Lang Resources & Evaluation (2009) 43:269–299
DOI 10.1007/s10579-009-9092-1

DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary

Bolette Sandford Pedersen · Sanni Nimb · Jørg Asmussen · Nicolai Hartvig Sørensen · Lars Trap-Jensen · Henrik Lorentzen

Published online: 14 August 2009

Abstract This paper is a contribution to the discussion on compiling computational lexical resources from conventional dictionaries. It describes the theoretical as well as practical problems that are encountered when reusing a conventional dictionary for compiling a lexical-semantic resource in terms of a wordnet. More specifically, it describes the methodological issues of compiling a wordnet for Danish, DanNet, from a *monolingual* basis, and not—as is often seen—by applying the translational expansion method with Princeton WordNet as the English source. Thus, we apply as our basis a large, corpus-based printed dictionary of modern

B. S. Pedersen (✉)
University of Copenhagen, Njalsgade 140-142, 2300 Copenhagen S, Denmark
e-mail: bspedersen@hum.ku.dk
URL: http://cst.ku.dk/

S. Nimb · J. Asmussen · N. H. Sørensen · L. Trap-Jensen · H. Lorentzen
Society for Danish Language and Literature, Christians Brygge 1, 1219 Copenhagen K, Denmark

S. Nimb
e-mail: sn@dsl.dk
URL: http://dsl.dk/

J. Asmussen
e-mail: ja@dsl.dk
URL: http://dsl.dk/

N. H. Sørensen
e-mail: nhs@dsl.dk
URL: http://dsl.dk/

L. Trap-Jensen
e-mail: ltj@dsl.dk
URL: http://dsl.dk/

H. Lorentzen
e-mail: hl@dsl.dk
URL: http://dsl.dk/

Danish. Using this approach, we discuss the issues of readjusting inconsistent and/or underspecified hyponymy hierarchies taken from the conventional dictionary, sense distinctions as opposed to the synonym sets of wordnets, generating semantic wordnet relations on the basis of sense definitions, and finally, supplementing missing or implicit information.

Keywords Wordnet · Dictionary · Lexical semantics · Semantic relations · Hyponymy · Nouns · Verbs

1 Introduction

During recent decades, a considerable amount of research within computational lexicography and computational lexical semantics has been devoted to examining the degree to which knowledge for computational semantic resources could be extracted from machine-readable dictionaries (cf. Boguraev and Briscoe 1989; Ide and Véronis 1995; Fontenelle 1997; Agirre et al. 2000; Kokkinakis et al. 2000; Ide and Wilks 2007; others). Containing an enormous amount of lexical and semantic knowledge, dictionaries are considered a likely source of information for use in computational semantic lexicons and semantic knowledge bases. Whereas Ide and Véronis (1995) conclude rather negatively that the results of reuse experiments are disappointing, and that in consequence the research community prefers to turn to text corpora as a source of semantic knowledge, other more recent experiments are more promising (e.g. Agirre et al. 2000; Kokkinakis et al. 2000; Vossen et al. 2008). One problem pointed out by Ide and Véronis (1995) is the difficulties involved in the automatic extraction of hierarchies and other semantic relations from dictionary definitions due to the fact that information is presented inconsistently even within the same dictionary. Another problematic factor, as also pointed out in several works related to automatic word-sense disambiguation (Kilgarriff 1997; Ide and Wilks 2007), relates to the fact that the sense distinctions given in dictionaries do not necessarily reflect actual usage. Finally, some types of information needed in lexicons for Natural Language Processing (NLP) do not exist in dictionaries at all since a considerable amount of knowledge is implicitly presupposed by the human dictionary user.

This paper contributes to the discussion of the three problem areas presented above in the sense that it refers to the theoretical as well as the practical problems that occur when a conventional dictionary is reused for compiling a lexical-semantic resource in terms of a wordnet. More specifically, it describes the methodological issues of compiling a wordnet for Danish, DanNet, on the basis of a large, corpus-based printed dictionary of modern Danish (Den Danske Ordbog, henceforth DDO) and discusses the issues of (1) readjustments of inconsistent and/or underspecified hyponymy hierarchies, (2) sense distinctions as opposed to the synonym sets of wordnets (synsets), and (3) supplementation of missing or implicit information.

In 2003 and 2004, two official Danish reports forcefully underlined the need for a lexical-semantic wordnet for Danish with the aim of facilitating flexible information

navigation in Danish text material in future computer systems. The publication of the reports coincided with the conclusion of two Danish projects: the aforementioned dictionary and a pilot version of a computational semantic lexicon for Danish comprising descriptions of 12,000 concepts in the so-called SIMPLE model [Semantic Information for Multifunctional, Plurilingual Lexicons, cf. Lenci et al. (2000) and Pedersen and Paggio (2004), henceforth SIMPLE-DK]. As a consequence, the compilation of DanNet was established as a collaborative project between the host of SIMPLE-DK: Centre for Language Technology at the University of Copenhagen, and the publisher of DDO: Society for Danish Language and Literature, an institution under the auspices of the Danish Ministry of Culture. DanNet can be downloaded under an open source license from www.wordnet.dk, and it currently contains 50,000 synsets and will be supplemented during the next 2 years to cover 70,000 of DDO's $\sim$100,000 word senses.

A widely discussed issue among wordnet developers concerns the choice between the 'expand approach' and the 'merge approach' when initiating a wordnet project (cf. Rigau and Agirre 2002; Fernández-Montraveta et al. 2008; Márton et al. 2008; Derwojedowa et al. 2008). It is generally accepted that the former approach—where a wordnet is produced by translating synonym sets from Princeton WordNet to the target language—is easier, cheaper and ensures better consistency between wordnets but on the other hand involves a genuine risk of linguistic bias. In contrast, the latter presents a more loyal picture of linguistic conceptualisation in a specific language but may for the same reason be less compatible with other wordnet structures; in addition, this strategy is more labour-intensive and thus correspondingly resource-demanding. Since the starting point of DanNet was a corpus-based, newly completed dictionary of Danish accessible in a machine-readable version with hyponymy information explicitly specified for each sense definition, the motivation for the merge approach was obvious. The fact that a wordnet for Danish could be semi-automatically built from carefully constructed sense distinctions where the set of senses was actually defined on the basis of corpus data, seemed to make it feasible to build a wordnet on monolingual grounds, which would be practically useful in NLP tools meant for Danish text material.

The paper is composed as follows: In Sect. 2, we discuss the semantic contents of a traditional dictionary in comparison with the kind of semantic data that a wordnet should ideally contain. This is followed by Sect. 3 where we look at DDO and describe the structure of the semantic part of this dictionary and sketch out the reuse perspectives of the information given there. In Sect. 4 we move on to the actual compilation of DanNet and discuss the necessary readjustment of hyponymies and the treatment of the so-called ISA-overload. It is shown how 1st Order Entities are treated differently than 2nd and 3rd Order Entities (Lyons 1977), and how the reuse perspective differs within each semantic class. In Sect. 5 we discuss how semantic relations other than hyponymy have been encoded in DanNet, and finally in Sect. 6 we discuss some evaluation issues of the resource. We present, in Sect. 5, two experiments that we have done on automatic extraction of relations from DDO definitions, and we discuss the many cases where reallocation and/or addition of information is needed in order to guarantee a consistent level of semantic description in the target resource.

2 Lexical semantic information in dictionaries and wordnets

Dictionaries rely heavily on human pragmatic knowledge and the language-user's ability to make assumptions without any explicit statements in the text (Svensén 1993: 133; Zgusta 1988). This implies that not all the information needed in wordnets can be extracted from a dictionary. Definitions for computational use have to make all information explicit and 'assumptions' in a wordnet can only be made by calculating semantic relatedness via the inheritance mechanism and the relations established by the specific linking between synsets. While it is normally not a problem for an editor compiling a monolingual dictionary to estimate the extent of the assumptions made by the reader, this is more challenging for the editor of a wordnet. In fact, in Veale and Hao (2008), it is argued that much of the knowledge needed in order to understand everyday language is not necessarily the kind of knowledge found in a dictionary. Much of it is based on stereotypes and culturally inherited associations and wordnets should be enriched with this type of information, for example that snakes are related to treachery and slipperiness, and that elephants have a good memory. This clearly lies outside the scope of DanNet at its current stage, but some of the knowledge which is not expressed in the dictionary definition, although clearly being part of the native speaker's lexical knowledge about the concept, should definitely be explicated.

