Recent Work in the Danish Computational Lexicon Project "STO"

Anna Braasch & Bolette S. Pedersen
Center for Sprogteknologi
Njalsgade 80
DK-2300 Copenhagen S
Denmark
anna@cst.dk bolette@cst.dk

Abstract

This paper describes a national co-operational project, STO, which has the aim of developing a large-scale Danish lexical database for computational use. We discuss some organisational aspects of the project and present the current project structure and main activities. Further, we discuss in more detail some of the linguistic issues that have required thorough consideration before a large-scale encoding could be initiated, encompassing topics such as the morphological encoding of compounds and proper names, as well as the syntactic encoding of there-constructions, phrasal verbs and reflexive verbs.

1 From a multilingual project to a national lexical database

1.1 About the STO project

The Centre for Language Technology (CST) is in charge of a national co-operational project with the aim of developing a large-scale Danish lexical database ('lexicon') for computational use, containing at least 50,000 lemmas. The short name of the project is STO, which stands for SprogTeknologisk Ordbase (i.e. Lexical Database for Language Technology). The intended core application area of this lexicon is natural language processing including both language technology applications and computational linguistic research purposes. The project receives funding from the Danish Ministry for Science, Technology and Development for a period of three years (2001-2004). The groundwork for STO was laid within the multilingual LE-PAROLE project (1996-98).

The aim of this project was to elaborate lexicons for 12 languages of approximately equal size (20,000 entry words), sharing linguistic specifications, descriptive model and information structure. The Danish lexicon was produced by CST in collaboration with The Danish Society for Language and Literature.

With respect to further developments, the initiation of new, national and co-operative projects was anticipated which would reuse and extend the material elaborated. Obviously, the STO project can be seen and understood within this context. To our knowledge, there are a few further similar national projects of various sizes running (e.g. in Sweden and Italy).

1.2 Project objectives

The objective of the STO project is the development of a computational lexicon of Danish for a broad practical application area. This lexicon will serve as a basic lexical data collection from which various dedicated lexicons can be derived. The project is mainly concerned with the formalised representation of existing linguistic and lexicographic knowledge for computational use, e.g. the treatment of inflectional morphology, noun compounding and valency (syntactic frames) of verbs. In the course of the design and development of the lexicon we also do pioneer work on some general linguistic tasks, if there is no systematic and exhaustive linguistic description that could be implemented straightforward. We shall return to this point later on (cf. in Sections 3 and 4).

1.3 Current project structure

CST started the STO project and is now responsible for project management and for co-ordination of work. Further, various central tasks such as software development, elaboration of linguistic specifications, encoding guidelines etc. are carried out at CST too. The co-operating project members, all part timers, are affiliated to three different institutes one being located at the University of Copenhagen, another at the Business School of Copenhagen and a third one being a self-supporting institute of University of Southern Denmark. They mainly carry out encoding-related tasks.

1.4 Recent and current main activity areas

The material developed and the experience gathered from the PAROLE project served as a point of departure for the development of the monolingual lexicon. The PAROLE linguistic specifications, descriptive model and information structure were designed to apply to 12 very different languages, such as Danish, Finnish, Greek, Italian, etc. The specifications comprised the basic features of all languages involved, thus being rather broad. In this connection we were faced with two problems to be solved before starting the encoding of the STO lexicon. First, the above specifications were not satisfactorily detailed as regards the specific properties of the particular languages. Second, because of the very broad and comprehensive linguistic specifications, the information structure also was extremely complex and bulky. These problems were dealt with during the first project phase, by tailoring and adapting the linguistic specifications to the monolingually determined requirements. The structure of the lexical database is simplified and reorganised accordingly.

In the ongoing project year, we are focusing on a number of working areas related to the large-scale production of lexicon entries. These areas are:

- General vocabulary: frequency based extension of the coverage
- Domain languages: composition of text corpuses, selection of lemmas, encoding
- Linguistic specifications: refinement, esp. improvement of language-specific features according to application-oriented requirements
- Coding manual for lexicographers: updates dealing with new features and newly developed tools
- Computational tools: further developments as required by lexicographer's work

- Database management, control of workflow, etc.

