SIMPLE LE4-8346

DANISH SIMPLE - LEXICON DOCUMENTATION

* * *

Document first version date			
Document date	21/06/2000		
Document ID	Danish Simple-Lexicon Documentation		
Version			
Doc. type			
Document status	***final***		
Validation type			
Comments			
	Name	***Organisation***	***Purpose***
From	Bolette Pedersen Sanni Nimb Sussi Olsen	COP	Documentation
To			

1. General design information

The aim of the EU-project SIMPLE (Semantic Information for Multifunctional Plurilingual Lexica) is to provide harmonised semantic lexicons for Natural Language Processing for 12 of the European languages. The project is an extension of the LE-PAROLE lexicons, which contain 20,000 entries with corresponding morphological and syntactic information for each of the 12 languages that participated in the project (cf. Ruimy *et al*, 1998).

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 experiences in preceding lexical projects such as Genelex, WordNet and EuroWordNet.

The SIMPLE project started in April 1998 and was completed in April 2000.

1.1. Lexicon population

The Danish SIMPLE-lexicon adds semantic descriptions to 8,200 of the 20,000 Danish PAROLE lexicon entries. These 8,200 morphological entries amounts to 10,000 semantic units because of cases of polysemy and homonomy. 7,000 of the semantic units are nouns; 2,000 are verbs, and 1,000 are adjectives.

The entries to be encoded in SIMPLE have been chosen on the basis of three different criteria:

- the Base Concepts have been selected for encoding when equivalents were present in the Danish PAROLE lexicon (cf. Lenci *et al.* 2000);
- words have been chosen that could illustrate the different ontological types if the SIMPLE model, i.e. almost all ontological types are represented;
- concrete nouns have had higher priority than abstract and event nouns although all three types are fairly represented in the final lexicon

In the case of nouns, we have sought towards a relatively 'closed approach' to lexicon population so that all relevant readings of the particular words were encoded. We have primarily based our reading distinction strategy on a medium-sized monolingual lexicon as well as on corpus examinations (i.e. in some cases we have deviated from the lexicon because the corpus revealed either less or other ambiguities than the ones represented in the lexicon).

In the case of verbs, a closed approach has not been plausible first of all because the Danish PAROLE lexicon has not adopted such an approach when describing the syntax of Danish verbs. For instance, Danish is characterised by a *very* high use of phrasal verb constructions (see also Section 2.7) and not all of these have been encoded in syntax.

In relation to lexicon population it is important for us to stress that the elaboration of a Danish computational lexicon does not stop with the PAROLE/SIMPLE project. An ongoing project at Center for Sprogteknologi is concerned with the task of scaling up the PAROLE/SIMPLE lexicon to 100,000 semantic units (see Braasch *et al.* 1998). In particular wrt. phrasal verbs our aim is to extent the existing phrasal verb descriptions into something that corresponds better to the presence of phrasal verbs in Danish corpora.

1.2. Background resources

Two background resources have played an important role in the building of the Danish SIMPLE data, namely corpora and a medium-sized Danish lexicon. First of all, the decision was made very early in the project that all data should be described on the basis of corpus examinations and that each semantic unit should be supported by an illustrative example from the corpus. This means that if a meaning of a word shows significant frequency in corpus we represent it in the SIMPLE lexicon - even if the particular meaning is not represented in the traditional dictionary we use as our other important background resource (for instance the metaphorical meaning of *puslespil* (puzzle)). Also, if a meaning is represented in the lexicon but with no occurrences in the corpus, the particular meaning has in most cases been omitted.

Our corpus examinations are primarily based on two corpora. The most important is the Berlingske corpus of about 20 mill. tokens, consisting of newspaper articles concerning various topics. In the cases where there are few or no examples of a given word in this corpus, the DK-korpus (Bergenholtz 1990), a balanced corpus of 4 mill. words composed of novels, newspapers, journals, magazines and miscellaneous, is used. We have chosen the corpus tool XKWIC (Christ 1993) for our corpus examinations. XKWIC is part of the IMS corpus toolbox developed at the University of Stuttgart and available on the Internet[1].

Nudansk Ordbog is a medium-sized Danish lexicon with a rather consistent reading distinction policy. We have achieved the right to exploit this resource as long as the material is not used with commercial perspectives[2]. Almost all definitions have been extracted from an electronic version of this source. All encoded words in our lexicon include a definition; in cases where we did not find an appropriate definition in Nudansk Ordbog - either because the word was not represented or because the definition for some reason or other was inappropriate - we have elaborated one. It has been of great help to have this resource as a reference point.

1.3 Introduction to information types in SIMPLE

One of the fundamental assumptions behind the SIMPLE model is that word senses differ in terms of their internal complexity and that this complexity can be described on the basis of an ontology established along different dimensions (cf. Lenci *et al.* 2000). Some word senses can be described by means of *simple* types, which means that they inherit their information from only one mother node in the ontology; others are more complex and thus inherit information from several mother nodes following the principle of orthogonal inheritance[i]. These types are called unified types. The multiple dimensions of meaning are represented in SIMPLE by means of an extended qualia structure model based on (Pustejovsky 1995) encompassing a set of semantic relations such as *is_a*, *used_for*, *part_of*, *has_as_parts*, *is_the_result_of* etc. for each qualia). Furthermore, regular polysemous classes are represented in SIMPLE via the additional type *complex* which establishes a link between systematically related senses.

First, as an illustration of a noun of the type 'unified', consider the four meaning components of the concrete sense of the Danish noun *puslespil* (puzzle):

[1] Http://www.ims.uni-stuttgart.de

[2] Rights have been achieved through Christian Becker, Politikens Forlag A/S, Copenhagen.

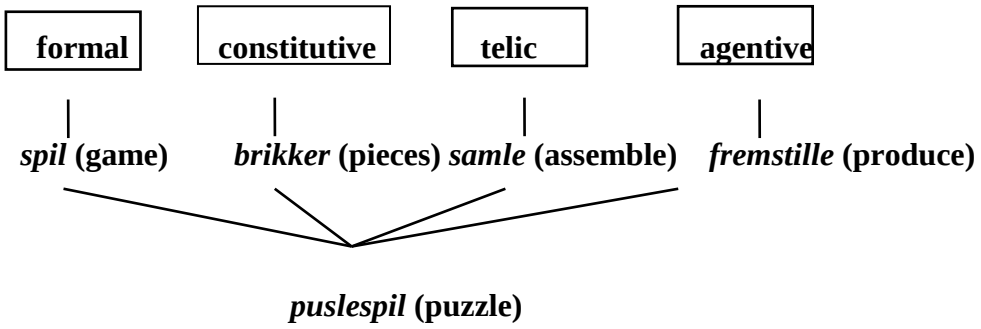


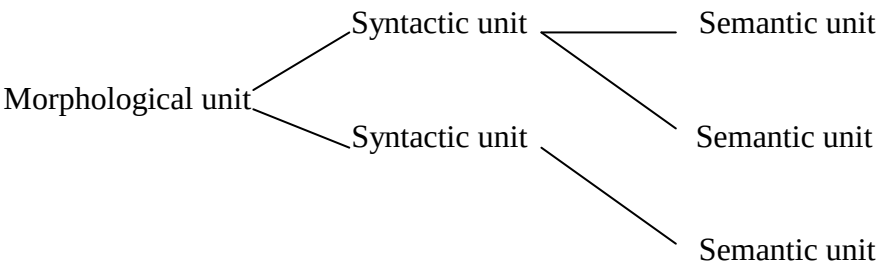
Figure 1. The meaning components of the noun *puslespil* (puzzle)

Four components are involved: (i) the formal role which provides information that distinguishes an entity within a larger set (in this case *is_a*), (ii) the constitutive role which expresses a variety of relations concerning the internal constitution of an entity (in this case *has_as_parts*), (iii) the telic role which concerns the typical function of an entity (here *used_for*), and (iv) the agentive role which concerns the origin of an entity (in this case *made_by*). These elements, plus a long list of additional information types such as definition, domain, corpus example, polysemy relations etc. are represented in the lexical entry, see below:

Semantic Unit	*Puslespil_ART* (puzzle – artifact reading)
Definition:	*et spil med træ- el. papbrikker i forskellige faconer som skal lægges sammen så de danner et hele* (NDO)[ii] (a game with wood or cardboard pieces in different shapes which must be assembled so that they make a whole)
Corpus example:	*nu var hun næsten ved at være færdig med det puslespil, hun var begyndt på lige efter påske* (now she had almost finished the puzzle she had started right after Easter)
Ontological type:	Artifact
Unification Path	Concrete_Entity|Agentive|Telic
Domain	General
Formal quale:	is_a = *spil* (game)
Agentive quale:	Created_by = *fremstille* (produce)
Telic quale:	used_for = *samle* (assemble)
Constitutive quale:	has_as_parts=*brikker* (pieces)
Complex:	ArtifactAbstract_entity= *puslespil_ABS* (puzzle – abstract reading)

1.4. Example of a full lexical entry – from morphology to semantics

One of the basic features of the PAROLE/SIMPLE model is its *modularity* with respect to morphological, syntactic and semantic information as illustrated below:



This division into layers with particular units connected to each implies that there exists no such thing as a *lexical* unit in the traditional lexicographical sense; in contrast each level of representation is described independently although coherently connected the one to the other. In this way, the model permits to distinguish different syntactic behaviours on pure syntactic criteria, and

independently of whether they share the same meaning or not. Furthermore, it permits the refinement of one level (i.e. syntax and semantics) without changing the description of others. *Genericity* and *explicitness* are two of the central requisites aimed at by choosing this architecture.