There are numerous reasons why some semantic information, which is in principle relevant in a dictionary, is often left out of the definition. In some cases, information is given implicitly in an example of language use or by some characteristic collocates. The sense descriptions in a dictionary entry are composed of elements that supplement each other without too much redundancy so that they can be read as a whole. These facts force the editor of a wordnet to look carefully for semantic information elsewhere in the dictionary entry before editing the wordnet entry. However, very often information is left out simply to avoid describing something that is common knowledge to all readers. For example, nothing is generally said about the human user when DDO describes the use of instruments and buildings since it is obvious to the reader. Only when the user belongs to a very restricted group is it mentioned in the definition. Some of the few examples are: 'police dog: used by the police'; 'diver's watch: used by divers' and 'doggy bag: used by customers in a restaurant'. In a wordnet, however, the user of an artefact should be described systematically for all concepts having a specific purpose. Thus, we add in DanNet that a lipstick is typically used by women even though this is not indicated in DDO, and that a shaving brush is used by men although DDO does not provide this information. In the quite similar cases of brilliantine and summer dress, DDO does in fact include information about the typically male and female user.

Furthermore, the substitution principle applied in many monolingual dictionaries according to which the definition should be phrased in such a way that it can replace the headword in a text, easily leads to the omission of semantic information. For complex concepts it is indeed difficult to forge a definition on this principle without leaving out information. This explains why nothing is said about the painter nor about the motif in the definition of a painting in DDO.

When information on display windows is not given in the definition proper of *butik* (shop), it is probably the substitution principle as well that leads to omission of

information. The collocation *se på butikker* (lit. look at shops, 'to go window shopping') as well as the example: 'We walked down the pedestrian street Strøget. Mona stopped in front of almost every shop. She loved looking at clothes' indirectly inform the reader of the fact that shops normally have a display window.

In DDO we also find cases where the dictionary is simply imprecise. A case in point is the word *bjælkehytte* (log cabin) that is defined as 'small house or hut built with beams'. This tells us nothing about the material of the beams, since *bjælke* can be made of wood, metal or concrete. It is evident that the semantic relation MADE_OF: *træ* (wood) should be added in DanNet. Another case is *registreringsattest* (vehicle registration certificate) where the information that the involved agent is a *motorkontor* (motoring office—the authority which issues this certificate) must be added to the DanNet entry. In DDO, there is no link whatsoever between these two words.

Interestingly, inheritance can facilitate the manual enrichment of semantic information. The inheritance mechanism ensures that relations are added systematically to all hyponyms (to be restricted to a narrower synset if necessary or blocked if inheritance is unwanted). For example all hyponyms of *butik* (shop) inherit the involved agent *handlende* (shopkeeper). Thus, the DanNet editor is prompted to identify the involved agent of the more restricted hyponym: that the shopkeeper of a pharmacy is a pharmacist, the shopkeeper of a bakery is a baker and so on. Such information is only rarely specified in DDO definitions (although sometimes provided implicitly as examples of word formation), but this information is seen as highly relevant in a wordnet. The semantic descriptions of artefacts in DanNet have been systematically supplemented with this type of information, creating links between synsets like *klaver/pianist* (piano/piano player) and *flycertifikat/pilot* (pilot licence/pilot).

3 Information types in DDO

DDO—the first and only corpus-based dictionary of Danish—is a printed dictionary in six volumes compiled by the Society for Danish Language and Literature and published 2003–2005. It comprises approximately 100,000 word senses described in about 63,000 entries. It gives detailed information on spelling, morphology, pronunciation, meaning, collocations, fixed phrases, syntax, usage, word formation and etymology, and thus addresses a wide variety of potential users. The dictionary is primarily based on the Corpus of the Danish Dictionary (DDOC), a reference corpus of contemporary Danish (Norling-Christensen and Asmussen 1998).

In order to achieve a high level of consistency in the semantic description, the dictionary entries were written in groups of semantically related words rather than in alphabetical order (cf. Lorentzen 2004). Templates for sense description were developed and applied for the individual groups. Function words were edited in groups of word classes. For the purpose of enabling future reuse of the data, a fine-grained microstructure was designed, which also included elements not meant for presentation in the printed dictionary. Even if the dictionary so far is only publicly available as a printed edition, it was edited in machine-readable format (SGML/XML), based on which an online version will be launched in 2009.

Fig. 1 Semantic description in DDO

The idea has been to transfer a substantial part of the sense definitions in DDO into synonym sets (synsets) in DanNet. In Fig. 1, an example is given to show the semantic description of the sense 'painting' in the entry for the noun *maleri*. Let us briefly comment on the different elements used in the entry: The first content element is ⟨Sysfag⟩ which is one of the above-mentioned non-printed elements in the dictionary, indicating information on subject or domain. This element has been filled in whenever possible, but it is only displayed in the printed dictionary if the sense in question is used as a professional or technical term. The information has been automatically transferred into DanNet by means of a semantic relation which links to the corresponding synset of the subject; in this case *maleri* is linked to the corresponding synset of the subject 'kunst' (art). In DDO, 617 senses carry information on the subject 'kunst', for instance lemmas like *portræt* (portrait), *galleri* (gallery) and *fortolkning* (interpretation) and the automatic reuse in DanNet of this broadly defined subject assignment links all these senses to the same synset in DanNet (*kunst_1* (art)) and thereby indirectly to one another.

The element ⟨Denbet⟩ contains the sense definition as well as two DanNet-related attributes that ensure a fixed linkage between the DanNet concepts and the definitions in DDO. ⟨Denbet⟩ is followed by yet another non-printed element ⟨Genprox⟩ indicating the hypernym (or *genus proximum*) of the sense as used in the definition. Wherever possible, definitions in DDO have been composed as so-called true definitions (Svensén 1993: 177) which are intensional and follow the classical scheme of giving the closest hypernym of the definiendum, the *genus proximum*, and the *differentia specifica* that distinguishes the definiendum from its co-hyponyms. The element ⟨Onym⟩ contains information on synonyms, near-synonyms and

antonyms which in DanNet evidently is very convenient information, particularly for the establishment of synsets. In Fig. 1, one near-synonym is given, *skilderi* (painted or drawn picture) and *skilderi* is consequently related to *maleri* in DanNet by the relation NEAR_SYNONYM.

In Fig. 1, the elements under ⟨Rel⟩ contain collocational information, and finally a usage example is given in the element ⟨Citat⟩. These last elements are, however, less relevant for DanNet purposes but because the sense descriptions in DDO should be readable as wholes they sometimes contain relevant information needed for the encoding of some of the relations as will be discussed in Sect. 5, such as information about typical agents or other semantic aspects of the word described.

4 The construction of hyponymies

4.1 Adjusting the hierarchy

As indicated above, all genus specifications from DDO have initially been extracted directly from the ⟨Genprox⟩ elements and stored in the DanNet encoding tool. A special interface (Fig. 2) allows the editors to manipulate these data in order to select which data should be related to which sense of the genus. In the unproblematic cases, the task of the editors when establishing the hyponymy structure of the wordnet has been solely to accept the suggested synset behind a DDO-given genus expression (the system suggests by default the first sense of the genus). In spite of this encoding facility which has significantly speeded up the start phase of the project, the general adjustment of the wordnet to ensure the construction of a reasonably consistent hierarchy turned out to be a somewhat cumbersome task.

First of all, the adjustment includes the disambiguation of the genus expression which in DDO is *not* a unique reference to a sense. As illustrated in Fig. 2, the term *celle* (cell) is used as genus proximum for both 'yeast cell', 'prison cell' etc. but clearly has to do with different senses of the word; a fact which is set out explicitly in DanNet by linking the hyponyms to different synsets, i.e. *celle_1*, *celle_2* and so forth.