2 The descriptive model and language

The linguistic information content of the lexicon is organised into three independent but coherently linked levels, i.e. the morphological, the syntactic and the semantic level. Each level is made up of a comprehensive system of the respective linguistic properties. Consequently, the full linguistic description of a lemma is structured in different sets of information i.e. in 'units' dealing accordingly with these three levels. The description model is based on a concept of such units. From the linguistic point of view, a unit represents a particular linguistic behaviour of a lemma at the level concerned, thus the description of the lemma comprises morphological, syntactic and semantic units. The complete linguistic description of a lemma can be constructed automatically by accessing the whole set of units linked to the lemma in question. The construction process can be seen as a progression through the three levels. The figure below illustrates the structured linguistic description of a lemma.

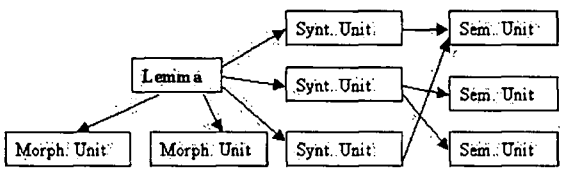

Figure 1: STO Linguistic description model

The above linguistic description method has both advantages and drawbacks from the lexicographer's, and user's points of view. From the user's point of view it is possible to derive customised lexicon material from the lexical data base by using scripts to select the material to be downloaded and format the output. On the other hand, modelling the linguistic information at three independent linguistic levels causes some difficulties, as it will be discussed in Section 4.3.

From the computational point of view a unit is a structured object containing a feature-based description, expressed in attribute/value-pairs. The requirement of explicitness and exhaustiveness is a well-known demand of natural language processing (NLP) as discussed in [Briscoe & Boguraev 1989] and [Van Eynde & Gibbon 2000], this leads to the 'atomisation' of information, i.e. to the division of linguistic description into very fine pieces, i.e. many features. Compared to traditional lexical work, the development of a lexicon for NLP applications is even more time consuming and therefore it is important to develop computational tools supporting the encoding work.

The basic element of description is the 'pattern'; each pattern is a unique combination of attribute/value pairs covering a particular linguistic behaviour type. A morphological pattern

comprises the inflectional properties, which are e.g. for declension of nouns: gender, number, definiteness and case.

In the following sections the paper will focus on a few particular tasks at the morphological, syntactic and semantic levels illustrating the recent working tasks in the project.

3 Particular working tasks at the morphological level

The concept of morphological units and the treatment of their properties are kept compatible - although not identical - with the principles common to all PAROLE lexicons. The current developments of our project are mainly concerned with the refinement and extension of the language-specific descriptions. In this section, we will discuss a few selected developments concerning the morphology of nouns.

3.1 Treatment of compound nouns

In Danish, noun compounding is a very productive word formation process (that is the combination of two or more independently existing words into a new one.) Compounds may be formed in all word classes; we focus in what follows, on compound nouns that make up the largest category. The internal structure of compounds is not fully predictable as regards the syntactic relationship between the components and the method of linking them. For NLP applications it is a necessary condition to treat compounds properly both in recognition and generation. In order to meet this demand, two new features are introduced which extend the linguistic coverage of the lexicon, allowing a dynamic utilisation of the material encoded, e.g. the recognition of newly coined compounds.

First, the treatment of noun compounds - i.e. the decomposition of compound lemmas into their immediate constituents, both independent words and linking element(s) - is focused on. Lexicalised compounds having non-compositional meanings are always encoded as entry words on a par with simplex words. Other compounds being frequent in the corpus are inserted as entry words, but they are treated differently according to their compositionality. The database entry of each compound has a separate field containing the decomposition of the compound into its two immediate constituents (which are the 'first element' and 'second / last element') and a linking element in between. The first element may possibly be a compound itself, and is treated accordingly.

The linking of constituents into a compound and the selection of linking element(s) cannot be described by general rules, but it must be individually recorded in the lexicon. This is illustrated with a few prototypical examples below, where the Danish compound is decomposed into its immediate constituents:

(1) *skolebog* => *skole* + **0** + *bog* (schoolbook)
(2) *forsamlingshus* => *forsamling* + **s** + *hus* (meeting house)
(3) *drengeskole* => *dreng* + **e** + *skole* (boys' school)
(4) *papirsblomst* => *papir* + **s** + *blomst* (paper flower)
(5) *æggebæger* => *æg***[ge]** + *bæger* (egg cup)
(6) *børnefødselsdag* => *b***[ørne]***arn* + *fødselsdag* (child's birthday party)

It is worth noting that the individual linking properties are registered consistently with the remove/add (henceforth R/A) method, which is used to compute the inflected forms of a lemma. The 'removal' part of the rule removes the part of a lemma (marked in square brackets), which has to be changed, leaving the radical for the generation of the inflected form in question. The 'add' part is the ending to be added and is introduced by a '+'. In the examples (5) and (6) the R/A method is used to trace the lemma form of the first constituent, viz. *b***[ørne]***arn* => *'barn'* (child).

