Nor Jnl Ling 27.1, 97–126 © 2004 Cambridge University Press DOI: 10.1017/S0332586504001179
Printed in the United Kingdom

Pedersen, Bolette Sandford & Patrizia Paggio. 2004. The Danish SIMPLE lexicon and its application in content-based querying. *Nordic Journal of Linguistics* **27.1**, 97–126.

The Danish SIMPLE lexicon and its application in content-based querying

Bolette Sandford Pedersen & Patrizia Paggio

This paper deals with the SIMPLE-DK lexicon, a computational lexicon for Danish developed at the Centre for Language Technology in Copenhagen within the European Union project SIMPLE. The general SIMPLE model, on which the Danish lexicon is based, is presented, and the way in which several specific aspects of Danish, such as nominal compounds and time expressions, are accommodated in this model is then described. Phrasal verbs – in particular phrasal motion verbs – are shown to be a challenging phenomenon since they are difficult to place in the SIMPLE event ontology, and pose problems regarding the interpretation of the directional particle they combine with. The encoding strategy that is proposed here accounts for compositional and non-compositional types of phrasal verb, and captures the relation between act-denoting and transition-denoting senses of the same verb in terms of regular polysemy. The final part of the paper deals with the exploitation of SIMPLE-DK as an ontological and lexical source in the Danish project on content-based querying OntoQuery. In the OntoQuery ontology, the structured concepts in SIMPLE-DK are combined with nutrition concepts, and the resulting ontology is used for matching evaluation. It is also discussed how selectional restrictions and qualia roles from SIMPLE-DK can be included in a conceptual grammar to be used for query and text analysis.

Keywords content-based search, Danish, ontology, semantic computational lexicons, SIMPLE

Pedersen, Bolette Sandford & Patrizia Paggio, Center for Sprogteknologi, Københavns Universitet, Njalsgade 80, DK-2300 S. E-mail: bolette@cst.dk, patrizia@cst.dk

1. INTRODUCTION

The aim of this paper is twofold. On the one hand the purpose is to present the Danish computational SIMPLE lexicon as an individual and self-contained result of the much larger European Union project SIMPLE (Semantic Information for Multifunctional Plurilingual Lexica), which aimed at providing harmonised semantic lexicons for Natural Language Processing for 12 of the European languages (Lenci et al. 2000a). On the other hand, we discuss the application of SIMPLE-DK to content-based querying in the Danish research project OntoQuery.

In order to illustrate the research problems that the SIMPLE model presents in relation to a Scandinavian language like Danish, we describe some of the central aspects of the lexicon regarding the semantic description of Danish nouns, such as noun-noun compounds and time entities where an extended qualia structure ensures a rather rich semantic description. Other abstract nouns, in contrast, prove more complex to describe semantically via the project's ontology and the qualia structure. Finally we examine the problems encountered when encoding the Danish verbal system in the SIMPLE model. Speaking in Talmy's terms (Talmy 1985), Danish is a satellite-framed language where prepositions and adverbial particles express what in many other languages is expressed by the verb stem. This aspect constitutes a challenge for a universal, strictly modular framework in which semantic information is anchored to the governing word classes and their arguments.

Computational lexicons are built for computer applications; therefore an additional aim of this paper, as already mentioned, is to describe the application of the lexicon to an actual computer application, thereby contributing to an initial evaluation of its practical usefulness. As can be read from the acronym, the idea behind SIMPLE is that the lexicons developed within the framework should be multifunctional; in other words it should be useful for many kinds of computer applications, for example, machine translation and content-based information retrieval. These two applications, which both require semantically rich language resources, traditionally rely on two different aspects of what can be labelled 'semantic information'. In machine translation, issues such as argument structure and word sense disambiguation often prove to be very critical factors, whereas the internal semantic structure – or ONTOLOGY – behind a given vocabulary has received more attention in recent approaches to text retrieval.

In this paper we only consider the latter application type, a system that deals with ontology-based querying and search: the OntoQuery Prototype (Andreasen et al. 2000, 2004). As we shall see, only a subset of the semantic information coded in SIMPLE-DK has been exploited in OntoQuery since the methodology adopted in the project mainly focuses on the ontological structure of the vocabulary. Thus, the ontology used in OntoQuery to determine how well the conceptual content of a text matches a given query, is extracted from SIMPLE-DK. The OntoQuery system deals with a database of nutrition texts and we explain how the SIMPLE lexicon – concerned primarily with general language vocabulary – has been merged with a domain-specific ontology. We show how the information types taken over from SIMPLE (domain specification as well as hyponymy and synonymy relations) are used by the searching algorithm and we finally discuss which additional information types (such as selectional restrictions and qualia structure) can profitably be built on in more advanced approaches to querying, ones involving also semantic annotation and sense disambiguation.

https://doi.org/10.1017/S0332586504001179 Published online by Cambridge University Press

2. THE DANISH SIMPLE LEXICON

2.1 SIMPLE and the internal complexity of lexical items

The SIMPLE project builds on the LE-PAROLE lexicons, which contain 20,000 entries with corresponding morphological and syntactic information for each of the 12 languages covered in the project, cf. Ruimy et al. (1998).[1]

The language-specific encodings in SIMPLE are performed on the basis of a unified, ontology-based semantic model – the so-called SIMPLE model – representing an extended qualia structure based partly on Pustejovsky (1995), partly on experience in previous lexical projects such as the Eureka project GENELEX (Antoni-Lay et al. 1994), WordNet (Fellbaum 1998) and EuroWordNet (Vossen 1999). A general design model is thus provided allowing for the encoding of a large amount of semantic information, such as ontological typing, domain information, semantic relations, argument structure, event structure and selectional restrictions.

One of the fundamental assumptions behind the SIMPLE ontology is that lexical items vary in their internal complexity (cf. Lenci et al. 2000b:15–19, 25–27). This can be understood in two ways: (i) how many dimensions of meaning are associated with an item, and (ii) how many senses (e.g. semantic types) the item incorporates. Pustejovsky's theory of lexical meaning (Pustejovsky 1995), relying on the qualia structure and on a highly structured lexicon in general, constitutes the backbone of the SIMPLE ontology since it proposes a strategy for accounting for exactly this internal complexity of meaning.

Regarding the meaning dimensions associated with an item, consider for illustration the word *biksemad* 'hash', defined in the following way in a medium-sized Danish lexicon (*Nudansk Ordbog*, henceforth *NDO*):

> en ret der laves af en rest kogt el. stegt kød og kogte kartofler der skæres i stykker og steges på panden med løg; serveres med spejlæg
> [a dish made of left-overs of boiled or fried meat and boiled potatoes which are cut into pieces and fried in a pan with onions; served with a fried egg]

The essential parts of these meaning dimensions can be structured by means of Pustejovsky's four qualia roles, as shown in Figure 1. In other words, at least four meaning dimensions come into play for *biksemad*: (i) the formal role, which provides information about its affiliation in the ontology (by means of the *is_a* relation, which corresponds to the genus part of the definition): hash is a kind of dish; (ii) the constitutive role, which expresses a large range of semantic relations typically concerning the internal structure of the concept, in this case *part of*: that it consists of different ingredients; (iii) the telic role, which describes the typical function of the item (here a *used for* relation: a dish is meant to be eaten), and finally (iv) the agentive role, which describes the origin of the item – basically whether it

https://doi.org/10.1017/S0332586504001179 Published online by Cambridge University Press

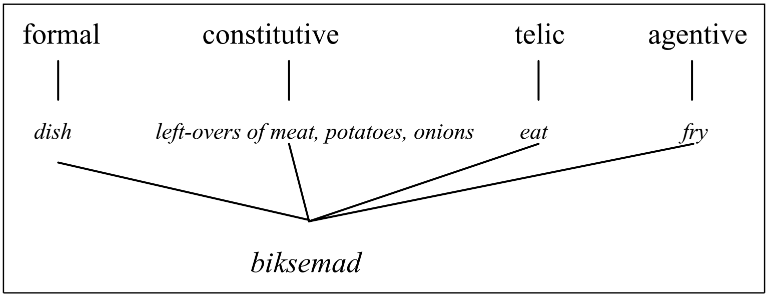


Figure 1. The qualia structure of *biksemad* 'hash'.

is natural or man-made and, if the latter, by means of what process (in this case a *made by* relation). The other complexity parameter, namely how many senses a word incorporates, relates to the classical notion of polysemy and in particular to regular polysemy, i.e. cases where more than one word from the same ontological class follow the same pattern of meaning change. This is typically the case for institutions like *universitet* 'university', which can relate either to the group of people making up or working in the institution or to the actual building of the institution.[2]

Three different types are established in the SIMPLE model in order to account for these different levels of complexity: SIMPLE TYPES, UNIFIED TYPES and COMPLEX TYPES. Simple types are applied to all so-called basic categories (corresponding more or less to natural kinds as referred to by Cruse (1991:140)) and to concepts with RIGID PROPERTIES, as presented in Guarino & Welty (2000:section 3.1) such as *himmel* 'sky', *blomst* 'flower' and *bakke* 'hill'. Basic categories are considered to be mono-dimensional and thus only inherit from the formal role. They can be defined in terms of a mono-dimensional hierarchy, by which we mean that they are organised uniquely by means of hyponymy relations. In contrast, unified types – examples of which are concepts like *biksemad* 'hash' and *lærer* 'teacher' – are multidimensional with multiple coordinates although also GROUNDED on a simple type. They are organised through the principles of orthogonal inheritance, which is a way of enriching a conventional inheritance structure by defining semantic relations on the four dimensions introduced by the qualia structure. In formal terms, orthogonal inheritance can be defined as multiple inheritance with the restriction that a node can only inherit from one mother node in the same partition – given that each of the four dimensions forms its own partition.