Furthermore, the genus expressions assigned in DDO were not taken from a predefined set of ontological concepts, but were rather decided upon by the individual dictionary editors on the basis of some general guidelines with the main purpose of communicating the sense to a human reader by one neat and well-turned phrase. Therefore, the DanNet editors often had to change the proposed genus and instead decide upon a closer or more precise hypernym; for instance all researchers, such as for instance *sprogforsker* (linguistic researcher/linguist) have been subsumed under *forsker* (researcher) in DanNet even if they have *person* (person) as hypernym in DDO. In other cases, the different genera prove to be synonyms and thus belong to the same synset in DanNet anyway, such as *anordning, indretning* (device, appliance). The inheritance of relations from a hypernym to its hyponyms furthermore helped to clarify whether the established hierarchy was reasonable, showing if the first choice of a hypernym was appropriate or not. As DanNet aims at facilitating the calculation of semantic similarity between concepts (for instance to be used in information retrieval), this harmonisation of hypernyms constitutes a very

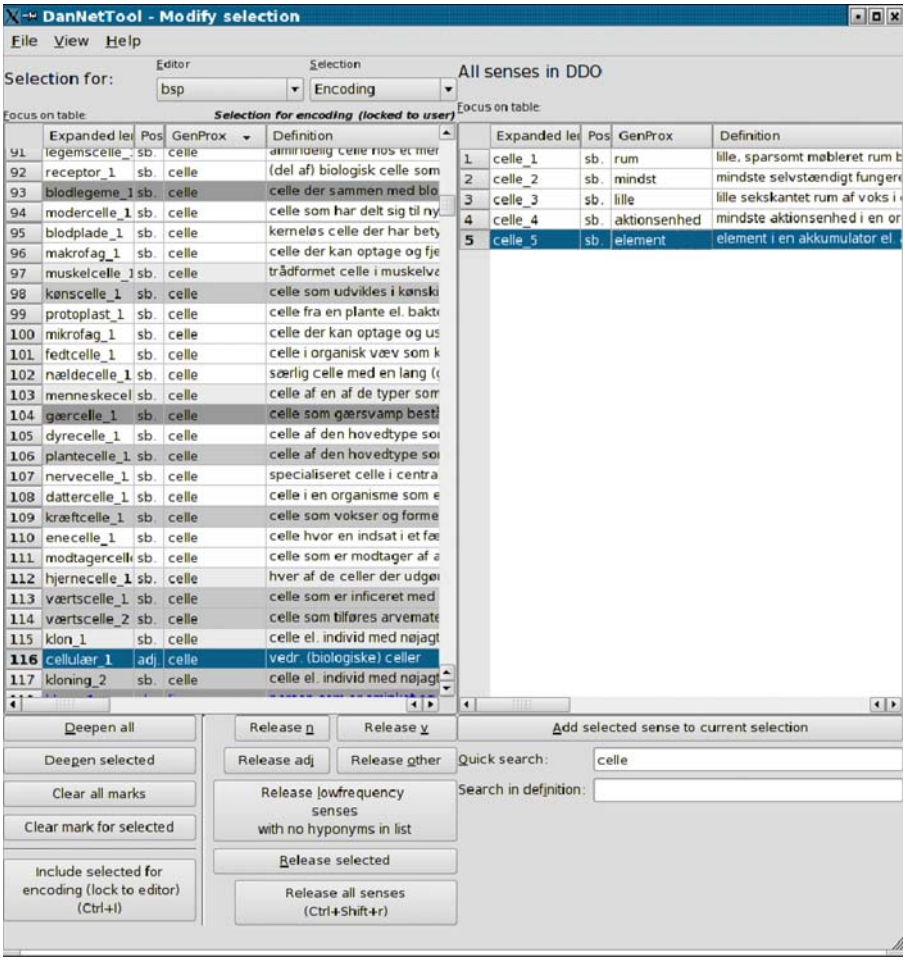

Fig. 2 Encoding interface where all subordinates of the genprox(es) for *celle* (cell) in DDO are presented to the editor for acceptance or modification. In the right column the editor is introduced to the different senses of *celle* in DDO

important part of the hierarchy building. In some cases, new patterns of synonymy were disclosed as some hypernyms proved to have semantically quite similar hyponyms in DDO. For instance, the terms *informatik* (informatics), *bromatologi* (food science), *samfundsfag* (social studies), and *datalogi* (computer science) were accidentally described under three different hypernyms in DDO; a situation which is treated in DanNet by merging the hypernyms *lære* (discipline), *fag* (subject), *videnskab* (science) into one synset.

4.2 Treatment of the ISA overload

A very central aspect which is not accounted for in DDO definitions since it is hardly relevant for the human user of a dictionary, is that lexical items prove to

expose hyponymy of great variety. Most wordnets—just like traditional dictionaries—generalise over this variation and do not distinguish between different kinds of hyponymy; a characteristic that has led to the problem of the so-called *ISA overload.* The ISA overload can be defined as a situation where sets of unequal hyponyms are grouped as simple sister terms under the same superordinate as in the case of the following examples from Princeton WordNet:

An oak HAS_HYPERNYM (ISA) *tree*
A bonsai HAS_HYPERNYM (ISA) *tree*

Oak refers to a specific kind of tree in botanical terms, whereas *bonsai* refers to any kind of tree (although some kinds are probably prototypical) grown in a fashion where the roots and the branches are kept small. Already in Miller (1998) this fact is acknowledged as a serious weakness of Princeton WordNet, and also several formal ontologists and wordnet editors have pointed out that the ISA overload constitutes a serious problem when a wordnet is used for inferencing, cf. Guarino (1998), Guarino and Welty (2002) and Huang et al. (2008). Expressed in another way, improperly structured taxonomies make models confusing and difficult to reuse or integrate, and this fact counts in particular for wordnets if the aim is to integrate them as part of computational models and not only to see them as lexicographical repositories organized in a slightly different way than traditional dictionaries. Computational models make heavy use of inheritance mechanisms, and such mechanisms are easily messed up if the taxonomy is not sound. The problem of the ISA overload has generally attracted more attention from formal ontologists than from the WordNet community itself. An exception is Huang et al. (2008) who suggests enriching wordnets semantically by introducing the relation of paranymy. This relation enables the builders of the Chinese wordnet to classify conceptually salient groups: a set of paranyms is defined as terms that are grouped together by one conceptual principle. In other words, the paranymy relation is used to refer to the relation between any two lexical items belonging to the same semantic classification.

Heterogeneity among subordinates constitute a severe practical problem when a wordnet is compiled directly from lexicographical vocabularies as is the case of DanNet. Consider for instance some of the co-hyponyms extracted from DDO under the hypernym *maleri* (painting), see Fig. 3.

According to Huang's approach, two sets of subordinates could be established here, one regarding the semantic dimension of the type of paint used in the painting (oil painting vs. water colour), and one concerning the object that is depicted (seascape vs. flower painting) as shown in Fig. 4.

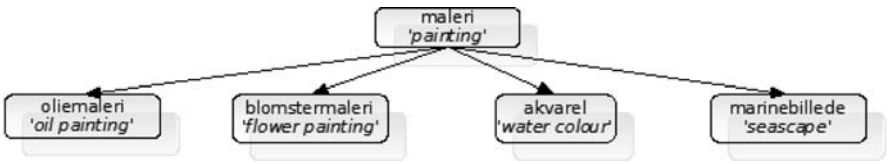

Fig. 3 Subordinates of *maleri* ('painting')

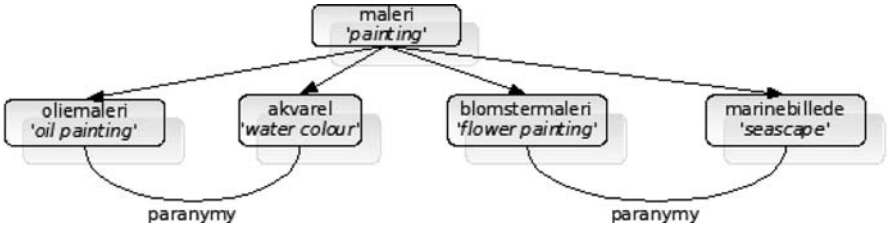

Fig. 4 Establishing paranymic relations between groups of synsets

A central characteristic of both semantic dimensions is that *taxonomical relations* can be established between the superordinate and the co-hyponyms within the same set of paranyms. In DanNet, taxonomy is defined as a specification of hyponymy, as generically described by Cruse (1991):

An X is a kind/type of Y

where co-hyponyms are mutually incompatible. In contrast, the more general hyponymy relation is described as *An X is a Y*. Taxonomical structures are—as opposed to the more general hyponymy—attractive because of their clearer ontological status with regard to inheritance mechanisms and other inferences. To put it another way, a water colour cannot at the same time be an oil painting; a seascape not a flower picture. But nothing prevents a seascape from being an oil painting at the same time.

With the lexicographical point of departure given in DanNet, however, the number of terms that *cannot* be defined as taxonomical in relation to their superordinate is overwhelming. The subordinates in Fig. 5 can hardly be defined as *kinds* of paintings but rather as derogative evaluations of these.

In search of an approach that can help define and distinguish such non-taxonomical terms from the taxonomical ones and define their status in the lexical hierarchy, the varieties of hyponymy must be examined further. Cruse (2002) provides such an examination and introduces three subdividing categories of terms, namely: *natural kinds*, *nominal kinds* and *functional kinds*.

Natural kinds are defined as naturally occurring things like animals, plants and naturally occurring materials and substances like wood, stone and water. Cruse (2002: 18) states that 'the names of natural kinds behave to some extent like proper names in that they show referential stability in the face of quite radical changes in the speaker's beliefs concerning the referent'. Put in another way, natural kinds generally possess what Guarino and Welty (2002) label *rigid properties*, i.e. properties that guarantee identity through change (thus according to Guarino and Welty *person* possesses rigid properties whereas *student* does not).