The easiest way to 'follow' a word from morphology to semantics in the sgml objects is to simply search on the word form throughout the lexicon file. For a verb like *læse* (study, read) this gives the following results (note that since the original Danish PAROLE lexicon covers 20,000 morphological units and around 60,000 syntactic units not all links are necessarily encoded in the semantic part of the lexicon which only covers 10,000 semantic units):

MORPHOLOGY

```
<MuS  (morphological unit)
                                id="UM029573"
                                naming="LÆSE"
                                gramcat="VERB"
                                gramsubcat="MAIN"
                                synulist="Usyn12 Usyn3796 Usyn3797 Usyn3798
                                Usyn3800 Usyn3801 Usyn3802 Usyn3803">
                                <Gmu
                                        attestation="RO86"
                                        inp="MFG0131">
                                        <Spelling>læse</Spelling></Gmu></MuS>
```

SYNTAX

```
<SynU  (syntactic unit)
                                id="Usyn3797"
                                naming="læse"
                                attestation="cn"
                                description="Dv2P-i"><correspSynUSemU
                                   targetsemu="USEM_V_læse_COE_1"
                                   correspondence="arg12i"</SynU>


<SynU
                                id="Usyn12"
                                naming="læse"
                                attestation="cn"
                                description="Dv2N0"><correspSynUSemU
                                   targetsemu="USEM_V_læse_COE_1"
                                   correspondence="arg12"><correspSynUSemU
                                   targetsemu="USEM_V_læse_COE_3"
                                   correspondence="arg12"></SynU>

<SynU
                                id="Usyn3800"
                                naming="læse"
```

```
                    attestation="cn"
                    description="Dv2P-paa"></SynU>

<SynU
                    id="Usyn3801"
                    naming="læse"
                    attestation="cn"
                    description="Dv2P-til"><correspSynUSemU
                       targetsemu="USEM_V_læse_COE_2"
                       correspondence="arg12til"></SynU>

<SynU
                    id="Usyn3802"
                    naming="læse"
                    attestation="cn"
                    description="Dv2xP0-op-til"><correspSynUSemU
                       targetsemu="USEM_V_læse_op_COE_1"
                       correspondence="arg12til"></SynU>

<SynU
                    id="Usyn3796"
                    naming="læse"
                    attestation="cn"
                    description="Dv3N0P0-for">
                    <correspSynUSemU
                       targetsemu="USEM_V_læse_SPE_1"
                       correspondence="arg122P">

<SynU
                    id="Usyn3803"
                    naming="læse"
                    attestation="cn"
                    description="Dv2t"><correspSynUSemU
                       targetsemu="USEM_V_læse_COE_1"
                       correspondence="arg12t"></SynU>

<SynU
                    id="Usyn3798"
                    naming="læse"
                    attestation="cn"
                    description="Dv2xN0-op">
                    <correspSynUSemU
                       targetsemu="USEM_V_læse_op_SPE_1"
                       correspondence="arg12"></SynU>
```

SEMANTICS
COGNITIVE EVENTS

<SemU
id="USEM_V_læse_COE_1"
naming="læse"
example=" Det er ikke en bog , man gider at læse to gange , men sjov er den . "
comment="full BSP"
freedefinition="se på og forstå en tekst (NDONY)" /look at and understand a text/
weightvalsemfeaturel="
WVSFTemplateCognitiveEventPROT
WVSFTemplateSuperTypePsychologicalEventPROT
WVSFEventTypeProcessPROT
TSVP_Cognition_TS_classificateur_de_verbe">
<PredicativeRepresentation
typeoflink="MASTER"
predicate="PREDhumsem_COE_1">
/selectional restrictions ARG1=human ARG2=semiotic /
<RWeightValSemU
weight="PROTOTYPICAL"
comment="Type-defining semantic relation"
target="USEM_N_erkendelsesproces_COE_1"
semr="SRIsa">
</SemU>

<SemU
id="USEM_V_læse_op_COE_1"
naming="læse_op (til)"
example=" På en videregående uddannelse kan man ikke , som på gymnasiet , bare læse op til eksamen "
comment="full BSP"
freedefinition="forberede sig til en eksamen" /prepare an exam/
weightvalsemfeaturel="
WVSFTemplateCognitiveEventPROT
WVSFTemplateSuperTypePsychologicalEventPROT
WVSFEventTypeProcessPROT
TSVP_Cognition_TS_classificateur_de_verbe">
<PredicativeRepresentation
typeoflink="MASTER"
predicate="PREDhum_COE_1">

/selectional restriction ARG1=human ARG2=unrestricted/

<RWeightValSemU
weight="PROTOTYPICAL"
comment="Type-defining semantic relation"
target="USEM_N_erkendelsesproces_COE_1"
semr="SRIsa">
</SemU>

<SemU
id="USEM_V_læse_COE_2"
naming="læse"
example=" En ordentlig arbejder , der ville frem i geledderne måtte helst læse til cand.polit "
comment="full BSP"
freedefinition=" være ved at tage en boglig uddannelse i noget (NDONY)" /take an education to become something/
weightvalsemfeaturel="
WVSFTemplateCognitiveEventPROT
WVSFTemplateSuperTypePsychologicalEventPROT
WVSFEventTypeProcessPROT
TSVP_Cognition_TS_classificateur_de_verbe">
<PredicativeRepresentation
typeoflink="MASTER"
predicate="PREDhumprof_COE_1">
/selectional restriction ARG1=human ARG2=profession/
<RWeightValSemU
weight="PROTOTYPICAL"
comment="Type-defining semantic relation"
target="USEM_N_erkendelsesproces_COE_1"
semr="SRIsa">
</SemU>
<SemU
id="USEM_V_læse_COE_3"
naming="læse"
example="Han trådte som 20-årig ind i redemtoristordenen og læste teologi hos Mauterne i Østrig "
comment="full BSP"
freedefinition=" være ved at tage en boglig uddannelse i noget (NDONY)"
weightvalsemfeaturel="
WVSFTemplateCognitiveEventPROT
WVSFTemplateSuperTypePsychologicalEventPROT
WVSFEventTypeProcessPROT
TSVP_Cognition_TS_classificateur_de_verbe">
<PredicativeRepresentation
typeoflink="MASTER"
predicate="PREDhumdom_COE_1">
/selectional restriction ARG1=human ARG2=domain /
<RWeightValSemU
weight="PROTOTYPICAL"
comment="Type-defining semantic relation"
target="USEM_N_erkendelsesproces_COE_1"
semr="SRIsa">
</SemU>

SPEECH ACTS

```
<SemU
        id="USEM_V_læse_op_SPE_1"
        naming="læse_op"
        example="jeg er heller ikke i stand til at læse op , hvad mine medarbejdere skriver"
        comment="full  SN"
        freedefinition="udtale noget skrevet, så andre kan høre det (NDONY)" /read aloud/
        weightvalsemfeaturel="
        WVSFTemplateSpeechActPROT
        WVSFTemplateSuperTypeActPROT
        WVSFEventTypeProcessPROT
        TSVP_COMMUNICATION_TS_classificateur_de_verbe">
        <PredicativeRepresentation
                typeoflink="MASTER"
                predicate="PRED2hum_sem_SPE_1">
       /selectional restriction ARG1=human ARG2=semiotic /
        <RWeightValSemU
                weight="PROTOTYPICAL"
                comment="Type-defining semantic relation"
                target="USEM_N_talehandling_SPE_1"
                semr="SRIsa">
    </SemU>


<SemU
        id="USEM_V_læse_SPE_1"
        naming="læse"
        example="han læste for pigen "
        comment="full  SN"
        freedefinition="læse højt af en tekst for nogen" /read aloud to somebody /
        weightvalsemfeaturel="
        WVSFTemplateSpeechActPROT
        WVSFTemplateSuperTypeActPROT
        WVSFEventTypeProcessPROT
        TSVP_COMMUNICATION_TS_classificateur_de_verbe">
        <PredicativeRepresentation
                typeoflink="MASTER"
                predicate="PRED3hum_sem_hum_SPE_1">
       /selectional restrictions ARG1=human ARG2=semiotic (can be ommitted)
       ARG3=human/
        <RWeightValSemU
                weight="PROTOTYPICAL"
                comment="Type-defining semantic relation"
                target="USEM_N_talehandling_SPE_1"
                semr="SRIsa">
    </SemU>
```

1.5. Lexicon Contents in the Danish SIMPLE lexicon

In the Danish SIMPLE lexicon we have encoded all required and recommended information and in some cases also optional information (for instance synonymy) for both verbs and nouns (cf. Lenci *et al.* 2000, see also section 3 for the full set of information types encoded in the Danish lexicon). For adjectives a more restricted strategy has been applied; thus for this word class only required information has been encoded; i.e. definition (gloss), template type (i.e. ontological type), semantic class, domain, predicative representation and selectional restrictions.