Second, the prototypical linking elements are registered for all simplex words as well as for lexicalised compounds - both of which can occur as the first constituent of a multiple compounded lemma. These entry words are provided with a field containing the prototypical linking element(s), which are ordered according to their frequency. Nouns usually have one or two, in a few cases three alternating linking elements like in (7)

(7) *mand* +e, +s, 0, (man) which allows for
(7a) *mandemåned* => *mand* + **e** + *måned* (man-month)
(7b) *mandsperson* => *mand* + **s** + *person* (male individual, lit: man's person)
(7c) *manddag* => *mand* + **0** + *dag* (man-day)

3.2 Treatment of the inflectional morphology of geographical and geo-political proper nouns

Another interesting language-specific task was the description of the morphosyntactic behaviour of proper nouns, especially geographical and geo-political names which are rather frequent in our newspaper corpus comprising 40 M tokens. We focus here on the morphological features of such proper nouns: names of states/countries, cities, mountains, rivers, seas and islands. On the one hand, they refer semantically to a designated entity, which differentiates them from common nouns, the so-called appellatives. On the other hand, they make up a subcategory of nouns having the formal, morphological properties of nouns: gender (common and neuter), number, case and definiteness and also morphosyntactic agreement is required by such proper nouns, although in somewhat different way. In a simplified formulation, the basic morphosyntactic agreement rules of Danish are the following:
(a) In a definite, attributive noun phrase, the word-initial article has to agree in form with the noun wrt gender and number, the attributive adjective is always definite (suffix: *-e*);
(b) An adjective in predicative function or a pronoun referring back to a noun has to agree in form with the noun wrt gender and number. Only the neuter gender of adjectives is marked by a suffix (*–t*); common is unmarked (in the table below: *Ø*).
In grammars, e.g. [Allan et al. 1995], the particular morphosyntactic properties of geographic and geo-political and proper nouns are mentioned in general terms - more or less as exceptions to the general rules; neither do dictionaries give sufficient information in this regard. For NLP applications it is important to describe these properties exhaustively.

According to the basic descriptive method dealing with the inflectional morphology, we develop patterns for all types of inflectional behaviours. This gives rise to the development of inflectional patterns treating proper nouns on a par with appellatives, but extended with

the particular features describing the deviations from general agreement rules. Thus, such patterns have thus to express explicitly the idiosyncratic properties, e.g. the change of form of definiteness (whether the word-final article is mandatory and fixed, or detachable) or the discrepancy of formal and logical number (e.g. formally plural, but semantically/logically singular). In most – although not all – cases, deviations from general morphosyntactic agreement rules arise from the meaning of the lemma: they are semantically determined. For instance, country and region names are semantically singular, neuter (word-initial article: *det*, word-final article *+(e)t*), according to the gender of the superordinate term *land* (country, land). Similarly, river names have common gender (word-initial article: *den*, word-final article *+(e)n*), *flod* (river), etc. More complicated is the case of city names, where the word-initial article in definite noun + adjective phrases is neuter according to *sted* (place), but in predicative and pronominal agreement common gender is used, according to the lemma *by* (city, town).

(8) Lemma: *Filippinerne* (The Philippines);
Formal properties: Plural (*-er*), word-final definite article (*–ne*) is fixed.
Agreement rules:
(a) Singular neuter. Ex.: ***Det** vestlige Filippinerne* ('The western Philippines')
(b) Singular neuter. Ex.: *Filippinerne er smukt. **Det** er også stort.*
('The Philippines is beautiful. It is also big'.)

(9) Lemma: *København* (Copenhagen)
Formal properties: (unmarked for gender, number and definiteness)
Agreement rules:
(a) Singular neuter. Ex.: ***Det** dejlige København* ('The beautiful Copenhagen'), a corresponding indefinite phrase: ***Et** sommerligt København* ('A summery C.')
(b) Singular common. Ex.: *København er smuk. **Den** er også stor.*
('Copenhagen is beautiful. It is also big.')