Such entities are generally assumed to be good candidates for the skeleton of a sound taxonomy and therefore constitute a good starting point for building the lexical network. For natural kinds it is generally true to say that *X is a kind of Y* as in *a pear is a kind of fruit*, and likewise the hyponyms 'apple' and 'pear' are mutually incompatible. Thus, they fulfil our restricted definition of taxonomy, and they obey

the general rules of inheritance according to which a subordinate inherits the characteristics of its superordinate. Another characteristic of natural kinds is that it cannot easily be defined what distinguishes one natural kind from another. What distinguishes an apple from a pear? Shape, colour and taste are relevant features, but these do not completely describe and distinguish the fruits from each other.[1] *Grøntsag* (vegetable), on the other hand, *can* be defined with a single feature, namely '(a part of) a plant that serves as food for humans'. It also inherits the characteristics of plants or parts of these, but cannot be described as *a kind of plant*. *Grøntsag* (vegetable) is therefore an illustrative example of a concept that cannot be classified as a natural kind, even though it refers to naturally occurring things. We shall return to these kinds of terms later.

The second category of taxonomical terms is referred to as *functional* terms. Functional terms typically refer to artefacts, the function of which plays a central role in their definition. They share certain features with natural kinds; mutual incompatibility is the most general characteristic: trumpets and trombones are incompatible since they are both *types of* musical instruments. Also, just as it is not easy to define what distinguishes a pear from an apple, it is not easy to define uniquely what distinguishes a trumpet from a trombone, both being subordinates of brass instrument. Differences in size, shape and sound come to mind, but again it is not possible to define the differences with a single feature. As a consequence, we describe such functional-kind terms as taxonyms in DanNet as in the case of natural kinds.

Contrary to natural and functional kinds, nominal kinds cannot be described as *a kind of* or *a type of* and are therefore not considered taxonomical. As a further characteristic, the relation between nominal kinds and their hypernyms can typically—unlike the two former categories—be captured in terms of a few differentiating features (Cruse 2002: 18) as we saw for *grøntsag* (vegetable).

Nominal kinds are found in both natural domains and artefactual domains. Since nominal kinds tend to confuse the natural and functional taxonomies, it seems obvious *not* to consider such senses as taxonyms, but as terms that are *orthogonal* to the taxonomy. For further examples of nominal kinds, consider lemmas with the hypernym *person*. DDO contains more than 4,000 lemmas with this hypernym [such as *passager* (passenger), *idiot* (idiot), *læser* (reader) and *medlem* (member)] and they are more or less all nominal kinds since they are not *kinds of persons*, but describe dimensions of persons with different characteristics highlighted, some of which are stable over time and some of which are only related to a specific situation (cf. Pedersen and Sørensen 2006 for a suggested qualia structure-based explanation model to such terms). As mentioned, they are generally easily definable by means of a single feature or very few features, in other words a passenger can be defined as a person that travels in a vehicle without being the driver, a reader as a person that reads etc.

Returning to the terms depicted in Figs. 3, 4 and 5, this distinction may be illustrated by introducing a graphical difference between the taxonomical and the

[1] Actually, Ruus (1995: 130) argues that some of these hyponyms are characterised by the fact that a *limited* set of features can distinguish them from each other. She uses Grandy's terminology and calls such hyponyms *contrast sets*.

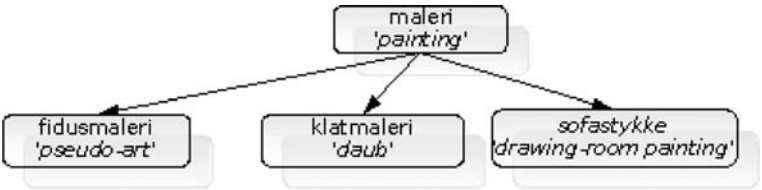

Fig. 5 Non-taxonomical subordinates of *maleri* (painting)

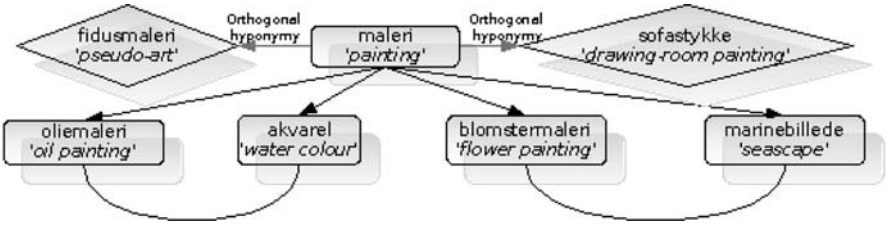

Fig. 6 *Maleri* ('painting') with two orthogonal hyponyms

non-taxonomical terms of *maleri* (painting). In Fig. 6, the rhombuses refer to terms that are classified as orthogonal, whereas the squares refer to taxonyms.

The nominal kinds are best regarded as belonging to the same level as their hypernym since practically any of the hyponyms of *maleri* could function as a specialisation of them; the pseudo-art painting could be either a water colour or a seascape and so forth.

4.3 Encoding the hyponymy structure of 1st Order Entities

Taking account of the findings accounted for in Sect. 4.2, DanNet distinguishes between taxonomical and orthogonal hyponymy by means of a feature (*ortho)* assigned to the relation HAS_HYPERNYM. As sketched out above, a set of linguistic tests can be applied to verify whether a given term (or set of terms) determines a *kind of* something, and whether sisters of the same superordinate are mutually incompatible. For practical reasons, and although it appears theoretically attractive and sound, the paranymic relation (cf. Huang et al. 2008) has not been implemented in DanNet. Alternatively, a pragmatic decision is made in each case regarding *which* dimension of meaning should be seen as the basic taxonomic one within a given subpart of the hierarchy. The choice of taxonomical viewpoint is in most cases straightforward, but as we have seen, there are cases where different semantic dimensions compete for the primary categorisation scheme. Especially if we move into the grey-zone areas of domain-specific vocabulary, it is often tempting to adopt a more specialized approach to the vocabulary. But since the main focus of DanNet is on capturing the characteristics of the general vocabulary, an intuitive layman approach is adopted in the basic taxonomy and as a consequence more specialized semantic dimensions are encoded as orthogonal. Consider for example the functional taxonomy of *møbel* (piece of furniture) to the left in Fig. 7.

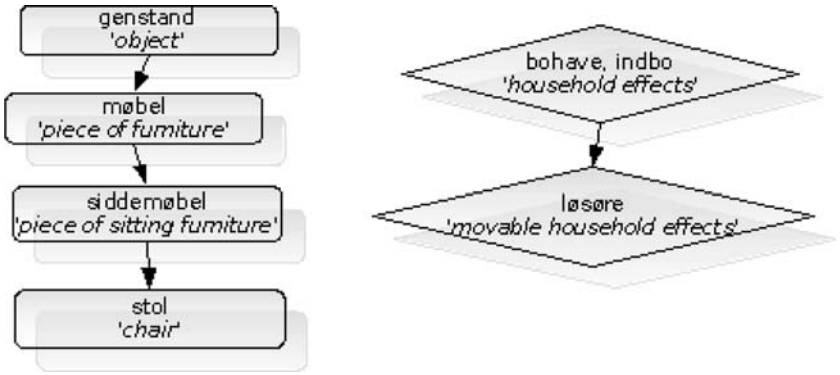

Fig. 7 The taxonomy of *møbel* (piece of furniture) and two grey-zone synsets *bohave, indbo* (household effects) and *løsøre* (movable household effects)

The two synsets to the right, in contrast, are frequently (but not only) used in the insurance domain but have still been included in the general vocabulary because they occur with a certain frequency in everyday language. They also relate to furniture, but from a different semantic perspective than their basic function, namely from the perspective of whether they can easily be moved (for instance stolen) or not. Such grey-zone terms that introduce another semantic perspective than the basic one, are generally classified as orthogonal in DanNet and not seen as part of the basic taxonomy. In this case, they are actually placed under *samling* (collection, group) and relate to furniture by the relation HAS_MERO_PART *møbel*.

In order to be ontologically compatible with other wordnets developed within this framework, all synsets of 1st Order Entities are furthermore assigned a top-ontological type originating from the EuroWordNet Top Ontology (see Fig. 8 for 1st Order Entities; Vossen 1999). The structure of the ontology relates to Lyons' three-category structure (Lyons 1977: 443ff.) of 1st, 2nd and 3rd Order Entities as well as to a four-dimensional structure comprising Origin, Form, Composition and Function. The ontology allows for complex ontological types such as for instance PLANT + PART + OBJECT + COMESTIBLE assigned to all edible fruits, or ARTEFACT + LIQUID + COMESTIBLE assigned to soups and drinks.