In the following table the overall statistics for the lexicon is given:

Table 1: Overall statistics

Number of full Semu's linked to syntax and morphology	10,000 semu's
Number of predicative Semu's	2,035
Semu per category Nouns: (required, recommended and optional information) Verbs: (required, recommended and optional information) Adjectives: (required information only)	 7,000 2,000 1,000
Number of dummies	approx. 1000

The following schemas show the amount of senses described under each template in the lexicon.

Table 2: Semu's per Template Type

CONCRETE NOUNS:

Entity	**21**
Part	**59**
Body part	**244**
Group	**19**
Human group	**609**
Concrete entity	**39**
Location	**49**
3D location	**116**
Geopol	**443**
Area	**77**
Openings	**38**
Building	**322**
Artifactual area	**131**
Material	**22**
Artifact	**412**
Artifact material	**52**
Furniture	**69**
Clothing	**137**
Container	**78**
Artwork	**60**
Instrument	**149**
Money	**92**
Vehicle	**131**
Semiotic artifact	**356**
Food	**17**

Artifact food	**129**
Flavouring	**17**
Physical object	**27**
Organic object	**15**
Living entity	**41**
Animal	**7**
Earth	**142**
Air	**77**
Water	**57**
Human	**114**
People	**109**
Role	**0**
Ideo	**53**
Kinship	**51**
Social status	**113**
Agent of temporary activity	**186**
Agent of persistent activity	**115**
Profession	**598**
Vegetal	**7**
Plant	**154**
Flower	**18**
Fruit	**39**
Microorganism	**11**
Substance	**37**
Natural Substance	**37**
Substance food	**52**
Drink	**3**
Artifactual drink	**23**

NON-CONCRETE NOUNS:

Property	**0**
Quality	**69**
Social property	**12**
Psychical property	**4**
Physical property	**25**
Colour	**11**
Physical power	**27**
Shape	**3**
Representation	**7**
Information	**236**
Language	**21**
Number	**34**
Sign	**24**
Unit of measurement	**52**
Abstract	**48**
Cognitive fact	**104**
Convention	**34**
Domain	**51**
Institution	**180**
Moral standards	**4**
Movement of thought	**30**
Time	**93**

EVENTS:

Event	**19**

Weather	**22**
Cause Aspectual	**37**
Aspectual	**25**
State	**42**
Exist	**13**
Relational state	**6**
Identificational state	**20**
Constitutive state	**6**
Stative location	**37**
Stative possession	**11**
Act	**88**
Non-relational act	**130**
Relational act	**250**
Purpose act	**135**
Move	**169**
Caused Motion	**9**
Speech act	**150**
Reporting event	**10**
Commisives	**2**
Cognitive event	**30**
Judgment	**30**
Caused experience event	**65**
Perception	**16**
Change	**10**
Relational change	**2**
Change possession	**4**
Transaction	**16**
Change Location	**51**
Natural transition	**10**
Change of State	**80**
Change of Value	**17**
Acquire knowledge	**42**
Cause Change	**34**
Creation	**13**
Physical creation	**33**
Mental creation	**12**
Symbolic creation	**22**
Copy creation	**1**
Cause relational change	**36**
Cause Change of State	**194**
Cause change of Value	**44**
Cause change of Location	**78**
Cause natural transition	**20**
Disease	**0**
Stimuli	**0**
Cooperative Act	**0**
Cause Act	**0**
Cooperative Speech act	**0**
Directives	**1**
Expressives	**0**
Declaratives	**1**
Psychological event	**2**
Experience Event	**0**
Modal event	**2**
Constitutive change	**0**
Cause constitutive change	**0**

Give knowledge	**0**

PROPERTY TEMPLATES

Modal	**2**
Temporal	**4**
Emotive	**12**
Manner	**7**
Emphasizer	**20**
Physical property	**628**
Psychological property	**155**
Social property	**137**
Temporal property	**12**
Relational property	**20**
Intensional	**3**
Object-related	**0**
Intensifying property	**0**
Extensional	**0**

1.6. Validation

In order to check the grammatical consistency of our encoded SGML templates we have adjusted an SGML parser which validates our files according to the document type definition (dtd).

Apart from the validation taken care of by the SGML parser; we have elaborated a few Unix procedures which help check other sources to mistakes. One procedure checks 'id' and 'naming' and produces a list of semantic units where the two are not identical. Another writes a list of target semu's referred to via the semantic relations in the qualia structure and check these towards the already encoded entries. This list is essentially a list of dummy candidates (i.e. words that have not been fully coded yet and should therefore be established as dummy semu's), but the list is checked manually and wrong references, misspellings, empty targets and other mistakes are sorted out. This can be done only because every 'id' is supplied with an abbreviation of the ontological type to which it belongs (i.e. USEM_V_bevæge_sig_**MOV**_1). Only when a word has more than one sense within the same ontological type the different senses receive subsequent reading numbers (i.e. USEM_N_kort_SEM_1 vs. USEM_N_kort_SEM_2).

As regards purely linguistic consistency checking, a great deal of work is still remaining. Although the lexical guidelines (Lenci *et al.* 2000) have ensured a large degree of consistency between the different parts of the lexicon by providing templates to each ontological type, many cases of inconsistency can still be found. A browser helps us ensure that the use of relations is appropriate; for instance hyponyms and hyperonyms are checked on the lexicon material in order to discover whether a homogenous semantic class refers to the same hypernym or not and whether the hyperonyms of a given hyponym really are hyperonyms at the same level of analysis.

2. Semantic encoding

2.1. Criteria for Syntax-Semantic linking

Non-predicative nouns are linked by simply relating to the semantic unit(s) to which a syntactic unit corresponds; in the case of *adresse*, two links are established from one syntactic unit, namely one to a 'representation' interpretation as in *brevet skal være forsynet med navn og adresse på bagsiden* (the letter should be supplied with name and address on the back) and one to a 'location' interpretation *folk afstår fra at flytte ind på visse adresser* (people desist from moving into to certain addresses):

<SynU
id="Usyn10003"
naming="adresse"
attestation="ns"
description="Dn0">
<CorrespSynUSemU
targetsemu="USEM_N_**adresse_REP**_1">
<CorrespSynUSemU
targetsemu="USEM_N_**adresse_LOC**_1"></SynU>

For predicative nouns and verbs, a more complex linking procedure between syntactic complements and semantic arguments has been established. Here we have followed the LINDA specifications (Underwood *et al.* 2000) where a principled analysis is given of the argument structure of Danish verbs and nouns. For a further description of the linking of predicative nouns and verbs, see Section 2.5.

2.2. Criteria for assigning Domain Features

Most of the vocabulary for this deliverable belongs to the domain: *General.* Specific readings belonging to particular domains have been assigned an appropriate domain from the domain list. Wrt. to domain assignment we have to a large degree followed the encodings made in Nudansk Ordbog. See Section 3 for the statistics for Domain.

2.3. Criteria for assigning Semantic Class and Template Type

Semantic Class and Template Types have been assigned according to the guidelines given by the Specification Group. In most cases, the templates are so well-defined in the guidelines that it has been more or less unproblematic to assign templates to the words. In some cases, however, the features proposed in the templates have been too specific as to count for all the words that would naturally fit into the template. This is in particular the case for events. To give an example, the template CHANGE_LOCATION has as a type-defining feature, the event type 'transition'. However, in the Danish lexicon we have encountered several 'change of location' verbs which denote processes rather than transitions such as *falde* (fall) and *dale* (descend) where the result phase is not expressed implicitly. One could argue that such verbs should therefore rather be encoded under the template MOVE. But the 'change of location' feature seems to be so essential for these two verbs that it doesn't seem convenient to encode them as 'manner of motion' verbs either.

Also for the group of abstract nouns we have sometimes found it difficult to assign templates to the words. Somehow too many words did not seem to fit into the seven more specific abstract template types and therefore simply had to be assigned the mother node "abstract entity". In this template group we therefore find very different words like *alibi* (alibi), *fødekæde* (food chain) and *harmoni* (harmony), which do not share much meaning content. We also found it difficult to distinguish between the groups "Moral Standards" and "Cognitive Fact", for instance in the case of the word *holdning* (attitude), which on the one hand just means a way of thinking about something, but on the other hand could be considered a question of moral. In the template group "Cognitive Fact" we have encoded words of "thinking": *tanke* (thought), *viden* (knowledge), but also words of feeling: *jalousi* (jealousy), *henrykkelse* (delight) etc., though one could discuss whether these words of 'feeling' are events more than cognitive facts.

2.3.1.Template subtyping for language specific encoding

The very large amount of semantic units represented under the ontological type ARTIFACT gives an indication of the fact that this category may require further splitting. We have felt the need for an additional subtemplate denoting electronic or mechanical devices

The interesting thing about electronic and mechanical devices is that they expose a different distribution than other artifacts in the sense that they can 'work by themselves' and thus can often fill in selectional slots which are very similar to human beings. This in particular counts for computers; consider for example the following corpus excerpt:

> *Så spørger computeren om cyklisten holder rigtigt og børnene skal så ved hjælp af musen klikke på enten 'ja' eller 'nej'*
> (then the computer asks whether the biker is in the right place and the kids are then to click on either 'yes' or 'no' with the mouse)

2.3.2. Criteria for encoding Semantic Relations

We have focused on *linguistically relevant* semantic relations. All type-defining, obligatory semantic relations have been encoded. Apart from this some essential relations have been encoded in cases where we believed them to have strong linguistic relevance. In most cases, we have followed the definition given in Nudansk Ordbog. This means that when a feature has been represented as part of the definition for a given word, we have included this feature as a semantic relation in the formal part of the semantic unit.