Finally, complex types are established in the model in order to account for regular polysemous classes such as institution/human group (*universitet*) and semiotic artifact/information (*bog* 'book'), etc. This can be seen as a first provisional way of expressing UNDERSPECIFIED semantic types as denoted by Pustejovsky's complex

https://doi.org/10.1017/S0332586504001179 Published online by Cambridge University Press

Semantic unit	*biksemad*
Definition:	*en ret der laves af en rest kogt el. stegt kød og kogte kartofler der skæres i stykker og steges på panden med løg; serveres med spejlæg* (*NDO*) 'a dish made of left-overs of boiled or fried meat and boiled potatoes which are cut into pieces and fried in a pan with onions; served with a fried egg'
Corpus example:	*om mandagen fik de ofte biksemad* 'on Mondays they often had hash'
Ontological type:	ArtifactFood
Ontological supertypes:	Concrete_Entity | Agentive | Telic
Domain:	General
Formal quale:	is_a = *ret* 'dish'[i]
Agentive quale:	created_by = *stege* 'fry'
Telic quale:	used_for = *spise* 'eat'
Constitutive quale:	has_as_parts= *kødrest, kartoffel, løg* 'left-overs of meat, potato, onion'
Complex (regular polysemy):	Nil
Synonymy:	Nil

[i] Note that relations are established among existing concepts in the lexicon; thus *ret* 'dish' is an existing unambiguous concept in the actual lexical database uniquely referred to by an identifier (e.g. USEM_N_*ret*_FOO_1) that expresses level of description (USEM for semantic unit in contrast to morphological or syntactic unit), word class (N for noun), lemma (*ret*), ontological type (FOO for Food) and reading number.

Figure 2. Semantic entry for *biksemad* 'hash'.

types (Pustejovsky 1995:118). In other words, complex types allow for two semantic items to be interrelated by an information slot called COMPLEX.

Consider, in Figure 2, a full lexical entry for *biksemad*, which constitutes a unified type grounded on the basic ontological type ArtifactFood, and inheriting also from the Telic and Agentive dimensions. It should be mentioned that most definitions in the Danish SIMPLE lexicon are taken over from *NDO*.

2.2 *Expressing the internal structure of concrete nominal compounds*

Other examples of concrete noun encodings are nominal compounds. Where the internal semantic structure of Danish deverbal nominals can to some extent be identified by the argument structure of the derived verb and the internal ranking of its arguments (Ørsnes 1995), non-deverbal nominal compounds, in contrast, display a much more arbitrary internal structure in Danish (cf. Paggio & Ørsnes 1993). This means that they require semantically more complex lexical entries: the qualia structure as expressed in the SIMPLE model provides a good basis for their semantic encoding (see also Pedersen & Keson 1999).[3]

https://doi.org/10.1017/S0332586504001179 Published online by Cambridge University Press

Semantic unit	*fedtdepot* ‘fat deposit’
Definition:	*depot i kroppen hvori der er ophobet fedt* ‘deposit in the body in which fat is accumulated’
Corpus example:	*Under graviditeten opbygger moderen især på lår og baller fedtdepoter.* ‘During pregnancy the mother accumulates fat deposits especially on thighs and buttocks’.
Ontological type:	Body_Part
Ontological supertypes:	Concrete_Entity | Constitutive
Domain:	Medicine/Health
Formal quale:	is_a = *depot* ‘deposit’
Agentive quale:	Nil
Telic quale:	Nil
Constitutive quale:	Is_a_part_of *krop* ‘body’ Contains *fedt* ‘fat’
Complex (regular polysemy):	Nil
Synonymy:	Nil

Figure 3. Semantic entry for *fedtdepot* ‘fat deposit’.

As an example, we can consider nominal compounds denoting artifactually made containers and therefore belonging to the ontological type Container. For instance, the encoding of the meaning of a word like *bæger* ‘cup’ can be further augmented with the telic role for such compounds as *målebæger* ‘measuring cup’, *raflebæger* ‘dice cup or *drikkebæger* ‘drinking cup’.

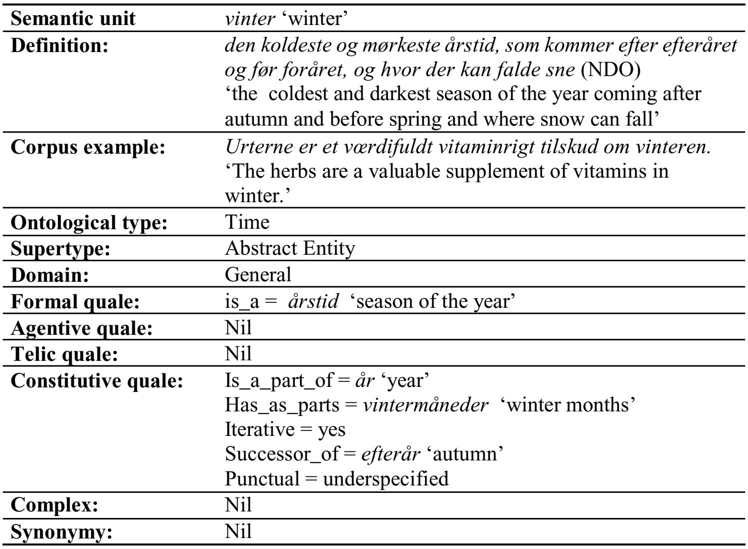
In the SIMPLE model, the encoding of the meaning of *bæger* can also be further specified by including information about the constitutive role for other types of compounds. The encodings can include additional information on (i) the material of which the container is made by means of a *made by* relation, e.g. *plasticbæger* ‘plastic cup’, *messingbæger* ‘brass cup’, or *papbæger* ‘paper cup’, or (ii) what the container (prototypically) contains, e.g. *askebæger* ‘ashtray’, *yoghurtbæger* ‘yoghurt cup’.

Also non-artifacts like body parts can include constitutive relations like the *contains* relation (in addition to their type-defining relation *part of*). An example from the nutrition domain is *fedtdepot* ‘fat deposit’, illustrated in Figure 3. The entry for *fedtdepot* is one of the 1,000 entries added to SIMPLE to cover the nutrition domain targeted in OntoQuery, and is thus an example of the flexibility of the semantic framework adopted in SIMPLE. The entry will be discussed also from the perspective of querying in section 3.3.

2.3 Abstract nouns

2.3.1 Expressing time entities by means of extended qualia structure

In SIMPLE, the possible value types of Pustejovsky’s qualia roles have been extended to encompass more fine-grained distinctions concerning a large set of semantic types.



Semantic unit	*vinter* ‘winter’
Definition:	*den koldeste og mørkeste årstid, som kommer efter efteråret og før foråret, og hvor der kan falde sne* (NDO) ‘the coldest and darkest season of the year coming after autumn and before spring and where snow can fall’
Corpus example:	*Urterne er et værdifuldt vitaminrigt tilskud om vinteren.* ‘The herbs are a valuable supplement of vitamins in winter.’
Ontological type:	Time
Supertype:	Abstract Entity
Domain:	General
Formal quale:	is_a = *årstid* ‘season of the year’
Agentive quale:	Nil
Telic quale:	Nil

Constitutive quale:	Is_a_part_of = *år* ‘year’
	Has_as_parts = *vintermåneder* ‘winter months’
	Iterative = yes
	Successor_of = *efterår* ‘autumn’
	Punctual = underspecified

Complex:	Nil
Synonymy:	Nil

Figure 4. Semantic entry for *vinter* ‘winter’.

For each qualia role a set of possible semantic relations and features have been established and so-called templates have been elaborated for each ontological type in order to suggest which type-defining relations and features this particular type should include. For abstract concepts like time concepts, this extended structure opens up for a fairly rich semantic description, as in the lexical entry for *vinter* ‘winter’ in Figure 4, where in particular the constitutive role is rich in information.

From this formal description we can deduce that winter is an iterative event, which takes place again and again, that it is underspecified with respect to punctuality (can refer both to a period and a point of time), is a season of the year which constitutes part of the year; comes after autumn, and has winter months as its parts.[4]

2.3.2 Other abstract nouns

In general, however, the assignment of ontological type and semantic relations and features to abstract nouns is a very difficult task; a fact reflected in the Danish encodings which, in the case of abstract nouns, are generally not as rich in information as the concrete nouns. Not surprisingly, the semantics of abstract nouns is generally of a much more complex nature than what can be structured through the existing relations and features of qualia structure in SIMPLE. An example is an abstract concept with a rather delimited meaning like *alibi* ‘alibi’. An alibi is something that proves that someone is innocent of some crime since the accused person was somewhere else at the moment of the crime. One could assign a telic role of *bevise* ‘prove’ to this

https://doi.org/10.1017/S0332586504001179 Published online by Cambridge University Press

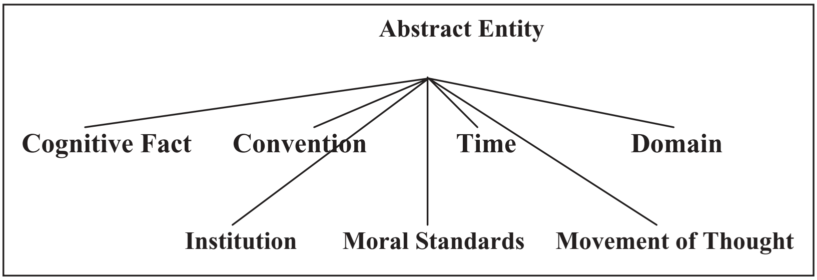


Figure 5. Ontology of abstract nouns.