4.4 Encoding troponymy of 2nd Order Entities

Compared to the encoding of 1st Order Entities, it is far more problematic to use the hypernymy information from DDO as the starting point for the encoding of 2nd Order Entities. These include a majority of verbs and many verb senses share the same few and very polysemous verb senses as genus proximum in DDO. By way of illustration, 4,755 verb senses (25% of the 19,000 verb senses in DDO) share the same 15 very common verbs, such as *gøre* (to do), *være* (to be), *give* (to give), *få* (to get), *bevæge* (to move), and in addition *blive* (to become). This is further complicated by the fact that these 15 verbs on average have no less than 22 main senses and subsenses each. As a consequence, the genus proximum given for verbs

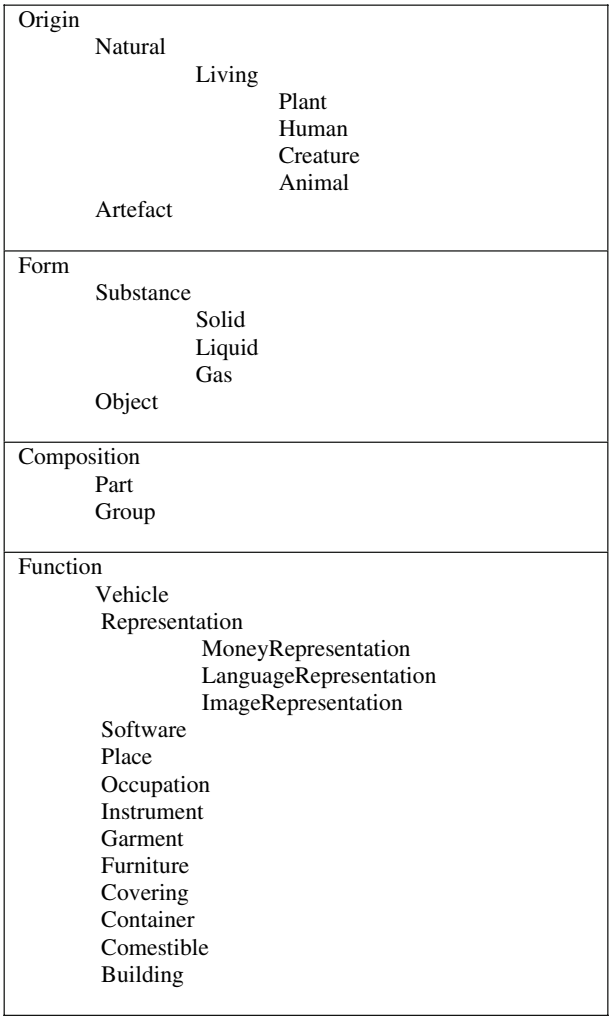

Fig. 8 Top-ontological assignments to 1st Order Entities (cf. Vossen 1999: 139)

in DDO does not automatically yield a reusable structure of a general hierarchy. Instead, information from the network of SIMPLE-DK has been integrated and a large set of the verb hyponymy relations given in DDO has been manually adjusted. This was furthermore complicated by the fine-grained sense distinction in DDO. About 1,500 of the 6,600 verbs in DDO have more than one sense, meaning that approximately 14,000 of the 19,000 verb senses come from these 1,500 verbs which have an average of 10 senses each. Due to doubt as to whether these fine-grained semantic distinctions in DDO are at all manageable in a wordnet, it was decided to merge a verb sub-sense with its main sense when the sub-sense represents (1) a more restricted sense, or (2) an extended sense. However, figurative sub-senses are generally maintained, belonging often to a different ontological type. To give an

example, consider the verb *skrælle* (to peel) which has a main sense with two subsenses related to it. The merging strategy results in two synsets in DanNet where the first includes the main sense (to peel a fruit) as well as the extended subsense (to remove the surface from something). The second synset includes the figurative subsense ('to disclose' in the figurative sense).

Although highly inspired by the SIMPLE-DK descriptions of events [built on Levin's classes (Levin 1993), which focus more on the semantic content of the event, cf. Lenci et al. (2000) and Pedersen and Nimb (2000)], DanNet applies the EuroWordNet Top Ontology of 2nd Order Entities (Fig. 9) as the upper level of ontological assignments to events. The main dividing principles are those of static versus dynamic events as well as that of telicity, as reflected in the situation types BoundedEvent and UnboundedEvent. However, in Danish, as in other Germanic languages, telicity is in most cases specified by means of verb particles and not—as in Romance languages—given in the verbal root. This can be seen for instance for the verb *spise* (eat), which seen in isolation denotes an atelic, unbounded event as opposed to the phrasal verb *spise op* (finish one's food, eat up), which denotes a telic, bounded event. This typological characteristic is reflected well in DDO, and the distinction between telic and atelic senses has been transposed more or less directly to DanNet. Phrasal verbs in general constitute a large part of the encoded senses in DanNet, and many verbs have parallel encodings as bounded and unbounded events depending on the presence or absence of a verb particle.

If building meaningful hierarchies of 1st Order Entities is a challenge, the construction of event hierarchies is an even tougher job and much less intuitive. Apart from the fact that there seems to be an extra measure of indeterminacy in the meaning of a verb, which complicates the issue, verb meaning seems to be better described along other dimensions than that of taxonomy, such as argument structure, event structure or meaning components/frame elements (cf. Levin 1993; Pustejovsky 1995; Fillmore et al. 2003). However, some level of hierarchical description of events does seem appropriate. The building of event hierarchies in wordnets is discussed in Fellbaum (2002), who argues that verb hierarchies are best described by means of the manner relations defined as *troponymy*. However, Fellbaum acknowledges that these cannot always be defined as mono-dimensional, a fact that is illustrated in English by examples like *move* and *exercise*. In this case Fellbaum proposes that parallel hierarchies should be established, allowing verbs like *run* and *jog* to act as subordinates of both *move* and *exercise*.

When treating events in DanNet, many similar cases of parallel hierarchies are found, and multiple inheritance is a way of accounting for subordinates with more than one superordinate. For instance, as seen in Fig. 10, the troponymy relation is generally established without implications for the physical domain itself (in this case food preparation) and allowing for multiple inheritance makes it possible to account in a flexible way for the fact that *salte* (to salt) can both be seen as a way to prepare food (*tilberede*) and to preserve it (*konservere*).[2]

[2] It should be made clear that multiple inheritance is also rather frequent with 1st Order Entities. For instance, in the previously mentioned example of *grøntsag* (vegetable) several vegetables are encoded partly with *plante* (plant) or *plantedel* (part of plant) as hypernym, partly with *grøntsag* as hypernym.

Fig. 9 The EuroWordNet Top Ontology for 2nd Order Entities (cf. Vossen 1999: 139)

SituationType
- Dynamic
  - BoundedEvent
  - UnboundedEvent
- Static
  - Property
  - Relation

SituationComponent
- Cause
  - Agentive
  - Phenomenal
  - Stimulating
- Communication
- Condition
- Existence
- Experience
- Location
- Manner
- Mental
- Modal
- Physical
- Possession
- Purpose
- Quantity
- Social
- Time
- Usage