Lemma	Final def.art	Gender	Number	Init. def.art. + attr.adj.	Pred. adj/ Pron.ref.
Donau	-	(Com.)	Sing.	**Den** brede Donau	Donau/**Den** er bredØ.
Tyskland	-	(Neu.)	Sing.	**Det** rige Tyskland	Tyskland/**Det** er rigt.
København	-	(Neu.)	Sing.	**Det** store København	København/**Den** er storØ.
Rhinen	Detachable	Com.	Sing.	**Den** snavsede RhinØ	Rhinen/**Den** er bredØ.
Elben	Fixed	Com.	Sing.	**Den** brunlige Elben	Elben/**Den** er bredØ.
Arresø	- /+**en**	Com.	Sing.	**Den** varme Arresø	Arresøen/**Den** er varmØ.
Sortehavet	Fixed	Neu.	Sing.	**Det** varme Sortehavet	Sortehavet/**Det** er varmt.
Atlanterhavet	Detachable	Neu.	Sing.	**Det** kolde AtlanterhavØ	Atlanterhavet/**Det** er koldt.
Atlasbjergene	Detachable	Ø	Plur.	**De** høje AtlasbjergeØ	Atlasbjergene/**De** er høje.
Filippinerne	Fixed	Ø	Plur.	**Det** vestlige Filippinerne	Filippinerne/**Det** er stort.
Færøerne (a) [region] (b) ['islands']	Fixed (a) Detachable (b)	Ø	Plur.	**Det** smukke Færøerne (a) **De** 18 FærøerØ (b)	Færøerne/**Det** er stort. (a) Færøerne/**De** er smukke. (b)
Christiansø	-	(Com.)	Sing.	**Det** smukke Christiansø	Christiansø/**Den/Det** er smukØ/t

Table 1: Selected combinations of properties of the lemma forms and agreement rules.

Presently, above 620 frequently used geographical and geo-political nouns are encoded with 15 patterns which cover all different attribute/value combinations observed up to now.

4 Elaborating syntax

4.1 Strategy for syntactic encoding

The method used for describing syntax in STO is consistent with that used for description of inflectional behaviour at the morphology level and again its main characteristics are adapted from PAROLE [Ruimy et al. 1998]. At the syntactic level, a pattern describes a particular syntactic structure compatible with a lemma. Basically, patterns in syntax describe the predictable and systematic syntactic behaviours of lemmas as found in the corpus, comprising primarily features related to subcategorisation properties (valency) and raising/control phenomena. In other words, the number of valency slots and the syntactic function of the arguments subcategorised are encoded; each of them is provided with information about their syntactic realisations i.e. phrase type [Navaretta 1997], [Braasch 1998].

One of the points that we have considered carefully concerns the relationship between linguistic rules and regular patterns on the one hand, and 'exceptions' to these rules on the other hand, to be accounted for in the lexicon. Several linguistic topics - related to the alternation types described for English by [Levin 1993] - call for a decision on this point; for example it had to be decided to which degree constructions with expletives and *der* (there) are lexically determined by the verb in Danish and therefore should be accounted for in the lexical verb entries, or whether these phenomena should only be described in the potential grammar. Since corpus examinations reveal a strong indication of a lexical choice in relation to a number of Danish verb classes, the choice has been made for these two phenomena that they should be encoded in the lexicon as a specific construction type [Braasch et al. 2002]. Thus, intransitive verbs like *løbe* (run), *restere* (remain), and *vaje* (wave) can occur in there-constructions as seen in example 10-12 below, in contrast to intransitive verbs like *vibrere* (vibrate) and *nyse* (sneeze) (example 13-14).

(10) *der løber en hund rundt ude i haven*
(lit: there runs a dog around out in the garden)

(11) *der resterer to sejladser*
(lit: there remain two runs (of ship navigation))

(12) *der vajer et flag på husets top*
(lit: there waves a flag on the top of the house)

(13) **der vibrerer en mur*
(there vibrates a wall)

(14) **der nyser en mand*
(there sneezes a man)

Similarly with transitive and ditransitive verbs, some verbs like *vente* (wait, await) (example 15) can occur with *der*, others like *købe* (buy) cannot (example 16).