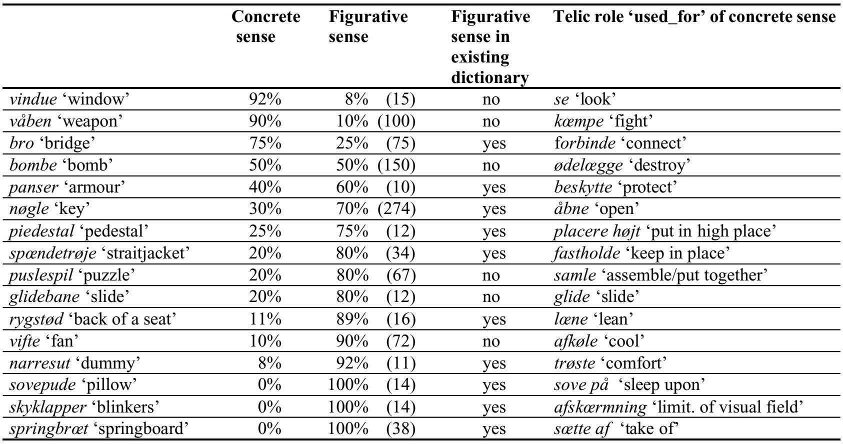
concept, but such a description would still be a very rudimentary approximation of the meaning of *alibi* since the dimension of being somewhere else than where a crime takes place can hardly be formalised with the formal apparatus at hand. The strategy adopted in the Danish SIMPLE lexicon with regard to abstract nouns other than time concepts is, therefore, a limited assignment of semantic information as well as a rather vague ontological labelling. The ontological node Abstract Entity in the SIMPLE ontology allows for the following sub-types (Figure 5):[5]

It turns out, however, that a considerable number of abstract concepts could not be assigned any of these types and are therefore encoded with the dominating type: Abstract Entity. An alternative to this strategy could be to develop a much more fine-grained ontology for these abstract concepts; the problem being, however, to formally distinguish these from one another. Given the current encoding of this section of the vocabulary, it is to be expected that this part of the underlying ontology will not have much discriminating power, for instance, when used by a searching algorithm.

Another remark concerns concrete entities used metaphorically and thus ACTING AS abstract entities. Since the Danish SIMPLE lexicon is very much based on actual word occurrences in Danish corpora, this phenomenon calls for considerable attention due to its frequency. Consider Figure 6, which shows the frequency of figurative senses of a given set of nouns in Danish newspaper corpora.[6] Note that only a little more than half of these figurative senses are at all mentioned in existing dictionaries.

The question is to what extent the telic role of the concrete sense is transferred to the metaphoric sense. It is precisely the telic role that predicts the meaning of the abstract sense to a considerable extent. However, we can only speak of a telic role associated with the abstract sense if we understand the relation itself as abstract. For instance, the figurative sense of ‘pillow’ denotes something that you use for sleeping in the *figurative sense* of the verb, i.e. a false security coming from previous achievements that makes you lazy or ‘blind’ regarding new initiatives. A similar example is *puslespil* ‘puzzle’, which in its figurative sense denotes a complex case where different bits and parts need to be PUT TOGETHER in order, for instance, to solve

https://doi.org/10.1017/S0332586504001179 Published online by Cambridge University Press



	Concrete sense	Figurative sense	Figurative sense in existing dictionary	Telic role 'used_for' of concrete sense
vindue 'window'	92%	8% (15)	no	*se* 'look'
våben 'weapon'	90%	10% (100)	no	*kæmpe* 'fight'
bro 'bridge'	75%	25% (75)	yes	*forbinde* 'connect'
bombe 'bomb'	50%	50% (150)	no	*ødelægge* 'destroy'
panser 'armour'	40%	60% (10)	yes	*beskytte* 'protect'
nøgle 'key'	30%	70% (274)	yes	*åbne* 'open'
piedestal 'pedestal'	25%	75% (12)	yes	*placere højt* 'put in high place'
spændetrøje 'straitjacket'	20%	80% (34)	yes	*fastholde* 'keep in place'
puslespil 'puzzle'	20%	80% (67)	no	*samle* 'assemble/put together'
glidebane 'slide'	20%	80% (12)	no	*glide* 'slide'
rygstød 'back of a seat'	11%	89% (16)	yes	*læne* 'lean'
vifte 'fan'	10%	90% (72)	no	*afkøle* 'cool'
narresut 'dummy'	8%	92% (11)	yes	*trøste* 'comfort'
sovepude 'pillow'	0%	100% (14)	yes	*sove på* 'sleep upon'
skyklapper 'blinkers'	0%	100% (14)	yes	*afskærmning* 'limit. of visual field'
springbræt 'springboard'	0%	100% (38)	yes	*sætte af* 'take of'

Figure 6. Figurative senses of concrete nouns (Nimb & Pedersen 2000:682).

a crime. In practice, frequent metaphors are encoded in SIMPLE-DK and linked to the concrete senses by means of the regular polysemy relation. The telic role is assigned to the abstract senses by applying, where possible, an abstract concept to denote the *used for* relation.

2.4 Expressing the Danish verbal system in a verb-framed model

Several aspects of the Danish verbal system constitute a challenge for a universal, strictly modular framework like the SIMPLE model, which focuses on the governing word classes and their arguments. As previously mentioned, Danish is a typical satellite-framed language, meaning that, for example, prepositions and adverbial particles express what in many other languages forms part of the meaning of the verb, cf. Herslund 1993, Durst-Andersen & Herslund 1996, Harder, Heltoft & Nedergaard-Thomsen 1996, Pedersen 1999. Thus, several of the most frequent verbs in Danish are relatively neutral with respect to semantic affiliation in the ontology as well as regarding event type, and their affiliation is determined rather by the particle or the preposition they combine with than by the verb stem itself. In fact, from our corpus examinations, we estimate that more than half of the verb senses relevant for SIMPLE (relevance is here based on frequency) is constituted by phrasal verbs which cannot be uniquely assigned a semantic type on the basis of the verb stem alone.

Representing this and other incorporation phenomena of Danish vocabulary where the verbal predicate is complex,[7] is not only a challenge to traditional lexicography but even a more serious challenge to computational lexicography, which typically strives for a modular composition of the lexicon which clearly distinguishes between morphology, syntax and semantics, and which focuses on the governing

https://doi.org/10.1017/S0332586504001179 Published online by Cambridge University Press

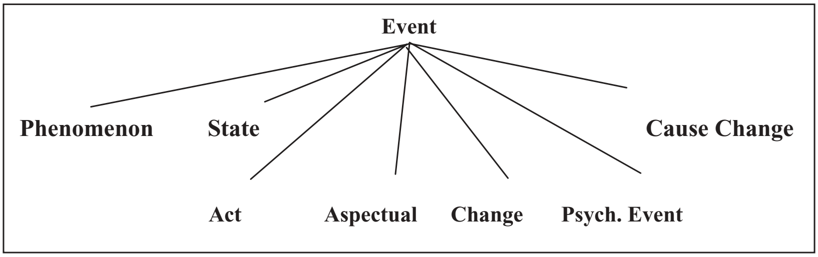


Figure 7. Event types in the SIMPLE ontology.

word classes: nouns, adjectives and verbs, and the arguments that they take. Such a model seems intuitively better-suited for a verb-framed language, which encodes the core meaning components in the verbal stem.

Before discussing how phrasal verbs are treated in SIMPLE-DK, we need to introduce the part of the ontology which deals with events. Verbs and event nouns are affiliated under the node Event, which again dominates a whole sub-hierarchy of types to be used when classifying different kinds of events (cf. Lenci et al. 2000b:29–30). The sub-ontology for events is influenced by several sources, including in particular WordNet (Miller et al. 1990), EuroWordNet (Alonge et al. 1998) and Levin's verb classes (Levin 1993). One of the aims has been to find a number of event classes which are richer than that of WordNet, comprising 15 classes, and less detailed than Levin's 234 classes. Thus, the SIMPLE event ontology comprises 59 classes grouped into seven core categories, as shown in Figure 7.[8] Three fundamental aspects have been considered in this classification:

- event type, i.e. basically whether a verb sense denotes a STATE (as the ontological types Phenomenon and State), an act (Act and Psychological Event) or a transition (Aspectual, Change and Cause Change)
- argument structure, i.e. arity and type of arguments selected by the verb sense
- causativity, i.e. whether a verb sense is causative or non-causative, the former always being represented by a unified type.

In order to discuss the specific problems of Danish satellite verbs, we need to see how the semantic lexicon is linked to syntax and morphology according to a model that SIMPLE has inherited from the previous lexicon-building project, LE-PAROLE. As described in Ruimy et al. (1998) and further specified for Danish in Braasch & Pedersen (2002:303), each concept (i.e. each semantic unit) is further linked to its corresponding syntactic and morphological units. Consequently, the model permits to distinguish different syntactic behaviours on pure syntactic grounds and independently of whether or not syntactic units share meaning. Figure 8 illustrates the linking

https://doi.org/10.1017/S0332586504001179 Published online by Cambridge University Press

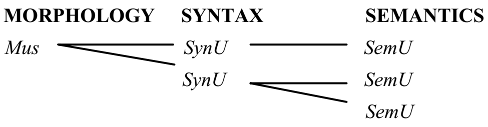


Figure 8. Linking of units at the three levels.

of units at the different levels (*Mus* = morphological unit, *SynU* = syntactic unit, *SemU* = semantic unit). As can be seen, one syntactic unit can very well link with two semantic units, which are then maybe assigned two different types in the event ontology.

2.4.1 Identification of phrasal verbs

When analysing Danish verbs and their satellites, unit accentuation plays a central role (see Scheuer 1995, Harder et al. 1996, Nedergaard-Thomsen & Herslund 2002). In fact, unit accentuation can be used as a linguistic test in order to distinguish phrasal verbs from other verbs combined with adverbial particles. Consider the examples in (1) and (2).

(1) Han 'blev 'væk.
he stayed away
'He didn't show up.'
(2) Han blev 'væk.
he stayed away
'He got lost.'

Example (1) has stress on both the verb and the particle, indicating the fact that we are dealing with a simplex verb *blive* 'stay' combined with an adverbial modifier *væk* 'away'.

In contrast, in (2), the absence of full stress on the verb indicates that the verb does not constitute a clausal predicate on its own but that the element that receives full stress (the particle) should be interpreted as integral part of the semantics of the predicate. In the case of (1), *blive* can be described as a state verb – i.e. as non-transitional – subcategorising for a locational argument; in the case of (2), we must consider *blive 'væk* as a phrasal verb and thereby a single semantic unit of the ontological type 'Change Location'.

A closer look at the group of verbs which can be categorised as phrasal verbs according to the test described above reveals a blurred picture of those phrasal verbs that are compositional in meaning on one hand and those that are not on the other. By compositionality we mean that both the host verb and the particle retain their core

https://doi.org/10.1017/S0332586504001179 Published online by Cambridge University Press

meaning. This is normally the case when directional particles are combined with motion verbs, as in (3) and (4).