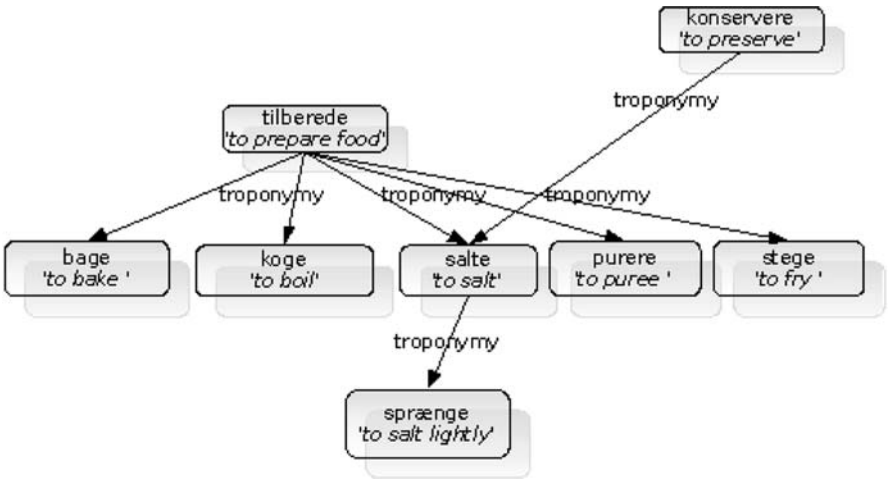

Fig. 10 *Tilberede* (prepare food, cook) with some of its troponyms

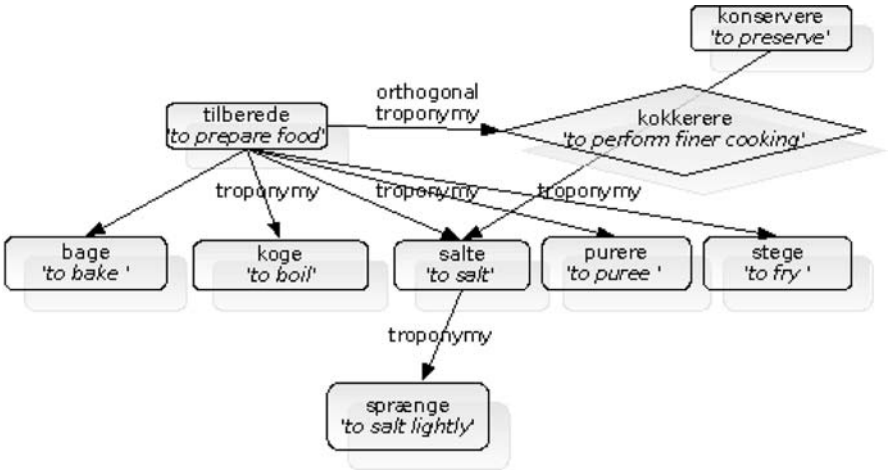

Fig. 11 *Kokkerere* (to perform finer cooking) as orthogonal to *tilberede* (to cook)

On closer study, however, several verbs in the domain tend to denote another dimension of the manner relation than the prototypical one. The verb *kokkerere* is another word for preparing food, but it usually gives a specific association to finer cooking. Therefore, it has been placed as a subordinate to *tilberede*, which seems at first glance unproblematic. However, it specifies a semantic dimension which is different from that of other cooking troponyms where focus is clearly on the manner component, describing the exact process that the food is undergoing. Therefore, in DanNet, *kokkerere* is encoded as orthogonal to the other hyponyms of *tilberede* (to prepare food/to cook), again visualised in the figure by a rhombus. Note that an orthogonal synset is characterised by its compatibility with its sisters; in finer cooking, the ingredients may both undergo boiling, salting and pureeing (Fig. 11).

In some physical domains, specific manner relations seem to hold. Subsumed under the verb *fjerne* (to remove), we find a series of verbs such as *affugte* (to dehydrate), *afkalke* (to descale/decalcify), *afluse* (to delouse), and *affarve* (to discolour, bleach) which specify *what* is removed from the object rather than *how* it is removed.

In the domains of mental verbs, even more subtle meaning dimensions are specified in the different hyponyms. Under the verb *tænke* (to think), verbs like *dagdrømme* (to daydream), *bekymre sig* (to worry), *forske* (to research) and *mindes* (to recall) are found; a very heterogeneous group of verbs organized along different meaning components that are not satisfactorily labelled as *one* manner relation. In such cases, we rely at this stage on a flat and to some extent underspecified structure, envisaging, however, to merge DanNet partly with a morphosyntactic resource (see Pedersen et al. 2008 for a pilot study), partly with a frame-like verb description model which is semantically richer.

4.5 Encoding hyponymies of 3rd Order Entities

3rd Order Entities—or abstract entities—comprise what Lyons (1977: 443, 445) defines as 'such abstract entities as propositions, which are outside space and time. Third-Order Entities are such that 'true', rather than 'real', is more naturally predicated of them; they can be asserted or denied, remembered or forgotten; they can be reasons, but not causes; and so on'. With regards to ontological assignments, 3rd Order Entities are assigned the same set of situation components that are used for 2nd Order Entities (shown in Fig. 9). The 3rd Order Entities encoded in DanNet generally fall into six subgroups referring to concepts like:

- Movements of thought (3RD ORDER + MENTAL + SOCIAL)
- Institutions (3RD ORDER + MENTAL + PURPOSE + SOCIAL)
- Sciences and subjects (3RD ORDER + MENTAL + PURPOSE(+DOMAIN[3]))
- Rules and methods (3RD ORDER + MENTAL + PURPOSE + SOCIAL)
- Values and thoughts (3RD ORDER + MENTAL)
- Time entities (3RD ORDER + TIME)

The hierarchical structure within each subdomain is generally flat with only a small set of abstract hypernyms. These coincide roughly with the categories presented above, such as *idé* (idea), *system* (system), *institution* (institution), *fag/videnskab* (subject/science), *ret* (right), *metode* (method), *måde* (manner), *tanke* (thought) and *tidsenhed* (unit of time). A few of the concepts within 3rd Order Entities are linked directly to an artificially constructed superordinate *abstract_entity* since there is no intuitive Danish hypernym to refer to; examples of this are terms like *virkelighed* (reality) and *formål* (purpose). Within the abstract domain, only the very conventionalised concepts are encoded as strictly taxonomical, such as *minut* (minute) versus *time* (hour), *mandag* (Monday) versus *tirsdag* (Tuesday) as well as *filosofi* (philosophy) versus *datalogi* (computer science); all others are encoded as orthogonal.

5 Other semantic relations: from DDO definitions to DanNet

5.1 DanNet relations and features

The set of semantic relations encoded in a wordnet is in most cases likely to change over time when the most basic information types have been encoded and the need for more detailed information types emerges. In the current version of DanNet, the most frequently used semantic relations established in EuroWordNet have been supplemented with a few relations from SIMPLE (Lenci et al. 2000) that seem to be

[3] 'Domain' is an ontological type that has been inserted by DanNet (not EWN ontology). Such additions are given in cases where large groups of synsets call for a more specific ontological type than what is given by the EuroWordNet Ontology. Another example of a DanNet extension of the ontology is the ontological type BodyPart.

Formal Role	Agentive Role (ORIGIN)	Constitutive Role (COMPOSITION)	Telic Role (PURPOSE)
has_hyperonym	made_by (from SIMPLE)	has_holo_made_of	used_for (from SIMPLE)
has_hyponym		has_holo_part	used_for_object
is_a way_of		has_holo_member	role_agent
		has_holo_location	role_patient
		has_mero_made_of	
		has_mero_part	
		has_mero_member	
		has_mero_location	
		concerns (from SIMPLE)	
		involved_agent	
		involved_patient	
		involved_instrument	

Fig. 12 Semantic relations in DanNet

important as they often occur in dictionary definitions. For a more detailed exemplification cf. Vossen (1999) and Lenci et al. (2000). As in the SIMPLE model, semantic relations in DanNet have been organized according to the four qualia roles (Pustejovsky 1995) which relate to inheritance structure, origin, composition and purpose, respectively; see Fig. 12.

Apart from these relations, DanNet encodes also near-synonymy between synsets (for cases where two synsets are closely related semantically but cannot be regarded as one single concept), antonymy, and regular polysemy (e.g. a country seen as a geographical place and as a human group, respectively), as well as a small set of features on synsets such as the orthogonal feature discussed in Sect. 4 and the features connotation and gender (male or female reference). Finally, we encode information on disjunctive relations (e.g. HAS_MERO_MADE_OF *glass* or HAS_MERO_MADE_OF *wood*) and negated relations such as 'does not have feathers' etc.

5.2 Experiments with automatic extraction of semantic relations from DDO definitions

At the initial phase of the DanNet project, the expectation was that dictionary definitions could be automatically converted into wordnet entries. The optimism was primarily based on the explicitly marked-up genus expressions in the DDO definitions. This seemed to make automatic analysis of the definitions more feasible than in previous attempts to turn dictionaries into lexical-semantic NLP resources. This section discusses an informal pilot study that was carried out in order to achieve a clearer picture of the potential for approaches to automatical extraction of semantic relations from DDO definitions.[4] Focus is on potential automatic methods for extracting physical attributes as well as typical use or function of artefacts from their definitions in DDO.

[4] A more detailed account of this is given in Asmussen (2007).

In this pilot study, all DDO definitions were transformed into a special type of corpus amenable to common corpus-analytical investigation that may help shed light on the structure of the definitions. Genus expressions were tagged and thus explicitly searchable, the definitions themselves were tagged with the associated lemma. No other tagging was used in this corpus.

The definition of *fjernsyn* (television set) reads: *kasseformet apparat der kan modtage tv-signaler og omsætte dem til bevægelige billeder på en skærm og tilhørende lyd i apparatets højttalere* 'box-shaped device which can receive television signals and transform them into moving pictures on a screen with accompanying sound in the speakers of the device'. The genus expression of this definition is *apparat* (technical device) whereas the modifier *kasseformet* (box-shaped) and the VPs *modtage tv-signaler* (receive tv signals) and *omsætte dem* […] (transform them […]) must be considered differentia information. Based on these observations, for artefact definitions, it is possible to hypothesize that:

1. Adjectives preceding the genus denote general (physical) properties of the definiendum
2. VPs in a relative clause which are headed by *kan* 'can' specify the function or use of the definiendum, i.e. the USED_FOR relation.