Consider the relation 'has_as_parts'. This is in many cases a semantic relation which describes what we would call a 'world-knowledge' aspect of a word. For instance, we would not encode a 'has_as_parts'-relation on the noun *hus* (house) since we believe that it is not linguistically crucial for this word that it contains walls, roof, floors, and windows etc.. This hypothesis is supported by the definition in Nudansk Ordbog for the word *hus* : *en bygning som udgør en selvstændig enhed, og som anvendes til beboelse* (a building which constitute an independent unit and which is used for habitation). In contrast, for the noun *trappe* (staircase) the definition does imply a 'has_as_parts'-relation: *et antal sammenhængende trin som man kan gå op el. ned ad* (a number of steps of which you can go up or down); thus this word is encoded with the relation *trappe* 'has_as_parts' *trin*:

<SemU
id="USEM_N_trappe_ART_1"

```
        naming="trappe"
        example=" Ruten i Leeds er uhyggelig hård - indeholder således en lang trappe, der skal forceres med cyklen på ryggen"
        comment="full BSP"
        freedefinition=" et antal sammenhængende trin som man kan gå op el. ned ad (NDO)"
        weightvalsemfeaturel="
                WVSFTemplateArtifactPROT
                WVSFUnificationPathConcreteentity-Agentive-TelicPROT
                TSVP_ARTIFACT_TS_classificateur_de_nom_C">
        <RWeightValSemU
                weight="PROTOTYPICAL"
                comment="Type-defining semantic relation"
                target="USEM_N_genstand_ENT_1"
                semr="SRIsa">
        <RWeightValSemU
                weight="PROTOTYPICAL"
                comment="Type-defining semantic relation"
                target="USEM_V_fremstille_1"
                semr="SRCreatedby">
        <RWeightValSemU
                weight="PROTOTYPICAL"
                comment="Type-defining semantic relation, gå op og ned"
                target="USEM_V_gå_1"
                semr="SRUsedfor">

        <RWeightValSemU
                weight="ESSENTIAL"
                comment="Semantic relation"
                target="USEM_N_trin_ART_1"
                semr="SRHasaspart">
</SemU>
```

A similar situation can be found with many compounds in Danish. Here an essential (non-type-defining) feature can often be used to express exactly the relation that holds between the two parts of the compound; consider for instance the examples below of two kinds of containers in Danish, *vinflaske* (wine bottle) which 'contains *vin*' (wine) and *blikdåse* (tin can) which is 'made of *blik*' (tin)

```
<SemU
        id="USEM_N_vinflaske_CON_1"
        naming="vinflaske"
        example="en vinflaske kan genbruges syv til otte gange"
        comment="full BKK"
        freedefinition="flaske til vin"
        weightvalsemfeaturel="
                WVSFTemplateContainerPROT
                WVSFUnificationPathConcreteentity-ArtifactAgentive-TelicPROT
                TSVP_NOTION_TS_classificateur_de_nom_C">
        <RWeightValSemU
```

```
                weight="PROTOTYPICAL"
                comment="Type-defining semantic relation"
                target="USEM_N_flaske_CON_1"
                semr="SRIsa">
        <RWeightValSemU
                weight="PROTOTYPICAL"
                comment="Type-defining semantic relation"
                target="USEM_V_fremstille_1"
                semr="SRCreatedby">
        <RWeightValSemU
                weight="PROTOTYPICAL"
                comment="Type-defining semantic relation"
                target="USEM_V_indeholde_1"
                semr="SRUsedfor">
        <RWeightValSemU
                weight="ESSENTIAL"
                comment="Semantic relation"
                target="USEM_N_vin_ARD_1"
                semr="SRContains">
</SemU>

<SemU
        id="USEM_N_blikdåse_CON_1"
        naming="blikdåse"
        example="en urtepotteunderskål, hvori man omvendt har sat en tom blikdåse, som fyldes med vand"
        comment="full BKK"
        freedefinition="dåse lavet af blik"
        weightvalsemfeaturel="
                WVSFTemplateContainerPROT
                WVSFUnificationPathConcreteentity-ArtifactAgentive-TelicPROT
                TSVP_NOTION_TS_classificateur_de_nom_C">
        <RWeightValSemU
                weight="PROTOTYPICAL"
                comment="Type-defining semantic relation"
                target="USEM_N_dåse_CON_1"
                semr="SRIsa">
        <RWeightValSemU
                weight="PROTOTYPICAL"
                comment="Type-defining semantic relation"
                target="USEM_V_fremstille_1"
                semr="SRCreatedby">
        <RWeightValSemU
                weight="PROTOTYPICAL"
                comment="Type-defining semantic relation"
                target="USEM_V_indeholde_1"
                semr="SRUsedfor">
```

```
    <RWeightValSemU
                weight="ESSENTIAL"
                comment="Semantic relation"
                target="USEM_N_blik_ARS_1"
                semr="SRMadeof">
</SemU>
```

In general, we have applied a template-driven approach in the sense that each encoder has been responsible for a specific set of templates in order to ensure as large a degree of consistency among encoders as possible as regards the semantic relations to be applied within a template type. For instance, we have striven towards a homogenous level of specificity as well as a consensus on which of the more general Targetsemu's to be applied for each relation.

2.3.3. Criteria for encoding Derivation Relations.

Derivation relations are not encoded in the Danish lexicon.

2.4. Encoding of synonymy and polysemy relations

2.4.1. Synonymy

We have chosen to give information on synonyms in the cases where a synonym is mentioned in the Danish dictionary we use to retrieve our definitions, as long as the synonym is represented in the PAROLE dictionary.
An example, seen below, are the two words *knække* and *brække* (both meaning "cause to break"), encoded in the template group "cause change of state":

```
1) <SemU
      id="USEM_V_brække_CCS_1"
      naming="brække"
      example="Jeg var målløs. Han sparkede på bilen, knuste lygterne og brækkede antennen"
      comment="full BC 200203548 SN"
      freedefinition="få noget til at brække(NDO)"
            weightvalsemfeaturel="
        WVSFTemplateCauseChangeofStatePROT
        WVSFTemplateSuperTypeCauseRelationalChangePROT
        WVSFEventTypeTransitionPROT
        TSVP_CHANGE_TS_classificateur_de_verbe_C">
            <PredicateRepresentation
                        typeoflink="MASTER"
                        predicate="PRED_brække_CCS_1">
            <RWeightValSemU
                        semr="SRAgentiveCause"
                        target="USEM_V_ændre_CCS_1"
                        weight="PROTOTYPICAL">
            <RWeightValSemU
                        semr="SRResultingState"
                        target="USEM_ADJ_itu_QUA_1"
                        weight="PROTOTYPICAL">
```

```
<RWeightValSemU
                weight="ESSENTIAL"
                comment="Synonym relation"
                target="USEM_V_knække_CCS_1"
                semr="SRSynonym">
```

</SemU>

2)

```
<SemU
    id="USEM_V_knække_CCS_1"
    naming="knække"
    example="hvis man knækker skaftet udleveres en ny spade"
    comment="full BC 200203548 SN"
    freedefinition="få noget til at knække (NDO)"
        weightvalsemfeaturel="
      WVSFTemplateCauseChangeofStatePROT
      WVSFTemplateSuperTypeCauseRelationalChangePROT
      WVSFEventTypeTransitionPROT
      TSVP_CHANGE_TS_classificateur_de_verbe_C">
        <PredicateRepresentation
                typeoflink="MASTER"
                predicate="PRED_knække_CCS_1">
        <RWeightValSemU
                semr="SRAgentiveCause"
                target="USEM_V_ændre_CCS_1"
                weight="PROTOTYPICAL">
        <RWeightValSemU
                semr="SRResultingState"
                target="USEM_ADJ_itu_QUA_1"
                weight="PROTOTYPICAL">
        <RWeightValSemU
                weight="ESSENTIAL"
                comment="Synonym relation"
                target="USEM_V_brække_CCS_1"
                semr="SRSynonym">
```

</SemU>

We imagine that links between synonyms in the dictionary could be very useful for many purposes, for instance in applications for information retrieval. It also helps to speed up the encoding process since the entries of two, or sometimes even three, synonymous words can be made easily at the same time.

