellaneous (all without particle): 4, which are *rime* ("rhyme"), *spøge* ("haunt viz. a place"), *klodse* ("help a lot"), *kildre* ("tickle").

The last group consists of single instances of rather different sense groups. The boundaries between the sound emission verbs and the senses compised by the miscellaneous group are not clear-cut, e.g. one of the verbs of the latter group (*gynge* "swing") usually being used about a rhytmic motion, is also used about rhythmic sound emission. On the other hand, some weather verbs are closely related to sound emission verbs being onomatopoetic words, e.g. *plaske* ("plash / patter"). When having a weather sense, these verbs in Danish are phrasal verbs and they also appear with a subject, i.e. in a monovalent variant, where the realised subject is *regn* "rain" e.g. *Regnen siler ned,* lit. "The rain is pouring down". Furthermore, some weather verb senses are closely related to other semantic fields like motion, and in such senses, including metaphorical use, they are mono- or divalent. English equivalents of some of the mentioned sound emission verbs are registered also in Levin's (1993) English Verb Classes and Alternations (e.g. plash, patter, thunder).

Verbs compounded of two elements, such as *styrtregne,* composed of the verbs *styrte* "pour" and *regne* "rain", lit. "pour-rain", and *pjaskregne, plaskregne* lit. "splash-rain" (all translated by "It is pouring with rain"), show that the first element modifies or intensifies the meaning of the basic verb *regne* "rain" in a particular way. A compound verb with a weakening/diminishing first element is also found, viz. *småregne* lit. "small-rain" (i.e. "drizzle"). Ex-

amining the largest group of the weather/climate verbs, a few sub-domains can be distinguished: verbs relating to different natural phenomena such as precipitation (rain, snow and hail), wind, sky, daylight, thunderstorm and temperature. An interesting lexical gap is to be mentioned: in modern Danish, the verb related to *slud* "sleet" has almost disappeared (there are only 4 occurrences found in a corpus of ~50 M tokens).

The largest subgroup of these verbs, viz. 16, relate to rain (possibly, this fact pictures the usual weather conditions in Denmark). Here, the purely sound emission and miscellaneous groups will not be discussed further, because their monovalent pattern includes a subject (source or agent), e.g. *alarmen bipper* lit. "the alarm beeps", so they are dealt with differently. The table below summarizes the result of the meaning component analysis in terms of a semantic grouping. It shows the verbs sharing certain aspects of meaning, according to the weather phenomenon they are related to. The grouping indicates also semantic relationships like synonymy, antonymy and differentiating semantic features as well. The hierarchical relationship between the *basic verb* and the other verbs of a semantic group is a hyperonymy/superordinate relationship, in Fellbaum's terms (WordNet 1998:79) *troponymy* for verb hyponym. Further, the material allows for the recognition of some interesting aspects of lexical relationships, e.g. the semantically conditioned word-formation of compound verbs, as described above.

Semantic domain related to natural phenomenon		Lemma (+syntactic unit no.)	Differentia specifica	Particle	Semantic grouping
'precipitation'	'rain'	*regne*	*None*	-	***Basic verb***
		småregne, støvregne, dryppe	Intensity: light	-	'drizzling' group
		skybrøgne, styrtregne, øsregne,	Intensity: heavy	-	'pouring' group
		pøse, skylle, styrte, øse, pladre, plaske_1, pisse,	Intensity: heavy (+ direction)	ned	'pouring' group
		pjaskregne, plaskregne,	Intensity: heavy (+sound)	-	'splashing' group
	'snow'	*sne*	*None*	-	***Basic verb***
		fyge	Intensity: heavy (+ direction)	-	'drifting'
	'hail'	*hagle*	*None*	-	***Basic verb***
'wind'		*blæse_1*	*None*	-	***Basic verb***
		hufte	Intensity: light	-	'blowing' group (light/fresh/ strong/gale)
		storme	Intensity: heavy	-	
		blæse_2	Intensity: increasing	op	
		løje	Intensity: decreasing	af	
		trække	Intensity: unmarked (+ direction)	-	'draughting'
'sky'		klare	Intensity: increasing	op	'clearing up' group
		lysne		-	
'thunderstorm'		blitze, lyne	Emission of: light	-	'lightening' group
		tordne, dundre, gungre	Emission of: sound:	-	'thundering' group
'temperature'		fryse	Degree: below zero		'freezing'
		tø	Degree: above zero	-	'thawing'
'daylight'		lysne, dages	Intensity: increasing	-	'dawning' group
		blåne, mørkne	Intensity: decreasing	-	'getting dark' group

Figure 1. Semantic grouping of weather/climate verb senses

3.2.2 Tetravalent verbs: syntactic descriptions and sense groups – a brief summary[3]

As regards the relationship between tetravalent verbs and their sense groups, only a few principal observations on the most salient sense groups are presented here. Tetravalent constructions comprise four valency-bound complements: beside the subject and object, an indirect object and/or prepositional object(s); the last ones represent the source/origin and goal/target arguments, respectively. Tetravalent verb descriptions are identified by the basic code *Dv4,* followed by the codes for each of its valency-bound elements, etc. (cf. subsection 2.2). Altogether, 51 verbs are encoded with a tetravalent syntactic description comprising three subtypes.