To find more definitions with these structural characteristics, the hypotheses can be reformulated as queries to be performed on the definition corpus: A rough approximation of the first hypothesis is to find all the definitions in the definition corpus with the genus expression *apparat* and exactly one word—an assumed prenominal adjective—immediately to the left of the genus expression. A quick search through the definition corpus reveals that *groups* of prenominal adjectives are quite common as well. The corpus query is therefore broadened to cover these cases also. The total inventory of prenominal adjectives used in conjunction with the genus expression *apparat* is (frequencies in square brackets): *elektrisk* [23] (electric), *elektronisk* [16] (electronic), *optisk* [5] (optical), *mekanisk* [4] (mechanic), *lille* [4] (small), *kasseformet* [3] (box-shaped), *transportabelt* [2] (portable), *ballonbåret* [1] (balloon-carried), *computerbaseret* [1] (computer-based), *programmerbart* [1] (programmable), *fladt* [1] (flat), *mindre* [1] (smaller), *teknisk* [1] (technical), *tryktluftdrevet* [1] (powered by air compression), *stort* [1] (large), *rørformet* [1] (tube-shaped)—a total of 16 different adjectives occurring (partly grouped together) in 57 *apparat* definitions.

The total of *apparat* definitions in the dictionary is 209. When examining the remaining 152 definitions it turns out that none of the 16 adjectives listed recur. This result is quite counterintuitive as one would expect some of the adjectives to be relevant property specifiers in at least some of the remaining 152 definitions as well. A closer look at the adjectives reveals that specification of physical attributes in the *apparat* definitions is indeed quite scattered and inconsistent: A *computer monitor* can be either *box-shaped* or *flat*, whereas a *television set* can only be *box-shaped*; an *oven* is *technical* but nothing is mentioned about its shape; a *hearing aid* is *small* (but not *electronic*) whereas a *pacemaker* is *electronic* (but not *small*). In a printed dictionary for humans this kind of inconsistency is hardly a big problem since the

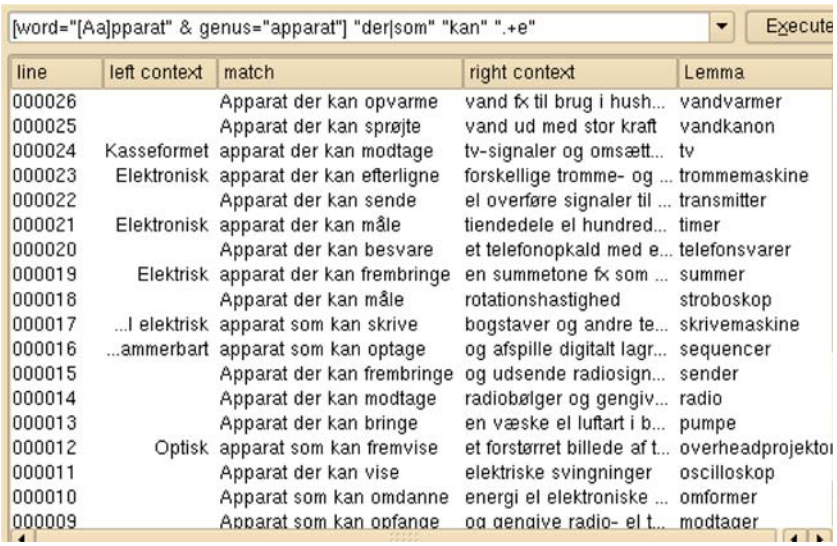

Fig. 13 *kan*-headed VPs denoting function or use of the definiendum

users will know how to interpret the information anyway, but it makes algorithmic exploitation of the material almost impossible.

When testing the second hypothesis on *apparat* definitions by rephrasing it as a simple corpus query, a total of 19 different verbs show up, of which the most frequent ones are *frembringe* (4) 'generate', *modtage* (3) 'receive', and *måle* (3) 'measure', some examples are shown in Fig. 13.

Each of the remaining 16 verbs occurs only once. The total number of occurrences is 26, a fact that clearly shows that the second hypothesis only covers a tiny part of the 209 *apparat* definitions. There may be two reasons for this: First, the definition is simply not captured by the rather rigid query. This is, for example, the case in the definition for the synonym pair *fjernskriver/telex* 'teleprinter'—intervening material that prohibits the query to match is surrounded by square brackets:

> *elektrisk apparat som [ligner en skrivemaskine og som tilsluttet et særligt netværk] kan sende og modtage skriftlige meddelelser*
> Lit.: 'electric device that [resembles a typewriter which connected to a special network] can send and receive written messages'

Second, the USED_FOR relation may emerge from other structural patterns than the one given in hypothesis 2. A quick review of some *apparat* definitions shows a variety of possibilities:

1. **Pattern** genus expression *der/som bruges til at* VP-inf *med*
   genus expression *that/which is used to* VP-inf
   Example *apparat som bruges til at spinde garn med*
   'device which is used to spin yarn'

Occurrences 3

2. **Pattern** genus expression *til at* VP-inf *med/på/i*
genus expression *to* VP-inf *with/on/in*
Example *apparat til at afspille cd'er med*
'device to play CDs on'
Occurrences 11

3. **Pattern** genus expression *der/som* VP-fin
genus expression *that/which* VP-fin
Example *apparat der måler og viser et køretøjs hastighed*
'device that measures and displays a vehicle's speed'
Occurrences 42

4. **Pattern** genus expression *til* NP
genus expression *for* NP
Example *apparat til optagelse og afspilning af lyd*
'device for the recording and playback of sound'
Occurrences 29

5. **Pattern** genus expression *der/som er specielt beregnet til at* VP-inf
genus expression *that/which is specially designed to* VP-inf
Example *apparat som er specielt beregnet til at optage og afspille tale*
'device that is specially designed to record and replay speech'
Occurrences 1

Patterns 1–5 cover 86 definitions. Together with the pattern from hypothesis 2, 70% of the *apparat* definitions are covered by six patterns. Once these patterns have been established, it becomes more feasible to extract the semantic information necessary to determine the USED_FOR relation automatically. But still, 30% of the definitions can probably not be processed automatically at all, as the variety of different syntactic ways to indicate semantic relations in definitions cannot be covered by a few algorithmic rules. In addition, the process of formulating these rules is in itself rather 'manual' and time-consuming. Furthermore, extraction with high precision would require a syntactically annotated definition corpus.

If dictionary definitions really are to be exploited automatically, they should be constructed in a fully predictable way with an explicitly defined syntax where syntactic patterns unambiguously resemble semantic relations. Even if many dictionaries use a controlled definition vocabulary and syntax, it does not seem likely that lexicographers would accept using such a thoroughly formal language for their dictionary definitions.

Another considerably more coarse way to isolate differentia expressions that could be used to identify the USED_FOR relation, is to apply a statistics-based approach where a frequency list of tokens in definitions with the genus expression *apparat* is compared with a frequency list of tokens in the definition corpus as a whole. Salient tokens from the *apparat* corpus can be determined by some statistical test such as log likelihood (Dunning 1994) or mutual information (Church and

Hanks 1989). Often, salient tokens express a central semantic feature, e.g. one that may indicate the USED_FOR relation to be included into DanNet. This is the case for some of the most salient tokens in *apparat* definitions, compared to all definitions as a whole, listed below in boldface. Lemmas with a definition that contains that particular salient token are listed as well:

- ***afspille*** '**to play back**': *grammofon*, *cd-afspiller* 'CD player', *afspiller* 'player', *sequencer*, *diktafon*
- ***afspilning*** '**play-back**': *kassettespiller* 'cassette player', *hjemmevideo* 'video cassette recorder', *kassettebåndoptager* 'cassette recorder', *båndoptager* 'tape recorder'
- ***måle*** '**measure**': *stroboskop*, *måler* 'measuring tool', *timer*, *løgnedetektor* 'lie detector', *ekkolod* 'sonar'
- ***måler*** '**measuring tool**': *gasmåler* 'gas meter', *speedometer* 'speed indicator', *omdrejningstæller* 'evolutions meter', *benzinmåler* 'fuel gauge', *fotofælde* 'speed camera'
- ***måling*** '**gauging**': *elmåler* 'electric meter', *trykmåler* 'pressure gauge', *luxmeter*, *spirometer* 'aeroplethysmograph', *gyrometer*, *alkometer*, *newton-meter*, *magnetometer*, *instrument*, *kalorimeter*
- ***målinger*** '**measurements**': *måleinstrument* 'measuring device', *radiosonde*, *satellit*, *fartskriver* 'tachograph'

By examining such automatically generated lists, the DanNet editor may get an idea of which synsets to supply with shared information regarding certain semantic relations, in the case of the extract shown above, the USED_FOR relation.