(3) Han løb 'ud.
he ran out

(4) Han gik 'op.
he walked up

Compositionality is an important parameter when deciding how to represent a phrasal verb in the computational lexicon. In fact, also in traditional dictionaries this distinction is usually maintained although several Danish dictionaries are not completely clear on this point. Normally, however, only the non-compositional phrasal verbs find their way into traditional dictionaries, usually indicated as SUBLEMMAS of the simplex verb as, for instance, in the cases of *vaske op* and *løbe ud*:

(5) Han vaskede 'op.
he washed up
'He did the dishes.'

(6) Fristen løb 'ud.
the-deadline ran out

In contrast, compositional phrasal verbs, which are predictable in meaning and often productive with respect to the directional particle they combine with, are rather described by means of valency patterns in the 'core' entry, as in *løbe op/ned/ud/. . .* 'he ran up/down/out/. . . ', resulting in the following valency pattern description: SUBJECT + DIRECTIONAL.[9]

However, this treatment of compositional phrasal verbs does not provide the means to distinguish them from 'regular' verb phrases, such as the one in (7) below, where the directional adverb is not incorporated in the verb.

(7) Han 'kiggede 'ned.
he looked down

The best way to capture the difference between (7) and phrasal verbs of the type illustrated in (3) and (4) above is to recognise that motion verbs constitute a very special semantic class. Motion verbs occurring as phrasal verbs are mostly predictable in meaning since the directional particle is usually to be understood in its core sense; motion verbs thus constitute a unique semantic class in that the directional marker they combine with is INCORPORATED in the verb, as is shown by the accentuation test. This relates well to the fact that other motion verbs inherently express direction without the need for a directional particle. It also makes it possible to view the expression of *direction* as a regular syntactic pattern which alternates with directional prepositional phrases:

(8) Han løb ud/ til bageren/ hen til skolen
he ran out to the-baker over to the-school

Another important feature of the directional particle is that it mostly acts as an aspectual marker, typically changing a process verb into a transition verb, as was seen in examples (3) and (4) above, where process motion verbs are changed into 'Change Location' verbs. In some cases, the aspectual marking is the only function of the particle – which then loses its directional meaning completely – as in examples (9) and (10) below, where process verbs like *spise* 'eat' and *drikke* 'drink' are changed into transition verbs.

(9) Han spiste 'op. TRANSITION
he ate up
(10) Han drak 'ud. TRANSITION
he drank out
'He drank up.'

This and several other features influencing the event type (such as definiteness of the object, the addition of a resulting state, etc.) must be taken into account when assigning ontological type to verb senses.

2.4.2 Representing phrasal verbs in the modular SIMPLE framework

The question is how to represent the semantics of phrasal verbs in a verb-framed model like the SIMPLE event ontology. It seems obvious that the idiosyncratic phrasal verbs must be fully lexicalised at the semantic level since their meaning is unpredictable and therefore requires a semantic description of its own. For the compositional phrasal verbs on the other hand, we can opt for either a fully lexicalised representation or for a directional slot representation of some kind. In any case, we need to consider more thoroughly the whole PAROLE/SIMPLE structure in order to decide for a convenient strategy since the identification of the LEMMA, as well as the representation of syntactic information, are both highly relevant for this discussion.

In other words, when deciding how to represent the semantics of phrasal verbs in SIMPLE, we need not only consider how to represent them at the semantic level, but also how to find a principled solution concerning their representation at the morphological and the syntactic level. An interesting aspect of the full lexicon model (including both morphology, syntax and semantics) is that there exists no 'lemma' as such in the traditional sense of the word. In order to identify what in traditional lexicography makes up a lemma, one has to start from the semantic unit and work

https://doi.org/10.1017/S0332586504001179 Published online by Cambridge University Press

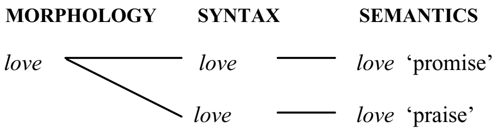


Figure 9. The representation of *love* 'promise, praise'.

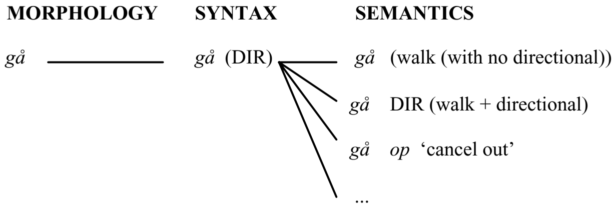


Figure 10. 'Split late' representation of *gå* 'walk'.

back through the syntactic and morphological levels, and gather all the relevant information.

This is due to the fact that the Danish PAROLE and SIMPLE lexicons consistently apply the so-called SPLIT LATE strategy. A split late strategy implies that only what can be identified at a particular level of analysis as two different units (morphology, syntax or semantics) should result in the entry being split into separate units. Thus, for the two homographs of *love* ('promise, praise') for instance, we find the representation in the Danish PAROLE and SIMPLE lexicons as in Figure 9.

To be more precise, even if we speak of homographs with different etymology and completely unrelated meanings, only one representation is given at the morphological level since the two are identical from a purely morphological point of view. The split into two units is realised at the syntactic level since the valency patterns of the two verbs differ, the former being ditransitive and the latter transitive.

For the representation of phrasal verbs several approaches can be adopted. We can either lexicalise a phrasal verb like *vaske op* 'do the dishes' at the morphological level and thus treat it as a completely different lexeme than *vaske*, or we can differentiate at the syntactic level and identify the phrasal verb *vaske op* in syntax. Or we can let the particle be treated as optional at the syntactic level. This is convenient, especially in cases of AMBIGUITY (i.e. where both a compositional and a non-compositional interpretations are possible, as in *gå op*, which can either be compositional in the sense of 'walk upwards' or non-compositional in the sense of 'cancel out'), since it prevents unnecessary redundancy at earlier levels and allows for a unified syntactic description of directionals at the syntactic level, irrespective of whether these are expressed as particles or as prepositional phrases.[10]

https://doi.org/10.1017/S0332586504001179 Published online by Cambridge University Press

Semantic unit:	*gå op_IDS* ‘cancel out’
Definition:	*(om regnestykke) løses så der ikke bliver nogen rest* ‘(about calculations) solve so that there is no remainder’
Corpus example:	*..men for at få regnestykket til at gå op måtte han indregne naboens grund* ‘but in order to make the calculation cancel out he had to include the neighbour’s garden’
Ontological type:	Identificational state
Ontological supertype:	Relational state
Domain:	Mathematics
Predicative rep:	Argument_1
Selectional restrictions:	Argument_1= Representational
Formal quale:	is_a = *tilstand* ‘state’
Agentive quale:	Nil
Telic quale:	Nil
Constitutive quale:	Relates = *tal* ‘numbers’
Complex (regular polysemy):	Nil
Synonymy:	Nil

Figure 11. Semantic representation of *gå op* ‘cancel out’.

Given this approach, the non-compositional phrasal verb *gå op* ‘cancel out’ can be given a specific, fully ‘lexicalised’ representation in the semantics and be assigned the type ‘Identificational State’, a sub-type of State which ascribes identity between numbers (Figure 11).

As for the compositional phrasal verbs, these are treated in the same way as constructions with weakly bound prepositions. Thus, the semantic unit in Figure 12 accounts for *gå* + directional, as in *gå op* ‘walk upwards’, as well as in *gå hen til bordet* ‘walk over to the table’, etc. The ontological type is ‘Change Location’.

In contrast, Figure 13 shows *gå* as a process verb, as in phrases like *han lærte snart at gå* ‘he soon learned to walk’, *han gik 2 km* ‘he walked 2 km’. This sense is assigned the ontological type ‘Motion’. The two semantic entries are interlinked by means of the complex type ‘Move/Change Location’, which indicates a relation of regular polysemy between the two. This approach ensures a proper event structure assignment, BUT it contradicts the guidelines given in the SIMPLE specifications (Lenci et al. 2000b), where it is stated that event type is meant to abstract away from the possible effects by complements and adjuncts. We claim, however, that it is NOT convenient to abstract away from this factor when assigning ontological types to Danish act verbs in general (of which motion verbs are a sub-type). In particular, seen in a multilingual perspective, it is a well-known fact that the lexicalisation border proves to differ radically between the Germanic and Romance languages allowing therefore a Danish motion + direction phrase to be translated into a Spanish directional verb lexeme, as seen in (11).

Semantic unit:	*gå_CHL* ('walk' – Change Location reading)
Definition:	*bevæge sig til fods fra et sted til et andet* 'move from one place to another by walking'
Corpus example:	*Vi skal* ***gå*** *hen til telefaxen , vente mens den kalder op osv.* 'we have to walk over to the fax machine, wait while it makes the call, etc.'
Ontological type:	Change_Location
Ontological supertype:	Cause_Change
Domain:	General
Predicative rep:	Argument_1, Directional

Selectional restrictions:	Argument_1= Human OR Animal
	Directional = Concrete

Formal quale:	is_a = *flytte_sig* 'move from place to another'
Agentive quale:	Nil
Telic quale:	Nil
Constitutive quale:	Manner = yes
Complex (regular polysemy):	Move/Change_Location = *gå_MOV* 'walk'
Synonymy:	Nil

Figure 12. Semantic representation of *gå* + directional.

Semantic unit:	*gå_MOV* ('walk' – Move reading)
Definition:	*komme frem ved at sætte den ene fod foran den anden* (NDO) 'proceed by putting one foot in front of the other'
Corpus example:	*han lærte at gå da han var 10 måneder gammel* 'he learned to walk when he was ten months old'
Ontological type:	Move
Ontological supertype:	Act
Domain:	General
Predicative rep:	Argument_1
Selectional restrictions:	Argument_1= Human OR Animal
Formal quale:	is_a = *bevæge_ sig* 'move'
Agentive quale:	Nil
Telic quale:	Nil
Constitutive quale:	Manner = yes
Complex (regular polysemy):	Move/Change_Location = *gå_CHL* 'walk/go'
Synonymy:	Nil

Figure 13. Semantic representation of *gå*.