2.4.2. Polysemy

Regular polysemy - when groups of related words display the same ambiguity - is handled in a uniform way in the SIMPLE model via the identification of a set of well-established regular

semantic classes for nouns, which are adjusted for each of the languages involved. While unsystematic ambiguous readings of a word are represented as totally unrelated semantic units, regular polysemous senses can be encoded as interlinked semantic units. This is represented by the information slot *complex*, whose value is the polysemous class to which the semantic unit belongs as seen below for *Dragør* (Dragør - Danish village) in the semantic unit for the human group sense of the word:

```
<SemU
        id="USEM_N_Dragør_HUG_1"
        naming="Dragør"
        example=" Dragør må i år af med godt 31 mill. kr. til den kommunale udligning”
          /This year Dragør must pay approx. 31 mill. crowns to the  community equalization /
        comment="full BSP"
        freedefinition="de mennesker der bor i Dragør eller som træffer belutningerne der"
        weightvalsemfeaturel="
                WVSFTemplateHumanGroupPROT
                WVSFTemplateSuperTypeGroupPROT
                TSVP_GROUP_NAMES_TS_classificateur_de_nom_C">
        <RWeightValSemU
                weight="PROTOTYPICAL"
                comment="Type-defining semantic relation"
                target="USEM_N_befolkning_HUG_1"
                semr="SRIsa">
        <RWeightValSemU
                weight="PROTOTYPICAL"
                comment="Type-defining semantic relation"
                target="USEM_N_indbygger_HUM_1"
                semr="SRHasasmember">
        <RWeightValSemU
                weight="PROTOTYPICAL"
                target="USEM_N_Dragør_GEO_1"
                semr="SRPolysemyHumanGroup-GeopoliticalLocation">
        </SemU>
```

In the Danish lexicon the most productive cases of regular polysemy involving concrete nouns are the following[3]:

- animal / food
- geographical location / human group
- fruit / plant
- human group / institution / building
- semiotic artifact / information

Other well-known polysemous pairs are not productive in Danish, as for example 'people / language' and 'flower / colour', where only a few examples of each can be found. This difference relates to the distinction made by Apresjan (apud Malmgren, 1988) between productive and regular

[3] see also Boje & Schøsler (ed.) (1992) pp. 11-12, Pedersen & Keson (1999) as well as Braasch & Pedersen (1999) for some considerations of regular polysemy on Danish nouns. Malmgren (1988) contains a more extensive study of regular polysemy in Swedish, a language which displays polysemous behaviour very similar to Danish.

polysemy. Here productive polysemy refers to cases where more or less the whole group of nouns within a semantic class display the same polysemy relations, whereas regular polysemy refers to cases where at least two words - but not the whole class - follow the same polysemy pattern.

A more extensive, empirically-based study of regular semantic polysemous classes of Danish nouns has not yet been carried out. However, the corpus-oriented approach used during the encoding of the Danish SIMPLE lexicon facilitates the identification of new polysemous classes, since the differences in distributional patterns of the encoded words senses are a good indication of whether a regular polysemy relation could be involved. It should be noted, however, that the common polysemy classes established in the project are not totally unproblematic in this respect. One would expect that the classes established would expose different distributional patterns in the corpus; however, this is not always the case. A well-established test for examining such patterns is the so-called zeugma test: two different senses of a word are expected to create a zeugma (i.e. nonsense) if they are put together in the same phrase, as is the case for the regular polysemy class that holds between geopolitical location and human group:

> **Danmark, som er et fladt og grønt land, nedlagde veto mod forslaget i Europakommissionen*
>
> (Denmark, which is a flat and green country, vetoed the proposal in the European Commission)

Nevertheless, for the semiotic artifact/information polysemy relation this is not the case as seen in the example below which clearly combines the two senses in one construction:

> *menukortet, der var dekoreret med en kopi af Arne Haugen Sørensens maleri 'Skovkentaur med dame', var varieret og ganske indbydende*
>
> (the menu, which was decorated with a copy of Arne Haugen Sørensens painting 'Forest centaur with lady', was varied and rather appetising)

This example leads to the discussion of the constraints that should be satisfied in order to establish two semantic units. If they are not distinguished in corpus via different distribution what are the criteria then for defining two senses ? In the particular case of semiotic artifact/information we are tempted to believe that this phenomenon should rather be categorised as a case of *semantic vagueness* than as a case of polysemy since we in a given context can refer to either meaning aspect OR both at the same time.

We have not encoded regular polysemy relations on verbs. It is characteristic for Danish that it has far less cases of regular polysemy for verbs that e.g. English, and we found that it would require a more detailed investigation to decide which of the many classes described in the guidelines would be relevant in the Danish lexicon. However, this work is foreseen in the Danish follow-up lexicon project.

2.5. Predicative nouns and verbs: linking between syntax and semantics and representation of predicative information.

Where non-predicative nouns, as described above in Section 2.1, are linked by simply linking the semantic and the syntactic unit to each other, nouns and verbs which take arguments also has to include information on the relation between the syntactic complements and the semantic arguments.

Apart from this all predicative words contain a predicate object in the semantic unit, in which the selectional restrictions on the arguments is precisely described.

For these predicative nouns and verbs, the linking procedure between syntactic complements and semantic arguments as well as the further representation of predicative information in the lexicon has been established on the basis of the LINDA specifications (Underwood *et al.* 2000) where a principled analysis is given of the argument structure of Danish verbs and nouns.

As regards the semantic arguments, the grammatical subject is in most cases assigned ARG1, the grammatical and prepositional object ARG2 and weakly bound prepositional complements are assigned the function ADJUNCT. In the case of trivalent verbs, the animate second participant (goal/receiver) is assigned ARG2P (as in the case of *drengen* (the boy) in *han gav drengen bogen* (litt. he gave the boy the book (he gave the book to the boy)). In cases where a trivalent verb is having an inanimate second participant (goal/origin/place), as in *han gav plankeværket maling* (he gave the hoarding paint), this participant (in this case the hoarding) is assigned ARG2E.

ARG0 is reserved for semantically empty subjects in the LINDA specifications, as in constructions like *det regner* ("it is raining"), and this kind of argument is not described in a predicate, but only taken care of in the syntactic description (at the syntactic level).

As regards selectional restrictions, we apply ontological types, only. When for an argument we want to express that it can refer human groups only, we simply refer to the ontological type 'human group' via the so-called Informarg objects:

```
<InformArg
        id="ArgHumanGroup"
        comment="human"
        status="CHECK"
        weightvalsemfeaturel="WVSFTemplateHumanGroupPROT">
```

These Informarg objects can contain any combination of selectional restrictions on an argument, e.g. the case where the argument can be both a concrete entity or an event:

```
<InformArg
        id="ArgConcreteEvent"
        comment="direction"
        status="DEFAULTCHECK"
        weightvalsemfeaturel="WVSFTemplateConcreteEntityPROT
WVSFTemplateEventPROT">
```

As regards the semantic roles, these are assigned to each argument according to the list in the guidelines on events. Only we have felt the need to introduce an additional role, "NonProtoAgent", for subjects of the type *flaget vajer* (the flag waves).

2.5.1 Linking of syntax and semantics:

As a an example of the linking procedure of a predicative word consider the noun *magt* (power) (in the sense *have magt over nogen/noget* (have somebody/something in ones power)).

The syntactic description Dn2G-PP-over is linked to the semantic description of the arguments (n_arg12over**)** by means of the feature ‘correspondence’:

<SynU
id="Usynn10407"
naming="magt"
attestation="ns"
description="Dn2G-PP-over">
<CorrespSynUSemU
targetsemu="USEM_N_magt_SOP_1"
correspondence="n_arg12over"></SynU>

The correspondence feature is further specified below where it can be seen how each complement (position) in syntax is linked to an argument in semantics; thus subject is linked to ARG1 and the valency bound prepositional phrase to ARG2:

<Correspondence
id="**n_arg12over**"
naming="mapping from genitive to arg1 and prepositional complement to arg2 with 'over'"
correspargposl="**ARG1_0P_CnGNP ARG2_P_CnPP-over**">

Another example on linking between syntax and semantics is the syntactic unit below for the verb *ride* (ride). Here we can see how the valency pattern Dv2P-paa in syntax is mapped onto the semantic frame arg12paa:

<SynU
id="Usyn4713"
naming="ride"
attestation="n"
description="**Dv2P-paa**"><correspSynUSemU
targetsemu="USEM_V_ride_MOV_1"
correspondence="**arg12paa**"></SynU>

Via the correspondence feature the subject is linked to ARG1 and the valency bound prepositional phrase to ARG2:

<Correspondence
id="**arg12paa**"
naming="mapping for divalent verb with prepositional object"
coresargpos1="**ARG1_P_CNPrsubj ARG2_P_CPP-paa**">

In some cases, more than one description is given in the syntactic unit which makes it necessary to specify which description links to which semantic unit. Below is given the case of *bevæge* (Dv4NPa0Pa0-fra-til) (‘move’ - causative) and *bevæge sig* (Dv4refNPa0Pa0-fra-til) (‘move’ reflexive, decausative). The two descriptions link to the semantic template MOVE and CAUSED MOTION, respectively:

<SynU
id="Usyn3515"
naming="bevæge"
attestation="cn"
description="a0-fra-tDv4NPa0Pil"
descriptionl="Dv3refNPa0Pa0-fra-til">
<correspSynUSemU
targetsemu=USEM_V_bevæge_sig_MOV_1"
correspondence="arg1_ADJ_ADJfratil"
description="Dv3refNPa0Pa0-fra-til">
<correspSynUSemU
targetsemu=USEM_V_bevæge_CAM_1"
correspondence="arg12_ADJ_ADJfratil"
descriptionl="Dv4NPa0Pa0-fra-til">
</SynU>

2.5.2 Predicative Representation

As an example of the predicative description of a word consider the semantic unit of *låse* (to lock) in which the predicative representation is described in the predicate PRED2hum_concrete_CCS_1. We have strived towards a systematic way of naming the predicates (although not all predicates have yet been systematised in this way): '2' means that there are two arguments, the abbreviations 'hum' and 'concrete' label the selectional restrictions which this predicate object describes: the first argument is human, the second is a concrete entity. CCS refers to the name of the template type: Cause Change of State.