(15) *der venter ham en spændende opgave*
(lit: there waits him an exciting task)

(16) **der køber ham en ny bog*
(there buys him a new book)

4.2 Covering varieties of syntactic behaviour

Another important issue concerning the establishment of syntactic patterns relates to the selection strategy adopted for deciding how many valency descriptions should be expressed at this level of representation. In the Danish verbal system, prepositions and adverbial particles express what in many other languages is part of the meaning of the verb. In the STO architecture, this generally has as a consequence that a verb represented at the morphological level splits into several units at the syntactic level. Furthermore, for some of the most frequent verbs, the number of possible distributional patterns is overwhelming and they are virtually semi-productive. This fact calls for a consistent corpus-based strategy with respect to the selection of patterns to be represented. Currently, a formula similar to the one adopted in the Senseval project [Kilgarriff 1998] is applied. The Senseval project is concerned with sense tagging of corpus samples and applies the following formula for the calculation of the number of corpus examples to be considered for each word: For each word (lemma): if it has n senses consider 75 + 15n instances; e.g. consider 120 instances for a 3-sense word. The problematic part in this approach lies in determining a priori how many senses a word has and therefore how much corpus material should be considered. In STO we work from a medium-sized Danish dictionary as our basis (Nudansk Ordbog) but since this dictionary is not fully corpus-based and misses several frequent distributional patterns, we estimate that the number of examples should be further increased. We therefore opt for 100 as the base factor for our formula and apply the formula 100 + 15n instances per lemma. After having undergone an introspective phase where incorrect or very strange patterns have been discarded, all other syntactic patterns that occur in this sample are considered relevant candidates for encoding in STO at the syntactic level.

Following this approach we get the following results for a verb like *røre* (touch, move, stir.). This verb has 8 senses in Nudansk Ordbog and thus requires 220 corpus instances. In these samples we find 12 syntactic patterns for the verb and 3 idiomatic expressions:

nok se, ikke ***røre*** (see but not **touch**)	SUBJ verb
derfor ***rører*** *jeg intet når jeg skal på scenen* (therefore I don't **take** anything (alcohol) when I go on stage)	SUBJ verb OBJ
de skal have handsker på når de ***rører*** *ved træet* (they are to wear gloves when they **touch** the tree)	SUBJ verb PREPOBJ (ved)
pestoen ***røres*** *tynd med noget af vandet fra pastaen* (lit. the pesto is **stirred** thin with some of the water from the pasta)	SUBJ verb OBJ ATROBJ (tynd/tyk)
det værk der ***rører*** *ham dybest* (the work that **touches** him most)	SUBJ verb OBJ MANNERADV (dybt)
en stor træslev der ***rører rundt***.. (lit. a big wooden spoon that **stirs** about)	SUBJ verb-rundt
hvad der ***rører sig*** *i befolkningen* (what **goes on** in the population)	SUBJ verb-sig
det er svært at ***røre ud*** (lit: it is difficult to **stir in**)	SUBJ verb-ud
jævn med kartoffelmel ***rørt ud*** *i koldt vand* (thicken with potato flour **stirred up** into cold water)	SUBJ verb-ud PREPOBJ (i)
æggeblomme ***rørt op*** *med en spsk. vand* (egg yolk **stirred up** with a tablespoon of water)	SUBJ verb-op OBJ PREPOBJ (med)
creme-fraiche ***rørt*** *med dijonsennep* (creme fraiche **stirred** with dijon mustard)	SUBJ verb OBJ PREPOBJ (med)
en ostecreme ***rørt*** *af gorgonzola* (lit: a cheese pasta **stirred** of gorgonzola)	SUBJ verb OBJ PREPOBJ (af)
røre *på sig* (**start moving**)	IDIOMATIC EXPRESSION
uden at ***røre*** *en finger* (without **stirring** a finger)	IDIOMATIC EXPRESSION
rørte *vande* (**troubled** waters)	IDIOMATIC EXPRESSION

Table 2: Syntactic patterns for *røre* (touch, move, stir)

Broadly, none of these syntactic patterns can be considered to be mistakes or aberrations, so all of them are good candidates for a syntactic unit representation (as well as a semantic representation) even though several of them are not mentioned in Nudansk Ordbog ; for instance the reflexive construction exemplified by the corpus excerpt *hvad der rører sig befolkningen* (lit: what is going on in the population), proved to be indeed very frequent in the corpus.

4.3 Particular problem cases: phrasal verbs and reflexive verbs

As already indicated above with the example of *røre*, the Danish verbal system in general challenges a strictly modular representation model as applied in STO. Speaking in Talmy's terms [Talmy 1985], Danish is a typical satellite-framed language, meaning that prepositions and adverbial particles express what in many other languages form part of the meaning of the verb (cf. [Harder, Heltoft & Thomsen 1996], [Durst-Andersen & Herslund 1996], [Herslund 1993] and [Pedersen 1999]). Thus, several of the most frequent verbs in Danish are relatively neutral with respect to syntactic and semantic affiliation as well as regarding event type; their affiliation being determined as much by a particle, a reflexive, or a preposition as by the verb stem itself. In fact, from our corpus examinations we estimate that more than half of the verb senses relevant for STO (relevance is here solely based on frequency) is constituted by phrasal verbs which cannot be uniquely assigned a syntactic or

semantic type on the basis of the verb stem alone (see also [Braasch & Olsen 2000] for a treatment of more complex idioms in STO).