(11) Han gik ud.
he walked out ⇨
salió
exited-3PERS

This phenomenon is known as the PATH-MANNER divergence (cf. Talmy 1985, Slobin 1996, Pedersen 1999 and others). It refers to the fact that Romance languages usually

https://doi.org/10.1017/S0332586504001179 Published online by Cambridge University Press

lexicalise the direction component (or ‘path’ in Talmy’s terms) in the verb stem whereas Germanic languages lexicalise the manner component. In Spanish the two verb stems *caminar* ‘walk’ and *salir* ‘go out’ are placed in two different parts of the ontology (under ‘Move’ and ‘Change Location’, respectively). In our approach, a similar semantic distinction holds between *gå* ‘walk’ and *gå ud* ‘go out’ in Danish, even if Danish lexicalises differently.

The unit accentuation test, which shows that both *gå ˈud* ‘walk/go out’ and *gå til staˈtionen* (go to the station) lose stress on the verb, is in our view a further indication of the fact that these two construction types belong together ontologically and should thus be affiliated under the same type.[11]

3. APPLYING THE LEXICON IN CONTENT-BASED QUERYING

3.1 Domain-specific extension of ontology and lexicon

In this section, we describe the application of the Danish SIMPLE lexicon in the Danish interdisplinary project OntoQuery, which deals with content-based querying, and the extensions to SIMPLE this has resulted in. The project aims at developing a methodology for content-based querying and retrieval which goes beyond superficial key word recognition, without, however, requiring a fully blown semantic analysis. Ontological knowledge plays a crucial role in this methodology since it is used to derive semantic descriptions from both queries and texts, and to determine the mutual closeness of these descriptions with respect to the concept ontology. The project focuses on one particular domain, namely the nutrition domain, and the project’s first prototype enables a user to query a text database of Danish nutrition texts taken from The Large Danish Encyclopaedia.

Before discussing the intended use of the ontology for text retrieval and the implementation of the OntoQuery methodology, we focus in this section on the ontology itself, and on the problems encountered when merging the ontology in SIMPLE with the domain-specific ontology required in OntoQuery. It has been discussed in the project whether the SIMPLE ontology and its language-specific instantiation could be seen as an ontology in the formal sense of the word, on which actual semantic inferences could be drawn, or whether it was rather a practical framework for lexicon building. Although some of the classifications in the SIMPLE ontology may require further analysis and development, the ontology has so far proven a good backbone to which further semantic restrictions may be added. In particular, this concerns, as we shall see, the formulation of a conceptual grammar capable of giving the ontology generative power (Nilsson 2001a).

The first issue to be discussed concerning the extension of the OntoQuery ontology from SIMPLE is the relation between the top part of the ontology, which

https://doi.org/10.1017/S0332586504001179 Published online by Cambridge University Press

- Top
    - Telic
        - *spise* 'eat'
    - Agentive
        - *stege* 'fry'
    - Entity
        - Concrete Entity
            - Food
                - Artifact Food
                    - *mad* 'food'
                        - *ret* 'dish'
                            - *biksemad* 'hash'

Figure 14. The concept *biksemad* 'hash' and its dominating nodes in the ontology.

has been carefully defined in SIMPLE on the basis of general semantic criteria, and the language-specific part. To appreciate the problem, it must be noted that, in each entry in SIMPLE-DK, a top-concept hyperonym (Ontological type) as well as a closest Danish hyperonym (Formal role) are defined, where the Danish hyperonym is derived from existing dictionary definitions. In other words, in SIMPLE-DK, further ontological structuring is added to the general language-independent model. Although this double typing made the extraction of the ontology for OntoQuery non-trivial, especially regarding Danish upper level concepts, it facilitated a COMPOSITE ontology of both language-independent top-concepts and language-specific concepts as seen in Figure 14 for the concept *biksemad* 'hash', discussed earlier.

Another issue concerns the addition of nutrition concepts to the original 10,000 semantic concepts. For this task, texts from a Danish encyclopaedia were used as a basis. This means that the nutrition ontology is built bottom-up mainly from the perspective of a layman, since this is the perspective adopted in the texts. Obviously, it also means that concepts are viewed (and structured) from a nutrition perspective and not, for instance, from a chemical perspective. Thus, minerals like iron and copper are considered sub-types of micro-nutrition stuff and not sub-types of elements in the chemical sense.

A few nodes are established in order to structure the ontology, such as *stof-i-krop* 'stuff in body' and *tilskudsstof* 'supplementary stuff'. The ontology links to two different nodes in the SIMPLE top-ontology, namely to Natural_Substance and Substance_Food, and is structured as shown in Figure 15.[12] In the figure, the concepts in square brackets indicate the inheritance path, and multiple inheritance is shown by a slash between two of the concepts in the path, e.g. [caroten|antioxidant].

https://doi.org/10.1017/S0332586504001179 Published online by Cambridge University Press

- **SUBSTANCE** [Concrete_entity]
    - **Natural Substance** [Substance]
        - ***NÆRINGSSTOF*** 'nutrition stuff' [Natural_Substance]
            - ***mikronæringsstof*** 'micro-nutrition stuff' [næringsstof]
                - ***Mineral*** 'mineral' [mikronæringsstof]
                    - ***mikromineral*** 'micro-mineral' [mineral]
                        - *jern* 'iron' [mikromineral]
                        - *zink* 'zinc' [mikromineral]
                        - *kobber* 'copper' [mikromineral] …
                    - ***salt*** 'salt' [mineral]
                    - ***makromineral*** 'macro-mineral' [mineral]
                        - *natrium* 'sodium' [makromineral]
                        - *kalium* 'pottasium'[makromineral] …
                - ***Ess-aminosyre*** 'ess amino acids'[mikronæringsstof]
                - ***Ess-fedtsyre*** 'ess. fatty acids' [mikronæringsstof]
                - ***Vitamin*** 'vitamin' [mikronæringsstof]
                    - ***vandopl. vitamin*** 'water soluble vitamin' [vitamin]
                        - *B-vitamin* [vandopl. vitamin]
                        - *C-vitamin* [vandopl.vitamin|antioxidant]
                    - ***fedtopl. vitamin*** (fat soluble vitamin) [vitamin]
                        - *A-vitamin* [fedtopl. vitamin]
                        - *D-vitamin* [fedtopl. vitamin]
                        - *E-vitamin* [fedtopl. vitamin|antioxidant] …
            - ***makronæringsstof*** 'macro-nutrition stuff' [næringsstof]
                - ***Fedt*** 'fat' [makronæringsstof]
                - ***Protein*** 'protein' [makronæringsstof]
                - ***Kulhydrat*** 'carbo hydrate' [makronæringsstof]
                    - ***kostfiber*** 'nutrition fibre' [kulhydrat]
                    - ***stivelse*** 'starch' [kulhydrat]
                    - ***alkohol*** 'alcohol' [makronæringsstof]
        - ***ANTIOXIDANT*** 'antioxidant' [Natural_Substance]
            - ***caroten*** 'carotene' [Natural_Substance]
                - ***Betacaroten*** 'betacarotene' [caroten|antioxidant]
        - ***STOF_I_KROP*** 'stuff-in-body' [Natural_Substance]
            - ***kolesterol*** 'colesterol' [stof-i-krop]
            - ***enzym*** 'enzyme' [stof-i-krop]
            - ***frie radikaler*** 'free radicals' [stof-i-krop]
            - ***hormon*** 'hormone' [stof-i-krop] …
                - ***Leptin*** 'leptine' [hormon]
    - **Substance_Food** [Substance|Food$_{\text{Telic}}$]
        - ***TILSKUDSSTOF*** 'supplementary stuff' [Substance_Food]
            - ***kosttilskud*** 'dietary supplements' [tilskudsstof]
            - ***lægemiddel*** 'remedies' [tilskudsstof]
                - ***Naturlægemiddel*** 'alternative medicine' [lægemiddel]
                - ***Cyanocobalamin*** 'cyanonobalamin' [lægemiddel]
            - ***appetitnedsættendem iddel*** 'appetite surpres. drug' [tilskudsstof]

Figure 15. OntoQuery nutrition ontology.

Thus, all types under Substance_Food are unified types together with three nodes under Natural Substance (*C-vitamin*, *E-vitamin* and *betacaroten*), which have multiple coordinates since they are both sub-types of vitamin and antioxidant (in contrast to e.g. *D-vitamin*). All other nutrition concepts are simple types.

https://doi.org/10.1017/S0332586504001179 Published online by Cambridge University Press

Approximately 1,000 nutrition terms have been detected in the texts from the encyclopaedia. These have been coded according to the ontology given in Figure 15 and the general SIMPLE guidelines.

3.2 Ontology-based querying

Searching in OntoQuery is performed by comparing the semantic description derived from the user query with the semantic descriptions extracted by the system off-line from the texts in a text database. As mentioned earlier, such semantic descriptions rely on the use of the ontology.