<SemU
id="USEM_V_lxaase_CCS_1"
naming="låse"
example="han låste døren"
comment="full SN"
freedefinition="lukke noget med en nøgle så det ikke kan åbnes uden den rette nøgle (NDONY)"
weightvalsemfeaturel="
WVSFTemplateCauseChangeofStatePROT
WVSFTemplateSuperTypeCauseRelationalChangePROT
WVSFEventTypeTransitionPROT
TSVP_CHANGE_TS_classificateur_de_verbe">

<PredicativeRepresentation
typeoflink="MASTER"
predicate="PRED2hum_concrete_CCS_1">
<RWeightValSemU
semr="SRIsa"
target="USEM_V_xaendre_CAC_1"
weight="PROTOTYPICAL">
<RWeightValSemU

```
                    semr="SRAgentiveCause"
                    target="USEM_V_gxoere_REA_1"
                    weight="PROTOTYPICAL">
          <RWeightValSemU
                    semr="SRResultingState"
                    target="USEM_A_lxaast_1"
                    weight="PROTOTYPICAL">

</SemU>
```

In the predicate 'PRED2hum_concrete_CCS_1', a list of argument objects is given:

```
<Predicate    id="PRED2hum_concrete_CCS_1"
                    naming="verber med 2 argumenter"
                    example="nogen aflåser noget"
                    type="LEXICAL"
                    multilingual="NO"
                    argumentl="ARG1PREDhum_CCS_1 ARG2PREDconcrete_CCS_1"
>
```

and finally thise argument objects describe the semantic role and the selectional restriction on each argument:

```
<Argument     id="ARG1PREDhum_CCS_1"
        comment="the first argument of the predicate is human"
        semanticrolel="Role_ProtoAgent"
        informargl="ArgHuman">

<Argument     id="ARG2PREDconcrete_CCS_1"
        comment="the second argument of the predicate"
        semanticrolel="Role_ProtoPatient"
        informargl="ArgConcrete">
```

The many different readings of the word *læse* (to read) given in Section 1.4, also give some illustrative examples on different predicative representations and selectional restrictions.

2.6 Linguistic problems

Phrasal verbs

Phrasal verbs have caused several problems during the encoding phase. Phrasal verbs are *very* frequent in Danish and therefore it is important to strive towards a principled treatment of these.

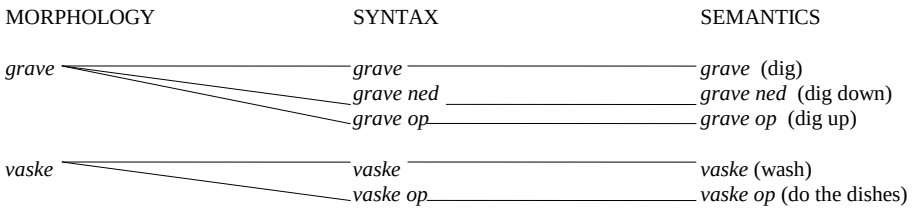
In Danish we can distinguish between three kinds of constructions involving directional particles (cf. Pedersen & Nimb 2000, Scheuer 1995 and Harder, Heltoft & Thomsen 1996):

- simplex verbs combined with adverbial modifiers i.e. *han 'så 'ned på sine tæer* (he looked down at his toes)

- phrasal verbs which are compositional, i.e. predictable in meaning i.e. *han løb 'ud* (he ran out)
- phrasal verbs which are non-compositional, i.e. idiosyncratic i.e *kabalen gik op* (the patience came out)

The second group of verbs is constituted by motion verbs; a unique semantic class in the sense that it admits the directional marker to be understood as *incorporated* in the verb in spite of the fact that the verb itself contributes to the meaning of the expression.

In the Danish Parole syntax such a distinction has not been established mainly due to the fact that the syntax does not really allow for such a distinction: irrespective of the internal nature of the particle construction, the particle is always expressed in the so-called 'self'. This gives an overall splitting strategy as follows:



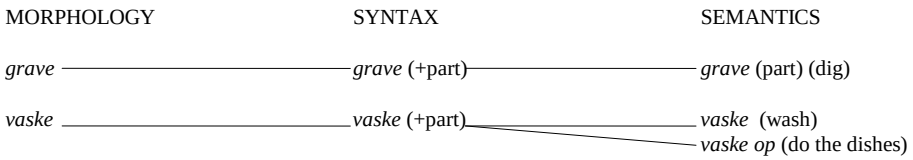
We interpret this as a kind of lexicalisation, having as a consequence that all phrasal verb/constructions in Danish are treated as lexicalisations. This lack of distinction provokes problems when dealing with semantics. As it is now we have been enforced to encode different semantic units to what is basically the same meaning of a word since the particles in such cases are not assigned a valency function but rather are considered as part of the lemma.

Consider the example below for the verb *løbe*. Two syntactic units have been established; the first one describes a construction like *han løb (fra Roskilde) (til København)* (he ran from Roskilde to Copenhagen); the second a construction like *han løb ud* (he ran out). Semantically, we would prefer to treat these as one semantic unit with a directional adjunct which can be expressed either as a PP or as a directional particle. However, as it is now we are enforced to encode - apart from the 'basic' sense of *løbe* - a phrasal verb construction of *løbe ud/ind/op/ned* (run out/on/up/down) which is fully predictable in meaning and which furthermore is considered to take only one argument (since the directional particle is considered to be a lexicalised part of the lexeme *løbe*).

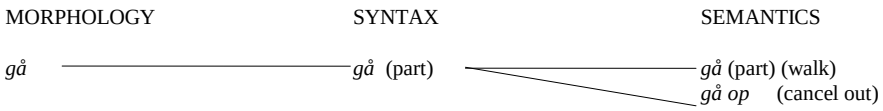
```
<SynU
        id="Usyn2016"
        naming="løbe"
        attestation="cn"
        description="Dv3Pa0Pa0v-fra-til"><correspSynUSemU
          targetsemu=USEM_V_løbe_MOV_1"
          correspondence="arg1_ADJ_ADJ" ></SynU>
<SynU
        id="Usyn5152"
        naming="løbe"
        attestation="cn"
        description="Dv1xdv-dir"><correspSynUSemU
```

targetsemu=USEM_V_løbe_ud_ind_op_ned_MOV_1"
correspondence="**arg1**" ></SynU>

We would have preferred a valency interpretation of all particles at the syntactic level leaving for the semantics to consider whether the meaning was predicable or not. This would also fit nicely into the 'split late' strategy adopted in the project and would leave the semantic distinction where it belongs: in semantics. Consider the figure below where such an approach is adopted for *grave* and *vaske* respectively:



Such a strategy would also be convenient for the really complex cases (which again are rather frequent in Danish) where both a predictable and a non-predictable meaning is found, as for *gå op* which can mean either 'go up' or 'cancel out':



Here the predictable sense (go up) is treated as one semantic unit together with the normal *gå* sense with an optional directional adjunct, whereas the 'cancel out' has its own semu belonging to a different node in the event ontology.

At a longer term, we will reorganise our lexicon wrt. this problems; however, within the scope of SIMPLE, we are not in capable of performing such a large change to the PAROLE lexicon.

Figurative senses

When using a corpus to find the distribution of the different meanings of words being encoded, we have noticed that in many cases the concrete meaning of a word is rarely represented in the text, whereas we often find a high frequency of a figurative sense of the word instead. Often these kinds of meanings are not described in the dictionary we use as our resource, and since these meanings are very abstract, they are quite difficult to place in the semantic hierarchy, at least in the case of abstract nouns.
We have made a small investigation on a group of concrete nouns which produce figurative meanings, namely words belonging to the group of artifacts in the SIMPLE ontology. The figure below shows how often a figurative sense was represented in our newspaper corpus of 20 mill tokens compared to its concrete counterpart:

	Con-crete Arti-facts	**Figurative Sense**	**figurative sense in existing dictionary**	Telic Role of concrete sense
vindue (window)	92 %	8% (15)	no	used_for: *se* (to look)
våben (weapon)	90 %	10 % (100)	no	used_for: *kæmpe* (to fight)

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

As can be seen, several of the figurative senses of these nouns are very frequent even if they are not mentioned in the existing dictionary that we use. In fact, in some cases, *only* the figurative sense is found in the corpus. The last column of the table shows the telic role of the concrete senses. It is remarkable how the verbs or deverbal nouns which are related via the semantic relation *used_for* more or less create the meaning of the figurative sense as can be illustrated by the following corpus examples demonstrating figurative use of the words *skyklapper* (blinkers) and *puslespil* (puzzle):

1) *valutahandlerne har skyklapper på i øjeblikket og vil kun se på de faktorer som vil føre til en styrket dollar*
(the currency brokers are wearing blinkers at the moment and only want to look at the factors which will lead to a strengthened dollar)