Representing this aspect in a lexicon is a challenge not only for traditional lexicography but even more for computational lexicography, which has a long tradition of a modular composition of the lexicon distinguishing strictly between morphology, syntax and semantics; and which is traditionally centralised around the governing word classes, nouns, adjectives and verbs and the arguments that they take.

Two questions are under consideration in order to propose a proper treatment of Danish phrasal verbs in STO: 1) Are phrasal verbs to be considered a morphological, a syntactic or a semantic phenomenon i.e. at which level of the lexicon model should the phrasal verb be registered and represented as a unit ? 2) How do we represent the syntax and semantics of phrasal verbs ? The compromise that we suggest is a so-called *split late* strategy meaning that phrasal verbs are only represented as such at the semantic level irrespective of whether they are compositional or non-compositional in meaning [Pedersen & Nimb 2000]. For a verb like *gå* (walk) this means that at the syntactic level we only account for the fact that it can be combined with a directional particle or prepositional phrase (abbreviated by 'DIR'). At the semantic level the non-compositional phrasal verbs are identified, such as *gå op* which apart from the literal meaning 'walk upwards' can mean also 'cancel out' (see Figure 2).

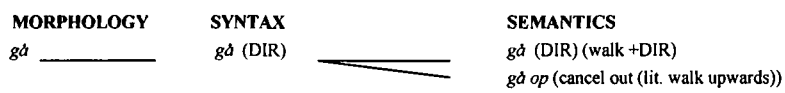

Figure 2: The representation of phrasal verbs

Another aspect concerns the treatment of reflexive constructions in Danish. In STO, true reflexive verbs are defined as verbs which are obligatorily reflexive i.e. *dumme sig* (make a fool of oneself), *korse sig* (cross oneself or be scandalised) and where a fully lexicalised (i.e. non-predictable) interpretation is required at the semantic level since the reflexive pronoun is more or less semantically empty. In several cases, however, true reflexives have a non-reflexive counterpart with a radically different meaning or a completely different syntactic potential: *genere sig* (feel embarrassed) vs. *genere ngn. (med ngt)* (disturb somebody (with something)).

Verbs that are not reflexive, but may potentially combine with a reflexive pronoun *sig*, also require a specific marking, since not all verbs have this possibility, i.e. *myrde* (murder). However, there is a group of verbs which is not easily defined as belonging to one of the two groups, because the meaning difference between the reflexive and the non-reflexive construction is subtle. A further element of complication relates to the fact that some reflexive verbs and some verbs which can occur with a reflexive pronoun, only occur with an emphatic pronoun *selv* (self) added to the reflexive pronoun. Since the use of this emphatic pronoun is obligatory and occurs in both groups of verbs it is necessary to account for this phenomenon in the encodings.

5 Tailoring SIMPLE Semantics for STO

A general design model for the semantic level of STO is provided by the SIMPLE model [Lenci et al. 2000] as effectuated under the SIMPLE project. This model consists of a core ontology with approx. 135 core ontological types allowing for the encoding of a large amount of semantic information such as ontological typing, domain information, qualia structure, semantic relations, argument structure, event structure, selectional restrictions, collocational information, as well as information on polysemy and synonymy. However, in STO, a priority ranking of semantic information types is required in order to tailor the outline in SIMPLE, which is extensively rich in information, to a large-scale project like STO with a relatively lean semantics.

In this context, we have elaborated a priority ranking with three levels of semantic specificity: for a substantial part of the vocabulary – especially the vocabulary for special purposes – only domain information is provided corresponding to a so-called 'one-level' semantics. 'Two-level' semantics is provided for most of the general purpose vocabulary - in particular verbs - comprising sense distinctions, ontological typing and selectional restrictions. 'Three-level' semantics corresponds directly to the full set of information types represented in SIMPLE and relates to what is already encoded for 10,000 word senses in STO (what constituted the original Danish lexicon). In the longer term, the plan is eventually to provide a 'three-level' semantics for a larger part of the STO lexicon depending, however, on the funding situation of the project.