To produce semantic descriptions, a number of NLP techniques are applied. Queries and texts are tokenised and part-of-speech tagged; then all NPs are recognised and the words occurring in them are lemmatised and replaced by the corresponding concepts in the ontology.[13] Each NP is represented as a set of concepts, and the relations between them are left undefined. For example, the semantic description of the NP *depoter af vitaminer* 'vitamin deposits' is the following set of concepts:

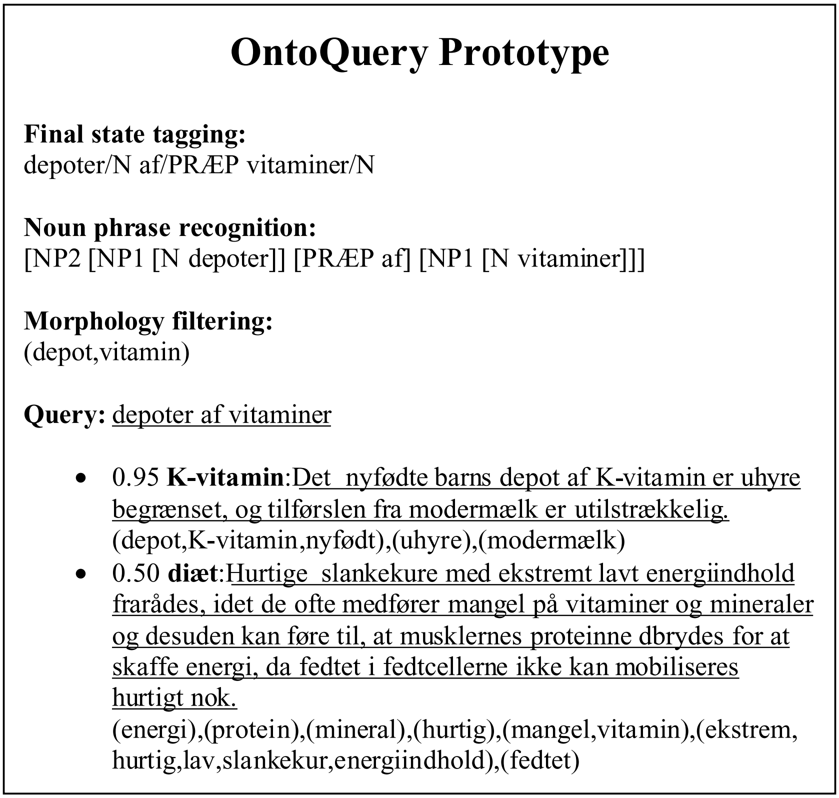
(12) (*depot vitamin*)

The analysis and querying results of this simple example are shown in Figure 16. After the POS-tagger has assigned part-of-speech tags to the words, the NP recogniser analyses the whole query as one single NP including the PP post-modifier; then the two nouns are lemmatised and the corresponding concepts found and collected into a set. This set is then matched against the semantic descriptions corresponding to the texts in the database, and the texts are retrieved and ranked according to how close these descriptions match the input description. In the figure, only excerpts from the two highest ranking texts are shown.

The similarity between the two representations to be matched is computed, based on the distance between different concepts in the ontology as proposed in Andreasen (2001), so that a query concept X matches another concept, Y, by 1 − d (X,Y)/10, where 'd' stands for distance. Thus, a concept in the query matches the same concept in a text by 1.0, an immediate sub-concept by 0.9, a second-level sub-concept by 0.8, etc. In the example under consideration, in the highest ranking paragraph, *depot* matches *depot* by 1.0, and *K-vitamin* matches *vitamin* by 0.9. The degree to which the set (*depot vitamin*) representing the content of the query matches the content of the paragraph is aggregated by simple average to yield 0.95. In the next best hit, only the concept *vitamin* matches a concept in the set (*mangel vitamin* 'lack vitamin'). This yields an aggregated average of 0.5, which is the score assigned to the text as a whole.

The interesting thing is that none of the texts in the database offers an exact match with the content words in the query. However, the first text, which deals with newly

https://doi.org/10.1017/S0332586504001179 Published online by Cambridge University Press



OntoQuery Prototype

Final state tagging:
depoter/N af/PRÆP vitaminer/N

Noun phrase recognition:
[NP2 [NP1 [N depoter]] [PRÆP af] [NP1 [N vitaminer]]]

Morphology filtering:
(depot,vitamin)

Query: depoter af vitaminer

- 0.95 **K-vitamin**:Det nyfødte barns depot af K-vitamin er uhyre begrænset, og tilførslen fra modermælk er utilstrækkelig.
(depot,K-vitamin,nyfødt),(uhyre),(modermælk)
- 0.50 **diæt**:Hurtige slankekure med ekstremt lavt energiindhold frarådes, idet de ofte medfører mangel på vitaminer og mineraler og desuden kan føre til, at musklernes proteinne dbrydes for at skaffe energi, da fedtet i fedtcellerne ikke kan mobiliseres hurtigt nok.
(energi),(protein),(mineral),(hurtig),(mangel,vitamin),(ekstrem, hurtig,lav,slankekur,energiindhold),(fedtet)

Figure 16. Query analysis in the OntoQuery Prototype.

fed babies and their deposits of vitamin K is intuitively relevant to the topic ‘deposits of vitamin’, and more so than the second text, which is about vitamin deficiency.

For further exemplification of the performance of the system, consider, in Figure 17, a number of queries, the corresponding sets of derived concepts and the texts retrieved by the system.

All of the four queries yield very similar results as regards the way in which the retrieved texts pattern together. Only in one case (query 1) does the system retrieve a text with a score of 1.00, and in all four cases very few (1 or 2) of the retrieved texts have a score between 0.90 and 0.95. In contrast, a relatively large number of the retrieved texts have a score of 0.50 and below, and are clearly much less relevant to the query. Thus, ontological similarity combined with the structure provided by the NPs constitutes useful additional knowledge to distinguish between relevant and less relevant texts in cases where there is no exact matching between query and texts.

https://doi.org/10.1017/S0332586504001179 Published online by Cambridge University Press

Query	**Sets of derived concepts**	**Matching concepts in text**		**Score**	**No of texts retrieved**
		Danish	English translation		
1. *Hvad har sygdomme at gøre med vitaminer?* 'What have diseases got to do with vitamins?'	(*sygdom*), (*vitamin*) (disease), (vitamin)	(*sygdom*), (*vitamin*)	(disease), (vitamin)	1.00	2
		(*sygdom*), (*thiamin*)	(disease), (thiamine)	0.95	1
		(*anæmi*), (*vitamin*)	(anemia), (vitamin)	0.95	1
		(*mangelsygdom*), (*vitamin*)	(deficiency disease), (vitamin)	0.95	2
		(*infektion*), (*B-vitamin*)	(infection), (vitamin B)	0.90	1
		(*vitamin*)	(vitamin)	0.50	22
		(*sygdom*)	(disease)	0.50	23
		(*følgesygdom*)	(complication)	0.45	5
		...		...	
2. *Hvordan relaterer hormonforstyrrelser sig til andre sygdomme?* 'How do hormone disturbances relate to other diseases?'	(*hormonforstyrrelse*), (*sygdom*) (hormone disturbance), (disease)	(*hormonforstyrrelse*), (*kræft*)	(hormone disturbance), (cancer)	0.95	1
		(*sygdom*)	(disease)	0.50	23
		(*mangelsygdom*)	(deficiency disease)	0.45	8
		...		...	
3. *Er der b-vitaminer i kornprodukter?* 'Is there vitamin B in corn products?'	(*b-vitamin*), (*kornprodukt*) (vitamin B), (corn product)	(*niacin*) (*kornprodukt*)	(niacin) (corn product)	0.90	1
		(*b-vitamin*)	(vitamin B)	0.50	7
		(*kornprodukt*)	(corn product)	0.50	5
		(*thiamin*)	(thiamine)	0.45	1
		...		...	
4. *sygdomme der følger af ensidig kost og har at gøre med tryptofan* 'diseases following from an unbalanced diet and related to tryptophan'	(*sygdom*), (*ensidig, kost*), (*tryptofan*) (disease), (unbalanced, diet), (tryptophan)	(*pellagra*), (*ensidig kost*), (*tryptofan*)	(pellagra), (unbalanced diet), (tryptofan)	0.93	1
		(*sygdom*), (*ensidig kost*)	(disease), (unbalanced diet)	0.67	1
		(*sygdom*), (*kost*)	(disease), (diet)	0.50	4
		(*kost*), (*tryptofan*)	(diet), (tryptophan)	0.50	2
		(*infektion*), (*kost*)	(infection), (diet)	0.47	1

Figure 17. Sample querying results.

https://doi.org/10.1017/S0332586504001179 Published online by Cambridge University Press

3.3 Discussion and future work

In the examples discussed so far, the ontology is used by the search engine to retrieve texts containing more specific sub-concepts than those in the query in cases where no exact match is found. Similar attempts have been made since the beginnings of automatic document retrieval without, however, showing an improvement in the results. It is noted already in Salton (1968) that the use of a thesaurus causes an accuracy loss. More recent experiments, where Wordnet (Voorhees 1994, Smeaton & Quigley 1996) and EuroWordNet (Gonzales et al. 1998) are used as semantic sources, show either loss of accuracy or no improvement compared to simpler methods. It is thus pointed out by Allan (2000), that query expansion using semantic relations does not produce more accurate results, and that given the costs associated with the construction of semantic repositories to be used for query expansion, purely statistical methods are more convenient. More recently, however, de Loupy & El-Bèze (2002) report more encouraging results from TREC-6. They maintain that accuracy can be increased if query expansion is used in combination with semantic sense disambiguation and by means of very specialised thesauri. They also point out that the properties of the semantic source used to enrich a query also have an effect on the final results. In particular, they mention (as is also pointed out by others) that WordNet has a number of drawbacks: (i) no semantic links between words belonging to different word classes; (ii) no link between different words of the same domain; and (iii) very fine-grained senses.

Therefore it is worth investigating, as OntoQuery is doing, whether better results can be achieved by using a semantic lexicon based on formal criteria for establishing sense distinctions. For the Danish SIMPLE lexicon, this formal approach amounts to the fact that if two senses of a word cannot be formally distinguished in the SIMPLE model, then they are merged into one sense. The sense-encoding strategy results in a much more coarse-grained lexicon than what is found in WordNet,[14] for instance, and will therefore yield different search results. However, the main explanation for the promising results obtained with the OntoQuery prototype are due to the fact that downward expansion within restricted domains is a much more reliable task than within general vocabulary. In this respect, the nutrition domain is a particularly 'well-suited' one since it has a long tradition of taxonomic structuring.