2) *det har været et puslespil at få udstillingen på benene*
(it has been a puzzle to arrange the exhibition)

In the first example the 'limiting of visual field' is what creates the new sense of *skyklapper*: the currency dealers sight is limited by their preoccupation with the dollar to such an extent that they cannot see anything else; they are blinded so to speak. In the second example, the metaphor *puslespil* is used to indicate that all the sub-events need to fall into place in order to establish an exhibition, so again the telic quale of parts coming together plays a central role.We would like to let this be reflected in the encoding of the abstract sense in the lexicon.
In our view, the qualia structure with its four meaning dimensions is best suited for the description of concrete nouns. This can also be seen from the fact that there are less type-defining quales to be expressed obligatorily in the abstract part of the ontology. For instance, for the type *abstract* (which we assign to all the senses in the figure above since the subtypes 'domain', 'time' etc. are not relevant) no type-defining quales are predefined apart from the formal role (*is_a*). This, we believe, is not just a particular problem of the SIMPLE model, but rather a general problem relating to the fact that abstract nouns are much more difficult to classify coherently and thus assign type-defining semantic components to. However, in the case of the figurative senses that we are dealing with here - which all originate from concrete artifact senses – we will let the relevant qualia roles be mapped more or less systematically onto the figurative senses; not as type-defining for the entire type 'abstract', but as an essential feature of these particular senses. However, since the *used_for*-relation is too restricted for the abstract senses because it indicates a volitional act with the concrete sense as the object, we suggest to broaden the quale and thus apply the more general semantic relation *object_of_the_activity*. The semantic relation is in any case rather vague and differs slightly depending on whether the figurative sense denotes something negative or something positive; what

seems most important is the information given by the related verb or the deverbal noun. The resulting figurative lexicon entry is shown below for *puslespil*:

Semantic Unit	*Puslespil_ABS* (puzzle – abstract reading)
Definition	*en kompleks sag der består af enkeltdele* (a complex matter which consists of separate parts)
Corpus example	*Det har været et puslespil at få udstillingen på benene* (it has been a puzzle to arrange the exhibition)
Ontological type	Abstract
Unification Path	Entity
Formal quale	is_a = s*ag* (matter)
Telic quale (essential)	object_of_the_activity = *samle* (assemble)
Constitutive quale	has_as_parts=*dele* (parts)
Complex	AbstractArtifact=*puslespil_ART (puzzle – artifact reading)*

As regards verbs, the event ontology seems to cover very well also the figurative senses. We have only been missing one ontological type, namely one to cover the metaphoric event senses 'to move in time' or 'time passing', which we have found were quite common figurative senses of motion verbs in our corpus. One example is with the verb *passere* (pass), which is encoded with the concrete sense 'Change of location', but which also has a figurative sense 'to move in time':

vi skal passere år 2000 , før alle danske biler kører med katalysator
(we will have to pass the year 2000 before all Danish cars run with catalytic converter)

As was the case for nouns, a Semu is established for a verb preferably on the basis of other dictionary sources and corpus examination. Looking at corpus examples of a group of 14 different motion verbs we discovered that many of them, as was the case for nouns, show a high number of figurative senses in real text. The verbs differed from the examined nouns which normally had only one figurative sense, in showing a surprisingly large variation of different metaphoric senses, but we also found that many of them often shared the same figurative sense.
This is for instance the case for the figurative senses 'Change', 'Change of value', 'Act', as well as for the time senses. Some verbs that produce the figurative sense assigned ontological type 'Change' are *bevæge sig* (move), *gå* (walk), *hoppe* (jump) and *springe* (jump). Consider some examples of metaphoric uses of these verbs with the meaning 'Change':

3) *Han sprang fra en akademisk karriere som universitetslærer i psykologi til DRs nye afdeling i Århus*
(lit: He jumped from an academic career as university teacher in psychology to DR's new division in Århus) (He changed from an academic career as a university teacher in psychology to the new division of the Danish Radio in Århus))

4) *Det sociale arbejde har i en årrække bevidst bevæget sig hen imod , at eksakt viden og teknisk kunnen skulle være det dominerende*
(lit. The social work has for a number of years deliberately moved against, that exact knowledge and technical know-how should be the most important (The social work has for a number of years deliberately been changing to a stadium where exact knowledge and technical know-how is the most important))

5) *Jeg tror, det inderst inde er et stort ønske hos mange forretningsfolk, at man gik tilbage til den gamle ordning med lovbestemte datoer*

(lit. I think that deep down it is many business people's big wish to walk back to the old system with dates fixed by law (I think that deep down it is many business people's big wish to change back to the old system with dates fixed by law))

Some verbs that produce the figurative meaning 'Change of value' are e.g. *bevæge sig* (move) *galoppere* (gallop), *hoppe* (jump), *springe* (jump), *kravle* (crawl), *passere* (pass) and *stige* (rise) . Here we see three examples of a metaphoric use of these verbs with 'Change of value' sense :

6) *Gajdars økonomiske reformer har fået priserne til at galoppere i Rusland*
(lit. Gajdar's economic reforms have made the prices gallop in Russia)
(Gajdar's economic reforms have made the prices rise in Russia)

7) *Temperaturen bevægede sig fra lidt over frysepunktet til lidt under*
(lit. The temperature moved from a bit above freezing point to a bit below)
(The temperature changed from a bit above freezing point to a bit below)

8) *Samtidig er også dollarskursen de seneste uger kravlet op ad*
(lit. At the same time also the dollar exchange rate has crawled upwards the last weeks)
(At the same time also the dollar exchange rate has slowly risen the last weeks)

We propose that these cases of systematic derivation of figurative senses should be reflected in the lexical entries as cases of regular polysemy, i.e. as complex types, as in the case of the nouns already described. And at least in some cases it is possible to map information from the qualia structure of the concrete sense into the qualia structure of the figurative sense. Consider for instance the case of motion verbs belonging to the 'Change of location' ontology type and mapping into the figurative sense 'Change of value'. When the direction of the movement is *forward* or *up* (constitutive quale) the constitutive role *resulting state* of the 'Change of value' sense will be *higher* and the direction will be *up*. In contrast, the directions *down* and *backwards* from the motion verb will result in a figurative 'Change of value' sense, where the resulting value is *lower* and the direction of the value change is *down.* See the two entries of the verb *passere* ('Change of location' reading and 'Change of value' reading) for an illustration:

Semantic Unit	*passere_CHL* (pass – Change of location reading)
Definition	*gå, rejse el. på anden måde bevæge sig forbi el. igennem nogen el. noget* (walk, travel or in another manner move by or through somebody or something)
Corpus example	*Hun passerer et hus* (she passes a house)
Ontological Type	Change_of_location
Unification Path	Change/Agentive
Predicative rep	ARG1 ARG2
Selectional Restrictions	ARG1= Living Entity; Vehicle ARG2= Concrete Entity
Event type	Transition
Formal quale	is_a = ændring (change)
Agentive quale	Agentive=flytte_sig (move)
Constitutive quale	Resulting_State = *være* (be) Direction= *forward*
Complex	Change_of_location_Change_of_Value= *passere_CHV*

Semantic Unit	*passere_CHV* (pass – Change of value reading)
Definition	*blive større end* (become bigger than)
Corpus example	*I år ventes på ny rekord , når salget formentlig passerer 300 mia. kroner* (this year a new record is expected, when the sale, as it is supposed, passes 300 bill. kroner)

Ontological Type	Change_of_value
Unification Path	Relational_change/Agentive
Predicative rep	ARG1 ARG2
Selectional Restrictions	ARG1 = Money;Number; Event ARG2 = Number
Event type	Transition
Formal quale	is_a = *ændring* (change)
Agentive quale	agentive = *årsag* (reason)
Constitutive quale	resulting_state = *større* (higher), direction = *up*
Complex	Change_of_Value_Change_of_Location = *passere_CHL*

A sense which was also found several times during our examination of the metaphoric verb senses, is the sense describing moving in time or time passing. Considering the high frequency in the corpus of the verbs in question, these senses are in fact quite common.

It is not easy to find an appropriate ontology type for these senses in the SIMPLE model. Some verbs that often have a metaphorical time sense are e.g.: *gå* (walk), *rende* (run), *sprinte* (sprint), *hoppe* (jump), *springe* (jump), *passere* (pass) and *komme* (come). The following corpus examples show some of these verbs used with a time sense:

9) *mens månederne gik blev hoppen tykkere og tykkere*
(lit. as the months walked the mare became thicker and thicker)
(as the months went by the mare became thicker and thicker)

10) *for hver dag kommer vinteren nærmere*
(each day the winter comes closer)

11) *faktisk mener jeg , at tiden er rendt fra ISAK-messen*
(lit. in fact I think that time has run from the ISAK fair)
(in fact I think that the ISAK fair has had its day)

12) *i næste måned vinker »Pärnu Postimees« endeligt farvel til blyet og hopper flere generationer ind i edb-alderen*
(next month »Pärnu Postimees« (a newspaper) finally waves goodbye to the lead (used for printing) and jumps several generations into the computer age)

13) *vi skal passere år 2000 , før alle danske biler kører med katalysator*
(we will have to pass the year 2000 before all Danish cars run with catalytic converter)

The fact that the SIMPLE model has no appropriate ontological type for these senses might be due to the fact that it is fully conventionalised that we use motion verbs both to describe moving in space and moving in time. Even dictionaries often seem to consider these two senses as one sense. As an example of this, the Danish dictionary 'Nudansk Ordbog' gives the following definition with both a concrete and a figurative example in the entry of the verb *passere* (to pass): *gå, rejse el. på anden måde bevæge sig forbi el. igennem nogen el. noget* (walk, travel or in another manner move by or through somebody or something*). De passerede grænsen. Han har passeret de 70.* (They passed the border. He has passed the age of 70).