- 49 verbs are encoded with the most prevalent tetravalent pattern that summarizes the syntactic behavior of *'X changes Y from Z to W'*. It reflects verb senses belonging to the externally controlled event type (domain: *cause_change*), distributed over various sense subgroups (some prototypic examples are shown in Figure 2.)
- 2 verbs of which one is provided with a description that summarizes the syntactic behavior of *'X pays Y to W for Z'* (semantic group: *pay*) and another with a description that summarizes the syntactic behavior of *'X sentences Y W for Z'*, with double object (semantic group: *bill/fine* see also in Levin 1993:47).

Semantic domain cause_change *of*	Lemma examples (+ syntactic unit no.)	Differentia Specifica	Comments
'location of the object'	*transportere* "transport" *bære 4* "carry"	with subject displacement	many subgroups
	sende 2 "send, transmit" *eksportere* "export"	no subject displacement	
'value of the object'	*opgradere* "upgrade"	increasing	subgroups on size, extent of the object
	nedjustere "cut down, reduce"	decreasing	
	omdefinere "redefine"	neutral	
'state of the object'	*opryklke* "move up"	upwards	subgroups on state-type
	degradere "degrade"	downwards	
	forandre "change, alter"	neutral	
'possession of the object'	*ekspropriere* "expropriate"	connotation (+, - or 0)	subgroups

Figure 2. Survey scheme of a semantic grouping of tetravalent verbs (Extract)

4 Conclusion and perspectives

The examination of zero- and tetravalent verbs illustrated a method of exploitation of the syntax-semantics relationships recognized. Although the verb lists were numerically small and semantically rather homogenous, the method can be employed for a full-scale analysis by applying it to systematically selected subsets of the large valency classes, e.g. divalent verbs. A full study will comprise all (890) different syntactic descriptions of 5,775 verbs, in total approx. 8,800 syntactic units. These patterns are unevenly distributed over the five arity classes. Obviously, the processing of large and semantically broad lists requires more steps of subdi-

[3] For a more comprehensive account for the result of this analysis and investigations into larger valency groups please see http://cst.dk/sto/referencer/index.html (Braasch, 2006).

vision and analysis refinement. A full-scale implementation of the method described has a promising prospective for a semi-automatic extension of the STO syntax with semantic information.

References

A. Dictionaries

Den Danske Ordbog (=DDO) (2003-2005). Gyldendal, Copenhagen.

B. Other Literature

Braasch, A., Olsen, S. (2000), 'Formalised Representation of Collocations in a Danish Computational Lexicon', in Heid, U. et al., (eds.) *The Ninth EURALEX International Congress, Proceedings,* Stuttgart, pp. 475-488.

Braasch, A., Sandford Pedersen, B. (2002), 'Recent Work in the Danish Computational Lexicon Project „STO"', in Braasch, A., Povlsen, C. (eds.) *Proceedings from the Tenth Euralex International Congress*, CST, Copenhagen, pp. 301-314.

Braasch, A. & Povlsen, C. (eds.) (2002), *Proceedings from the Tenth Euralex International Congress*, CST, Copenhagen.

Fellbaum, Ch. (1998), 'A semantic Network of English Verbs', in WordNet.

Hanks, P. (2004), 'Corpus Pattern Analysis', in Williams, G. & Vessier, S. (eds.) *Proceedings of the Eleventh EURALEX International Congress,* Lorient, pp. 87-97.

Helbig, G. & Schenkel, W. (1980), *Wörterbuch zur Valenz und Distribution Deutscher Verben*, Leipzig, VEB Bibliographisches Institut.

Hunston, S. & Francis, G. (2000), *Pattern grammar. SCL*, Amsterdam/Philadelphia, John Benjamins.

Kilgarriff, A., Rundell, M. (2002), 'Lexical Profiling Software and its Lexicographic Applications – a Case Study', in EURALEX (2002), pp. 807-818.

Kohl, K.T., Jones, D.A., Berwick, R.C. & Nomura, N. (1998), 'Verb Alternations in WordNet', in WordNet.

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

Moon, R. (1987) 'The analysis of Meaning', in Sinclair, J. (ed.) *Looking Up. An Account of the COBUILD Project in lexical computing*, London & Glasgow, Collins LT, pp. 87-103.

Olsen, S. (2002), 'Some Aspects of the Syntactic Encoding of Nouns, in a Computational Lexicon – the STO project', in EURALEX (2002), pp. 159-168.

Sinclair, J. (1991), *Corpus, Concordance, Collocation*, Oxford, Oxford University Press.

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

STO Linguistic Specifications (V.1.1, 2005), Internal document. Available together with the data or on request (cf. http://cst.dk/sto)

WordNet = WordNet. An Electronic Lexical Database. (1998). Fellbaum, Ch. (ed.) The MIT Press, Cambridge, Massachusetts, London.