Moreover, it is the aim of the project to differentiate itself from the experiments quoted above also by going beyond the approach discussed so far, and to this end, to take advantage of the other types of semantic knowledge encoded in the Danish SIMPLE lexicon. Whilst the semantic description of NPs used in the prototype is a set of unordered concepts, a more complete rendering of the ontological content of NPs must also take the relations that hold between the concepts into account. Within such an approach, a semantic description is defined as an algebraic term, or ontotype, associated with a node in a lattice of concepts (Nilsson 2001b). For example, the

https://doi.org/10.1017/S0332586504001179 Published online by Cambridge University Press

ontotype corresponding to the NP *fedtdepoter hos børn* ‘fat deposits in children’ is the complex concept resulting from the combination (expressed by the meet operator x) of the atomic concept *depot* (deposit) with the two concepts *fedt* ‘fat’ and *barn* ‘child’ by means of the relations *contains* (CON) and *located in* (LOC).

(13) (*depot* x (CON: *fedt*) x (LOC: *barn*))

The two relation-concept pairs (CON: *fedt*) and (LOC: *barn*) are valid restrictions of the concept *depot*, and the resulting complex concept can thus be regarded as a sub-type of it.

Another example, shown below, is the complex concept expressing the meaning of the NP *mangel på D-vitamin om vinteren* ‘lack of vitamin D in winter’, where *på D-vitamin* adds a generic WRT (*with respect to*) relation to the concept of *mangel*, whereas *om vinteren* adds a temporal (TMP) specification to it.

(14) (*mangel* x (WRT: *D-vitamin*) x (TMP: *vinter*))

Such complex concepts can be generated by a conceptual grammar defining valid combinations of concepts. Examples of such general combinations of a concept with semantic relations valid in the nutrition domain are:

(15) (*substance* x (LOC: *concrete-entity*)
x (CBY: *event*)
x (SRC: *concrete-entity*)
x (POF: *concrete-entity*))

In addition to LOC, the semantic relations used are CBY for *caused by*, SRC for *source*, and POF for *part of*. In contrast, the combination (*substance* x (TMP: *time*)) is not allowed by the ontological grammar (see Nilsson 2001b for a discussion).

A conceptual grammar for the nutrition domain has been defined based on the same Encyclopaedia texts that have been used to derive the domain ontology, and is being implemented and tested in the LKB system (Copestake 1999).[15] The interesting issue from the point of view of the present paper is how the conceptual grammar can build on the semantic information provided by the SIMPLE lexicon, particularly the qualia structure. In fact, two different information types are relevant. First of all, traditional semantic selectional restrictions as also encoded in SIMPLE are part of the conceptual grammar, where they provide the correct detection and semantic interpretation of valency-bound complements. More interestingly, the qualia roles provide additional information which can be used especially for concepts that do not correspond to a verbal event from which selectional restrictions can be derived. One such example is *depot* ‘deposit’. From its qualia structure we can derive the valid semantic relations the concept can be associated with. Thus, the PP *af vitaminer* in the query shown in Figure 16 fills out the *contains* relation (CON in OntoQuery) in the CONSTITUTIVE qualia role of *depot*. Likewise in *fedtdepoter*, also discussed

https://doi.org/10.1017/S0332586504001179 Published online by Cambridge University Press

previously, the same role is filled out by the first noun component of the compound. All the examples in (16) display the same basic semantic structure.

(16) (*container* x (CON: *substance*))

a. depoter af vitaminer
deposits of vitamins

b. vitamindepoter
vitamindeposits

c. fedtdepoter
fatdeposits

d. depoter af K-vitamin
deposits of vitamin K

The task of defining a conceptual grammar of the domain building on the semantic constraints expressed by the qualia roles is by no means a trivial task. Indeed, it constitutes an extremely interesting test bed for the SIMPLE model. It remains to be seen, of course, to what extent a more complex weighted matching procedure that takes into account the semantic relations yielded by such conceptual generalisations improves querying.

4. CONCLUSION

The establishment of semantic lexicons for Language Technology purposes is an extremely costly project even if existing traditional lexicons are used as a starting point. The SIMPLE model constitutes a sound and well-tested basis for the development of such a lexicon – providing, as it does, a multifunctional focus on lexical information. In this paper, we have presented the Danish SIMPLE lexicon: we have dealt with some of the most interesting and novel aspects of the SIMPLE model and showed how they fit the Danish data, as well as discussed some of its more problematic aspects. As we have seen, Danish phrasal verbs in particular constitute a challenge because of the role the adverbial particle plays in their semantics – a fact that the theory of events proposed in SIMPLE does not account for. By revising the theory on this point, however, we have shown that the semantics of phrasal verbs can indeed be represented in the SIMPLE framework. This is done by considering the adverbial particle as incorporated into the verb and by applying complex types to capture the relation of regular polysemy that holds between verb pairs belonging to the ‘Act’ class (in particular motion verbs) and the ‘Change’ class (in particular ‘Change Location’ verbs).

We believe that SIMPLE constitutes a valuable platform for future semantic lexicon projects. Further funding has in fact been obtained from the Danish Ministry of IT and Research to scale-up the Danish PAROLE and SIMPLE lexicons to produce a large dictionary suitable for large scale applications (see www.cst.dk/sto/uk).

https://doi.org/10.1017/S0332586504001179 Published online by Cambridge University Press

In spite of the fact that the coverage of SIMPLE-DK is not very large, the lexicon is already being used in the Danish research project on content-based querying OntoQuery. Thus, in the final part of the paper, we have explained how the project has built an ontology based on SIMPLE-DK and extended it with a domain-specific sub-ontology, and how the searching algorithm uses the ontology to guide the matching procedure. The methodology developed in OntoQuery does not take advantage of the whole SIMPLE model in that only the ontological backbone provided by the *is_a* relation is used. However, the project is investigating how to include the other types of semantic information coded in SIMPLE – selectional restrictions and qualia structure – in a conceptual grammar of the domain. In conclusion, so far SIMPLE has proven a flexible and very rich lexical and ontological source.

ACKNOWLEDGEMENTS

The work presented in this paper builds on work carried out by a large group of researchers. In particular, the Danish SIMPLE lexicon has been developed at CST (Center for Sprogteknologi, København) by the following main staff members: Sanni Nimb, Sussi Olsen, Anna Braasch, Britt Keson, and Bolette S. Pedersen. The OntoQuery prototype has been developed at the University of Roskilde under the leadership of Troels Andreasen, whereas the nutrition ontology has been developed mainly by Nikolaj Oldager at the Danish Technical University. The language technology components of the OntoQuery prototype have been developed at CST by Dorte Haltrup and the present authors, with contributions from Hanne Thomsen and Bodil Nistrup Madsen from the Copenhagen Business School. Finally, the formal logic used by the search engine is the work of Jørgen Fischer Nilsson from the Danish Technical University, whereas the ideas concerning a conceptual grammar have been developed by Jørgen Fischer Nilsson in cooperation with Per Anker Jensen, Southern Danish University, and Carl Vikner, Copenhagen Business School.

NOTES

1. To be precise, the Danish SIMPLE lexicon currently contains 8,000 noun concepts (typically concrete and abstract concepts), 2,000 verb concepts (typically events) and 1,000 adjective concepts (typically properties). The lexicon is being augmented in the ongoing STO project (http://www.cst.dk/sto/index.html).
2. Cf. Pedersen & Keson (1999:50f.) for an account of the treatment of Danish regular polysemy in SIMPLE, partly based on the extensive studies by Malmgren (1988).
3. We ignore for the moment the general problem of deciding when to encode a compound in the lexicon and when to leave it for production rules in the grammar. In SIMPLE, frequent compounds are generally encoded as lemmas.
4. In theory, additional and more detailed information could be included in the constitutive quale, e.g. that winter is generally cold and is associated with snow, as long as an adequate

https://doi.org/10.1017/S0332586504001179 Published online by Cambridge University Press

semantic type or semantic feature can be found to express it. For the practical purpose of harmonisation between lexicons, however, the SIMPLE specifications currently include a fixed set of approximately 60 semantic relations and 70 semantic features (Lenci et al. 2000b) to choose from.

5. Examples of abstract concepts encoded to each sub-node are for Cognitive Fact: *viden* ‘knowledge’, Convention: *lov* ‘law’, Time: *juleaften* ‘Christmas eve’, Domain: *medicin* ‘medicine’, Institution: *skole* ‘school’, Moral Standards: *frihed* ‘freedom’, Move of Thought: *kommunisme* ‘communism’.

6. The percentages are based on 100 corpus examples for each noun.

7. We refer to Nedergaard-Thomsen & Herslund (2002) for a study of complex predicates in Danish including also, in addition to phrasal verbs, constructions like *læse aˈvis* lit. ‘read paper’and *træffe beˈslutning* lit. ‘make decision’. A characteristic feature of complex predicates in Danish is that the verb loses its stress, in contrast to examples like *ˈlæse aˈvisen* ‘read the paper’ and *ˈtræffe en beˈslutning* ‘make a decision’.

8. Examples of event concepts encoded under each sub-node are: Phenomenon: *influenza* ‘influenza’, State: *blive* ‘stay’, Psychological Event: *tænke* ‘think’, Aspectual: *begynde* ‘begin’, Act: *spise, løbe* ‘eat, run’, Change: *aftage* ‘decrease’, Cause Change: *bringe* ‘bring’.

9. We must note here, however, that frequency also plays a role in the construction of most modern lexicons; thus very frequent phrasal verbs do tend to figure as sub-lemmas even if their meaning is predictable.

10. Note that at lower representation levels in the database, DIR is ‘unfolded’ into the accepted directional prepositions and particles. This ensures correct mapping of non-compositional phrasal verbs to semantics.

11. We do leave (at least) one problem unsolved in this approach, namely that not all directional prepositional phrases result in loss of stress on the motion verb; thus, we distinguish in Danish between *han gik til staˈtionen* ‘he went to the station’ and *han ˈgik til stationen* ‘he walked to the station’. The latter example, which is NOT a case of incorporation, is in our approach not distinguished from the former.

12. Figure 15 shows only the individual top-concepts of the nutrition ontology. Seen as a whole, the ontology includes also a so-called conceptual grammar, which defines valid combinations of concepts, as reported in section 3.3 (cf. Nilsson 2001a, b).