6 Application areas

Since STO is a computational lexicon under development, it has not yet been applied and tested in full-scale applications. However, some ongoing experiments with the application of STO in particular research prototypes can be reported. Two main application types are considered here, namely information retrieval and machine translation.

Since 2000, STO and Danish SIMPLE has been exploited in a Danish research project on content-based information retrieval, called OntoQuery (Ontology-based Querying) (cf. [Andreasen et al. 2000], [Paggio, Pedersen & Haltrup 2001], http//: www.ontoquery.dk). The aim of this project is to investigate the possibilities of a retrieval system that goes beyond superficial key word recognition but on the other hand does not require a *full* semantic analysis of queries and texts. The nutrition domain has been selected as a first test domain, and an ontology for this domain has been established and merged with the ontology applied in STO. Several information types have been taken over directly from STO - such as morphological information (applied in the lemmatising of texts and queries), domain specification, ontological typing and synonymy. In the current version of the OntoQuery Prototype the extended ontology is exploited in a downward weighted expansion of the queries, resulting in a content-based *ranking* of the search results that is based on a calculation of the semantic distance between the concepts in the query and the concepts in the texts. Ongoing research in the project is concerned with the exploitation of the ontology in the analysis phase, leading to a further use of the semantic information encoded in STO, such as qualia structure and selectional restrictions [Pedersen & Paggio 2002].

Another research project, the EU-project TQPRO (Translation Quality for Professionals) which deals with machine translation and other translation tools, has experimented with the use of STO. This experiment has taken place in the machine translation system PaTrans, which is adjusted for automatic translation of patent texts between English and Danish [Bech 1992]. STO is here used in the transfer phase in order facilitate disambiguation of ambiguous words. In the experiment a transformation of data has taken place, since the Patrans formalism (building on the former EUROTRA formalism) differs substantially from the typed feature structure tradition with inheritance hierarchies on which STO is built. In the experiment use is made of so-called preference rules which are added 'on top of' the grammar so to speak, evaluating for each possible analysis the best result from a semantic point of view.

7 Summing up

The STO project is primarily a production project which implements lexicographical knowledge adjusted to potential language technology applications of which we have briefly mentioned two: machine translation and information retrieval. This means that several pragmatic decisions are made in the project and that frequent linguistic phenomena and properties (verified by corpus examinations) which are important to automatic analysis and generation have a high priority in the project as opposed to less frequent (or to NLP less crucial) phenomena.

Nevertheless, the very explicit information types that are required by a lexicon made for computational use, also calls for pioneer work in some fields of Danish lexicography and linguistics. We saw this in the case of geo-political proper names, as well as in the case of verb alternations where it had to be decided whether the alternations in question should be lexicalised or described in general terms by grammar rules.

Acknowledgements

Many thanks go to the members of the STO project group: Sussi Olsen, Sanni Nimb, Nicolai Hartvig Sørensen, Dorthe Haltrup Hansen and Anna Sophie Liebst for their valuable contributions to the work preceeding this article. Furthermore, the STO project gratefully acknowledges the support given by colleagues at The Danish Dictionary (DDO) project.

References

[Allen et al. 1995] Allan, R., P. Holmes, T. Lundskær-Nielsen, 1995. Danish: *A Comprehensive Grammar.* Routhledge, London and New York.

[Andreasen et al. 2000] T. Andreasen, J. Fischer Nilsson, & H. Erdman Thomsen: Ontology-based Querying, in H.L. Larsem et al. (eds.) *Flexible Query Answering Systems, Flexible Query Answering Systems, Recent Advances*, Physica-Verlag, Springer, 2000. pp. 15-26.

[Bech 1992] Bech, A., 1992. 'PaTrans-projektet', in: *Skriften på Skærmen no. 6, Datalingvistisk Forenings Årsmøde nr. 2 pp.107-115,* Handelshøjskolen i Århus, Denmark.

[Braasch 1998] Braasch, A., 1998. 'Corpus-Supported Modelling of Syntactic Information on Nouns in the Danish PAROLE Lexicon', in: *Conference Abstracts from ALLC/ACH 1998* pp.22-26, Lajos Kossuth University, Hungary.

[Braasch & Olsen 2000] Braasch, A., S. Olsen, 2000. 'Formalised Representation of Collocations in a Danish Computational Lexicon',in: U. Heid & al., (eds.) *Proceedings of the Ninth EURALEX Congress* p.475-488. Stuttgart, Germany.