Another reason might be that only very few words (at least in the case of Danish) have a time sense only. This makes time senses easy to overlook as a semantic group.Some examples from Danish verbs with a time sense only are *rinde, forløbe* and *hengå* ('elapse', 'slip by', 'pass'*), vare* ('last') and *henslæbe* ('drag on (a miserable existence)'). Other examples where time is an aspect of the meaning of a word are verbs like *feriere* (to holiday), *overnatte* (to stay the night), *tilbringe* (to spend (the night)). But these verbs also have the sense of stative location, and have therefore

without problems been assigned the state ontological type ‘Stative location’ in the Danish SIMPLE lexicon.

We have chosen to assign the time senses the top ontology type ‘Event’ in our lexicon, but in a future extension of the Danish SIMPLE lexicon, we do feel a need for developing the treatment of this kind of figurative senses of verbs.

3. Statistics of information types in SIMPLE-DK

The following tables shows the information types that have been used in the Danish SIMPLE lexicon.

Domains applied in the encodings:

Agriculture
air_transport
arts
astronomy
baby_care
biochemistry
botany
bus_transport
business
car_transport
chemistry
civil_law
commerce
computing
diplomacy
drink
economics
education
entomology
ethnology
fashion
film
finance
fishing
food
freshwater_fishing
furnishing
geography
geology
geometry
gymnastics
health
history
home_and_garden
hotel_business
inland_waterway_transport
law
librarianship
life sciences
linguistics
livestock_farming
logic
mail
mathematics
mechanical_engineering
media
medicine
military
mineralogy
music
ornithology
physical sciences
physics
physiology
poetics
politics
politics and government
psychology
publishing
rail_transport
religion
restauration
road_transport
sailing_yachting_and_boating
sciences

sea_transport
ship_building
sociology
sports and leisure
subway_transport
taxation
transport
trucking
zoology

Semantic Classes applied in the encodings:

ABSTRACT
AGENCY
AMPHIBIAN
ANIMAL
ARTIFACT
ATTRIBUTE
BIO
BIRD
BODY
BODY_PART
BUILDING
CHANGE
COGNITION
COGNITIVE_FACT
COLOR
COMMUNICATION
COMPETITION
CONCRETE
CONSUMPTION
CONTACT
COULEUR
CREATION
CURRENCY
DAY
EMOTION
ETHNOS
FISH
FLOWER
FORM
FRUIT
FURNITURE
GARMENT
GEOG
GEOGRAPHY
GROUP_NAMES
HUMAN
IDEO
INANIMATE
INSECT
INSTRUMENT
LETTER
LIVING_BEING
LOCATION
MAMMAL
MATTER
MEASURE_UNIT
MICROORGANISM
MOLLUSC
MONTH
MOTION
MUSHROOM
NOTION
OBJECT
OCCUPATION
OCCUPATION_AGENT
PART
PERCEPTION
PERIOD
PERIODE
PLANT
POSSESSION
PSYCHOLOGICAL_FEATURE
REPTILE
SHRUB
SOCIAL
STATIVE
SUBSTANCE
TIME_PERIOD
TREE
VEHICLE
WEATHER

List of Polysemy Relations applied:

Agentofpersistentactivity-Profession
Animal-Food
Animal-Material
Area-Humangroup
Area-Institution
Building-HumanGroup
Building-Institution
Container-Amount
Convention-Semioticartifact
Flavouring-Plant
Flower-Colour
Flower-Plant
Food-Animal
Fruit-Plant
GeopoliticalLocation-HumanGroup
HumanGroup-Building
HumanGroup-GeopoliticalLocation
HumanGroup-Institution
Information-Semioticartifact

Institution-Building
Institution-HumanGroup
Language-People
Location-HumanGroup
Material-Animal
Material-Plant
Opening-Artifact
People-Language
Plant-Flavouring
Plant-Flower
Plant-Fruit
Plant-Material
Plant-Substance
Plant-Substancefood
Semioticartifact-Container
Semioticartifact-Information
Substance-Colour
Substance-Plant

List of Semantic Relations applied in the encodings:

Agentive
AgentiveCause
Concerns
Constitutiveactivity
Contains
Createdby
Derivedfrom
Hasascolour
Hasasmember
--Hasaspart
Indirecttelic
Instrument
Isa
Isafollowerof
Isamemberof
Isapartof
Isin
Istheabilityof
Istheactivityof
Isthehabitof
Livesin
Madeof
Measuredby
Objectoftheactivity
Producedby
Produces
Propertyof
Purpose
Quantifies
Relates
Relatedto
ResultingState
Resultof
Successor
Successorof
Synonym
Telic
Usedas
Usedby
Usedfor

4. Concluding remarks

Attempts to harmonise linguistic descriptions of different European languages into a universal model constitutes a challenging task but they also bring linguistic research further. Thus, the scope of the SIMPLE project makes it a truly pioneering project for Danish and considering the current status of language technology for the 'small' European languages, the development of these harmonised large-scale semantic lexicons is a first step in the right direction for creating advanced language technology also for less widely spoken European languages. Thus, the Danish SIMPLE lexicon constitutes the first large-scale attempt to give formalised, semantic descriptions of Danish word vocabulary.

Bibliography

Braasch, A., A. B. Christensen, S. Olsen & B.S. Pedersen (1998) 'A Large-Scale Lexicon for Danish in the Information Society', in: *Proceedings from First International Conference on Language Resources & Evaluation*, Granada 1998.

Braasch, A., B. Pedersen (1999). ‘En stor sprogteknologisk ordbog for dansk - med særligt fokus på håndtering af flertydighed i den niveaudelte ordbog’, in: P. Widell (ed.) *7. Møde om Udforskning af Dansk Sprog*, Århus Universitet.

Bergenholtz, H., (1990). 'DK87-DK90: Dansk korpus med almensproglige tekster', in: M. Kunøe & Erik Larsen (eds.) *3. Møde om Udforskning af Dansk Sprog,* Aarhus Universitet.

Boje, F. & L. Schøsler (ed.) (1992). ‘DISEM - A Semantic MT-Component’ in: *CST Working Papers no. 1*, Center for Sprogteknologi, Copenhagen.

Christ, O. (1993) : *The Xkwic User Manual*. Institut für maschinelle Sprachverarbeitung, Universität Stuttgart.

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.

Kjærulff Nielsen: Engelsk- Dansk Ordbog, Gyldendal, Copenhagen.

Lenci, A. F. Busa, N. Ruimy, E. Gola, M. Monachini, N. Calzolari, A. Zampolli, J. Pustejovsky, E. Guimier, G. Recourcé, L. Humphreys, U. Von Rekovsky, A. Ogonowski, C. McCauley, W. Peters, I. Peters, R. Gaizauskas, M. Villegas, O. Norling-Christensen (2000). *SIMPLE Linguistic Specifications*, University of Pisa.

Malmgren, S. (1988). ‘On Regular Polysemy in Swedish’, in: *Studies in Computer-Aided Lexicography*, Almquist & Wiksell, Stockholm.

Nimb, S. & B. Pedersen (2000). ‘Treating Metaphoric Senses in a Danish Computational Lexicon – different cases of regular polysemy’, in: *EURALEX 2000*, Stuttgart, Germany.

Pedersen, B., & S. Nimb (2000) ‘Semantic Encoding of Danish Verbs in SIMPLE – Adapting a Verb-framed Model to a Satellite-framed Language’, in *Second International Conference on Language Resources and Evaluation*, Athens, Greece.

Pedersen, B. & B. Keson (1999). ‘SIMPLE - Semantic Information for Multifunctional Plurilingual Lexicons: Some Examples of Danish Concrete Nouns’, in M. Palmer (ed.) SIGLEX ’99, ACL Workshop, Maryland, USA.

Politikens Store Nye Nudansk Ordbog, Version 2.1, Politikens Forlag, Copenhagen.

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

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.

Scheuer, J. (1995). *Tryk på Danske Verber*, RASK Supplement, Vol. 4, Odense Universitetesforlag, Odense.

Underwood, N., C. Povlsen, P. Paggio, A. Neville, B.S. Pedersen, L. Jørgensen, B. Ørsnes, A. Braasch (2000). *LINDA, Linguistic Specifications for Danish*, CST Working Papers Report No.1, Center for Sprogteknologi, Copenhagen.

---

[i] By 'orthogonal inheritence' we understand multiple inheritance with the restriction that a feature can only inherit its value from *one* mother node from the same partition. Thus, in SIMPLE each meaning dimension (each qualia role) establisshes its own partition.

[ii] NDO indicates that the definition is taken from the Danish dictionary 'Nudansk Ordbog'.