13. In Paggio, Pedersen & Haltrup (2003) more details on each individual analysis step are provided.

14. See also (Roventini, Ulivieri & Calzolari 2002) for a comparison of ItalWordNet and SIMPLE- IT.

15. A discussion of how the conceptual grammar can be expressed in the typed feature structure formalism supported by LKB can be found in (Paggio 2001).

REFERENCES

Allan, James. 2000. Natural Language Processing for Information Retrieval. Tutorial notes presented at the NAACL/ANLP Language Technology Joint Conference, Washington.

Alonge, Antonietta, Nicoletta Calzolari, Piek Vossen, Laura Bloksma, Irene Castellon, Amria Antonia Marti & Wim Peters. 1998. The Linguistic Design of the EuroWordNet Database. In Piek Vossen (ed.), *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Press, 91–115.

Andreasen, Troels. 2001. Query evaluation through concept descriptions. In Hanne Thomsen (ed.), *OntoQuery Workshop Proceedings, LAMBDA*. Copenhagen Business School, 29–35.

Andreasen, Troels, Jørgen Fischer Nilsson & Hanne Thomsen. 2000. Ontology-based Querying. In Henrik Legind Larsen et al. (eds.), *Flexible Query Answering Systems: Recent Advances*. Berlin: Springer/Physica-Verlag, 15–26.

Andreasen, Troels, Per Anker Jensen, Jørgen Fischer Nilsson, Patrizia Paggio, Bolette Sandford Pedersen & Hanne Thomsen. 2004. Content-based text querying with ontological descriptors. *Database and Knowledge Engineering Journal* **48**, 199–219.

Antoni-Lay, Marie-Helene, Gil Francopoulo & Laurence Zaysser. 1994. A generic model for reusable lexicons: the GENELEX project. *Literary and Linguistic Computing* **9.1**, 47–54. Stanford:

Braasch, Anna & Bolette Sandford Pedersen. 2002. Recent work in the Danish Computational Lexicon Project STO. In *EURALEX Proceedings 2002*. Copenhagen: Center for Sprogteknologi, 301–315.

Copestake, Ann. 1999. *The (new) {LKB} system – version 5.2*. Stanford: CSLI.

Cruse, David Alan. 1991. *Lexical Semantics*. Cambridge: Cambridge University Press.

de Loupy, Christophe & Marc El-Bèze. 2002. Managing synonymy and polysemy in a document retrieval system using WordNet. In *Proceedings from the Workshop on Using Semantics for Information Retrieval and Filtering*. Las Palmas de Gran Canaria: LREC, 20–27.

Durst-Andersen, Per & Michael Herslund. 1996. The syntax of Danish verbs: lexical and syntactic transitivity. In Engberg-Pedersen et al. (eds.), 65–102.

Engberg-Pedersen, Elisabeth, Michael Fortescue, Peter Harder, Lars Heltoft & Lisbeth Falster Jakobsen (eds.). 1996. *Content, Expression and Structure: Studies in Danish Functional Grammar*. Amsterdam: John Benjamins.

Fellbaum, Christiane (ed.). 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Gonzales Julio, Fernando Verdejo, Carol Peters & Nicoletta Calzolari. 1998. Applying EuroWordNet to cross-lingual text retrieval. *Computers and the Humanities* **31**, 185–207.

Guarino, Nicola & Christopher Welty. 2000. Ontological analysis of taxonomic relationships. In Alberto Laender & Veda Storey (eds.), *Proceedings of ER-2000: The International Conference of Conceptual Modeling*. Berlin: Springer Verlag, 210–224. [Available at http://citeseer.nj.nec.com/correct/309633]

Harder, Peter, Lars Heltoft & Ole Nedergaard-Thomsen. 1996. Danish directional adverbs, content syntax and complex predicates: a case for host and co-predicates. In Engberg-Pedersen et al. (eds.), 159–198.

Herslund, Michael. 1993. Transitivity and the Danish verbs. *LAMBDA* **18**, 41–62, Copenhagen Business School.

Jensen, Per Anker & Peter Skadhauge (eds.). 2001. *Ontology-based Interpretations of Noun Phrases: Proceedings from the First International OntoQuery Workshop*. Kolding: University of Southern Denmark.

https://doi.org/10.1017/S0332586504001179 Published online by Cambridge University Press

Lenci, Alessandro, Nuria Bel, Federica Busa, Nicoletta Calzolari, Elisabetta Gola, Monica Monachini, Antoine Ogonowski, Ivonne Peters, Wim Peters, Nilda Ruimy, Marta Villegas & Antonio Zampolli. 2000a. SIMPLE: a general framework for the development of multilingual lexicons. *International Journal of Lexicography* **13**, 249–263.

Lenci, Alessandro, Federica Busa, Nilda Ruimy, Elisabetta Gola, Monica Monachini, Nicoletta Calzolari, Antonio Zampolli, James Pustejovsky, Emilie Guimier, Lee Humphreys, Ursula Von Rekovsky, Antoine Ogonowsky, Clair McCauley, Wim Peters, Ivonne Peters, Rob Gaizauskas, Marta Villegas & Ole Norling-Christensen. 2000b. *SIMPLE Linguistic Specifications*. SIMPLE report, ms., University of Pisa.

Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago & London: The University of Chicago Press.

Malmgren, Sven-Göran. 1988. On Regular Polysemy in Swedish. In *Studies in Computer-Aided Lexicography*. Stockholm: Almquist & Wiksell, 179–200.

Nedergaard-Thomsen, Ole & Michael Herslund. 2002. Complex predicates and incorporation: a functional perspective. Ole Nedergaard-Thomsen & Michael Herslund (eds.), *Travaux du Cercle Linguistique de Copenhague* **XXXII**, 7–47, Copenhagen: C. A. Reitzel.

Nilsson, Jørgen Fischer. 2001a. Generative ontologies, ontological types, and conceptual grammar. In Hanne Thomsen (ed.), *Proceedings from OntoQuery Workshop on Ontologies and Search, LAMBDA*. Copenhagen Business School, 45–53.

Nilsson, Jørgen Fischer. 2001b. A logico-algebraic framework for ontologies. In Jensen & Skadhauge (eds.), 11–43.

Nimb, Sanni & Bolette S. Pedersen. 2000. Treating metaphorical senses in a Danish computational lexicon: different cases of regular polysemy. In *Proceedings from The Ninth Euralex International Congress*. Universität Stuttgart, 679–691.

Paggio, Patrizia & Bjarne Ørsnes. 1993. Automatic translation of nominal compounds: a case study of Danish and Italian. *Rivista di Linguistica* **5.1**, 129–156. Torino: Rosenberg & Sellier.

Paggio, Patrizia. 2001. Parsing in OntoQuery: experiments with LKB. In Jensen & Skadhauge (eds.), 89–102.

Paggio, Patrizia, Bolette Pedersen, Dorte Haltrup. 2003. Applying language technology to ontology-based querying: The OntoQuery Project. *Applied Artificial Intelligence Journal* **17, 8–9**: *Artificial Intelligence for Cultural Heritage and Digital Libraries*, 817–833.

Pedersen, Bolette Sandford. 1999. Systematic verb polysemy in MT: a study of Danish motion verbs with comparisons to Spanish. In Harold Somers (ed.), *Machine Translation* (vol. 14). Dordrecht: Kluwer Academic Press, 35–82.

Pedersen, Bolette Sandford & Britt Keson. 1999. SIMPLE – Semantic Information for Multifunctional Plurilingual Lexicons: some examples of Danish concrete nouns. In *SIGLEX 99: Standardising Lexical Resources*. ACL Workshop, University of Maryland, 46–51.

Pustejovsky, James 1995. *The Generative Lexicon*, Cambridge, MA. MIT Press.

Roventini, Adriana, Marisa Ulivieri & Nicoletta Calzolari. 2002. Integrating two semantic lexicons, SIMPLE and ItalWordNet: what can we gain? In *The Third International Conference on Language Resources & Evaluation*. Las Palmas, Gran Canaria, 1473–1477.

Ruimy, Nilda, Ornella Corazzari, Elisabetta Gola, Antonietta Spanu, Nicoletta Calzolari & Antonio Zampolli. 1998. The European LE-PAROLE project: The Italian Syntactic Lexicon. In *First International Conference on Language Resources & Evaluation*. Granada, 241–249.

https://doi.org/10.1017/S0332586504001179 Published online by Cambridge University Press

Salton, Gerard. 1968. *Automatic Information Organization and Retrieval*. New York: McGraw-Hill.

Scheuer, Jan. 1995. *Tryk på Danske Verber* [Stress on Danish verbs] (RASK Supplement **4**). Odense Universitetesforlag, Odense.

Slobin, Dan. 1996. Two ways to travel: verbs of motion in English and Spanish. Masayoshi Shibatani & Sandra A. Thompson (eds.), *Essays in Semantics*. Oxford: Oxford University Press, 195–217.

Smeaton, Alan & Ian Quigley. 1996. Experiments on using semantic distances between words in Image Caption Retrieval. In *Proceedings of the 19th International Conference on Research and Development in IR*. Zurich, 174–180.

Talmy, Leonard. 1985. Lexicalisation patterns: semantic structures in lexical forms. In Timothy Shopen (ed.), *Grammatical Categories and the Lexicon* (vol. 3). Chicago: Press Syndicate of the University of Chicago, 57–149.

Voorhees, Ellen. 1993. Using WordNet to disambiguate word senses for text retrieval. In Robert Korfhage, Edie Rasmussen & Peter Willett (eds.), *Proceedings of the 16th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, Pittsburgh, 171–180.

Voorhees, Ellen. 1994. Query expansion using lexical-semantic relations. In William Bruce Croft & Keith van Rijsbergen (eds.), *Proceedings of the 17th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, New York: Springer Verlag, 61–69.

Vossen, Piek (ed.). 1999. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers.

Ørsnes, Bjarne. 1995. *The Derivation and Compounding of Complex Event Nominals in Modern Danish: An HPSG Approach with an Implementation in Prolog*. Ph.D. dissertation, University of Copenhagen.

https://doi.org/10.1017/S0332586504001179 Published online by Cambridge University Press