[Braasch et al. 2002] Braasch, A., C. Navaretta, S. Nimb, S. Olsen, B.S. Pedersen, N. Sørensen 2002, *SprogTeknologisk Ordbog – Lingvistiske Specifikationer*, Technical Report version 5, Center for Sprogteknologi, Denmark.

[Boguraev & Briscoe 1989] Boguraev, B. & T. Briscoe, 1989. *Computational Lexicography for Natural Language Processing.* Longman, London and New York.

[Durst-Andersen & Herslund 1996] Durst-Andersen, P. & M. Herslund, 1996. 'The syntax of Danish verbs: Lexical and syntactic transitivity', in: E. Engberg-Pedersen et al. (eds.) *Content, Expression and Structure. Studies in Danish Functional Grammar.* John Benjamins, Amsterdam.

[Harder, Heltoft, & Thomsen 1996] Harder, P., L. Heltoft & O.N.Thomsen, 1996.'Danish directional adverbs, content syntax and complex predicates: A case for host and co-predicates', in: E. Engberg-Pedersen et al. (eds.) *Content, Expression and Structure. Studies in Danish Functional Grammar.* John Benjamins, Amsterdam.

[Herslund 1993] Herslund, M., 1993. 'Transitivity and the Danish Verbs', in *LAMBDA no. 18,* Copenhagen Business School, Copenhagen.

[Kilgarrif 1998] Kilgarrif, A. 1998. SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation Programs. In *Proceedings from First International Conference on Language Resources and Evaluation pp.581-588*, Granada, Spanien.

[Lenci et al. 2000] Lenci, A., N. Bel, F. Busa, N. Calzolari, E. Gola, M. Monachini, A. Ogonoski, I. Peters. W. Peters, N. Ruimy, M. Villegas, A. Zampolli. 2000.'SIMPLE – A General Framework for the Development of Multilingual Lexicons', in: T. Fontenelle (ed.) *International Journal of Lexicography Vol 13.* pp. 249-263. Oxford University Press.

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

[Navaretta 1997] Navarretta, C., 1998. 'Encoding Danish Verbs in the PAROLE Model'. In: R. Mitkov, N. Nicolov & N. Nikolov (eds*) Recent Advances in Natural Language Processing*, Tzigov Chark, Bulgaria, 1997, p. 359-363

[Paggio, Pedersen & Haltrup 2001] Paggio, P; Pedersen, B.S.; and Haltrup, D., 2001. 'Applying Language Technology to Content-based Querying, The Ontoquery Project', in *Proceedings from Workshop on Artificial Intelligence for Cultural Heritage and Digital Libraries* pp. 75-79. Università di Bari, Italy.

[Pedersen 1999] Pedersen, B.S., 1999.'Systematic Verb Polysemy in MT: A Study of Danish Motion Verbs with Comparisons to Spanish', in H. Somers (ed.): *Machine Translation Vol.14, Iss. 1 p 39-86.*, Kluwer Academic Publishers, Dordrecht.

[Pedersen & Nimb 2000] B.S. Pedersen & S. Nimb (2000) 'Semantic Encoding of Danish Verbs in SIMPLE - Adapting a verb-framed model to a satellite-framed language'. *Proceeding from Second International Conference on Language Resources and Evaluation, LREC 2000*, Athens.

[Pedersen & Paggio 2002] Bolette S. Pedersen & Patrizia Paggio, 2002. Semantic Lexical Resources Applied to Content-based Querying - the OntoQuery Project, in *Third International Conference on Language Resource and Evaluation 2002*, Las Palmas, Gran Canaria.

[Ruimy et al. 1998] Ruimy, N., O. Corazzari, E. Gola, A. Spanu, N. Calzolari, A. Zampolli, 1998. 'The European LE-PAROLE Project: The Italian Syntactic Lexicon', in: *First International Conference on Language Resources & Evaluation*, Granada, Spain.

[Talmy 1985] Talmy, L., 1985. 'Lexicalisation Patterns: Semantic Structures in Lexical Forms', in T. Shopen (ed.) *Grammatical Categories and the Lexicon, Vol. 3*, Press Syndicate of the University of Chicago, Chicago.

[Van Eynde & Gibbon 2000] Van Eynde, F. & D. Gibbon (eds.), 2000. Lexicon Development for Speech and Language Processing. Kluwer Academic Publishers, Dordrecht/Boston/London.