The brief examples given in this section indicate that it is possible to some extent to use some approaches from corpus linguistics to get a first impression of the structure and contents of dictionary definitions, but that the interpretation of the correlation between elements in the differentia of the definition and their appropriate semantic functions can only be performed by an editor. A fully automated transformation of dictionary definitions into wordnet entries seems hardly possible although certain corpus-analytical methods may prove useful in some cases. Thus, to determine the USED_FOR relation, the established patterns could be used to extract verbs from the definitions expressing this relation and these could be offered to the editor as possible descriptors among which the editor could then choose. But even so, this way of exploiting the differentia part of the definitions proved too tedious.

Hence, the compilation of semantic relations other than hyponymy has been performed manually, and the only current semi-automatic facility in the DanNet tool for speeding up the encoding process is an automatically established link between the words occurring in the differentia and the corresponding DanNet synsets as well as a facility enabling the encoding of synonyms extracted from DDO.

5.3 Manual encoding of semantic relations from DDO definitions

Figure 14 gives three examples of how DDO definitions are manually transformed into a set of semantic relations eased by the facility just described. As can be seen,

English translation of DDO definition	DanNet relations
kunstgenstand; kunstværk **(work of art)**: an **object** or a visible or audible **expression** which is the result of an **artist**'s creative ability, e.g. a painting, a sculpture, a ballet or a piece of literature	hypernym: **object** MANUALLY ADDED RELATIONS: INVOLVED_AGENT: **artist** MADE_BY: **to create** USED_FOR**: to express** used_for**: to expose**
maleri **(painting)**: a **work of art** in the form of a **painted** picture, typically made on a **canvas** and **framed**, and intended to be **hung** on the **wall**	HYPERNYM: **work of art** → RELATIONS INHERITED FROM **work of art**: INVOLVED_AGENT: **artist**; MADE_BY: **to create**; USED_FOR: **to express**; USED_FOR: **to expose** RESTRICTED TO → INVOLVED_AGENT: **painter** (Danish: *kunstmaler* (instead of **artist**, *kunstmaler* being the term for 'a painting artist' in Danish)) MADE_BY: **to paint** (instead of **to create**, **to paint** being a hyponym of **to create**) MANUALLY ADDED RELATIONS: HAS_MERO_PART: **canvas** MADE_BY: **to frame** HAS_HOLO_LOCATION: **wall** CONCERNS: **motif**
stilleben (still life): a **painting** or a **drawing** of placed objects, e.g. **fruits, flowers** or **jugs**	HYPERNYM: **painting** → RELATIONS INHERITED FROM **painting**: MADE_BY: **to paint** MADE_BY: **to frame** INVOLVED_AGENT: **painter** HAS_HOLO_LOCATION: **wall** HAS_MERO_PART: **canvas** USED_FOR: **to express** USED_FOR**: to expose** CONCERNS: **motif** MANUALLY RESTRICTED TO → CONCERNS: **fruit**; CONCERNS: **flower**; CONCERNS: **jug** (instead of CONCERNS: **motif** (**motif** being a (non-taxonomical) hyponym of **object**, meaning that any **object**, including **fruit**, **flower** and **jug** can be a **motif**))

Fig. 14 DDO definitions transformed into semantic relations in DanNet (for ease of reading in English translation)

information has to be moved up or down in the hierarchy in several cases to its appropriate place. For example the INVOLVED_AGENT relation given between *work of art* and *artist* is also relevant for describing (or rather restricting) paintings, even if it is not mentioned in the definition of painting itself (i.e. INVOLVED_AGENT *painter*). Likewise, the definition of *still life* strongly indicates that the motif might also be an important piece of information to give for the hypernym *painting* although this is not mentioned in the dictionary. Therefore, it is added to the synset of *painting* and later inherited to *still life* and all the other hyponyms of *painting*.

All three definitions are examples of intensional or true definitions and therefore, as mentioned previously, the most common type of definition in DDO. It is basic practice for all serious work on terminology to use only intensional definitions and to include as many distinctive features as needed in order to distinguish a concept from each of its co-hyponyms (Svensén 1993: 122–123). This is also DanNet's

ideal, but, as is well-known because of unclear distinctions between senses in general language, it is not always possible to distinguish all hyponyms from one another on the basis of semantic relations. It has been a general principle that, whenever possible, the distinctive features of the intensional definitions should be expressed as a corresponding wordnet relation. Obviously, however, the DanNet model proves too restricted in several cases. For instance, for the word *fresko* (fresco) the distinctive features in the DDO definition are expressed by adjectives like *humid*, *newly limed*, and *durable*, and these features cannot be expressed in the current DanNet lexicon since relations and features for all the meaning aspects related to properties have not been established in the data model yet.

The true definition is just one of several ways of defining the meaning of a word in a dictionary (cf. Svensén 1993: 116–117; Jackson 2002: 93–96). Monolingual dictionaries sometimes simply give a paraphrase consisting of synonymous, but more common words than the headword. Seen from a wordnet perspective, this type of definition presents possible members of the same synset as the lemma itself but does not describe the word by any other semantic relation. Paraphrase definitions in DDO mostly fall into two categories. The first concerns the definitions of relatively rare lemmas. This is easily handled in the DanNet encoding process by simply inserting the lemma into the synset of the synonyms from the definition. The other concerns paraphrase definitions for words belonging to the very general top level of the DanNet hierarchy. Here the concepts have no hypernym and it is impossible to avoid circular definitions since more or less synonymous words are used to define one other. In such cases, the editor must decide on a top hypernym and attach other senses to it, either as members of the same synset or as its direct hyponyms. Examples of lemmas with circular definitions in DDO are *område* (area), *sted* (place) and *plads* (place).

In DDO, as well as in other monolingual dictionaries, another common way of defining concepts belonging to the general level of the language is to present the range or the extension of the concept by listing the typical hyponyms. For example, the concept *garment* belongs to the general level, as opposed to the basic-level concept *trousers* and the specific-level concept *jeans* (Dirven and Verspoor 1998). In Fig. 13, *kunstværk* (work of art) is partly described this way by the phrase 'e.g. a painting, a sculpture, a ballet or a piece of literature'.

Another problem in the use of the definitions from DDO for specifying DanNet concepts concerns the common use of ad hoc compounds with a compositional meaning as genus proximum. We estimate that 10% of the approx. 8,000 different nouns used as genus expressions in DDO definitions are ad hoc compounds. Some examples are: *kulthandling* (cult act), *hovedzone* (main zone), and *tidsafsnit* (time section), compounds that are too rare in the language to occur as headwords in a dictionary. The human reader is capable of understanding the compounds although they are not common words, but DanNet has adopted the strategy of inserting the head of the compound as hypernym and expressing the semantics of the other part of the compound as a relation. For example, a word defined in DDO by the genus proximum *kulthandling* (cult act) is given the hypernym *handling* (act) and the semantic relation CONCERNS*: kult* (cult) in DanNet.

6 Evaluation issues

6.1 Manually assigned versus inherited relations

The 41,000 synsets currently encoded in DanNet establish a total of 125,000 links to other synsets by means of semantic relations. A closer investigation of a subset of DanNet, namely 8,836 concrete artefact objects, gives an idea of the distribution of the different relation types in DanNet, both with regards to the manually encoded ones and those inherited. The 8,836 synsets include 56,638 relations (an average of 6.4 relations per synset) 47,802 of which are other relations than hyponymy. Of these, 38,859 (81%) are inherited relations while 8,943 (19%) are manually assigned. For the distribution of manually assigned versus inherited relations, see Figs. 15 and 16.

From the figures it can be seen that meronymy is the most frequent relation, both assigned manually and inherited (apart from hyponymy which is not in the figure) but also that the telic USED_FOR relation plays a crucial role in the description of artefacts. The USED_FOR relation has in most cases been directly deducible from the DDO definition, confirming its type-defining character. In contrast, the information in DanNet on the involved user of artefacts as well as on the many parts of more

Semantic relation	**Example**	**Number of times the relation is manually assigned**	**Percentage of 8,836 artefacts manually described with the relation**
meronymy (HAS_HOLO_PART, HAS_MERO_PART, HAS_HOLO_MEMBER, HAS_MERO_MEMBER, HAS_HOLO_MADE_OF, HAS_MERO_MADE_OF, HAS_HOLO_LOCATION, HAS_MERO_LOCATION)	*bog/side* (book/page); *side/bog* (page/book) etc.	2,529	29%
USED_FOR	*bog/læse* (book/to read)	2,417	27%
USED_FOR_OBJECT	*brødkniv/brød* (bread knife/bread)	1,780	20%
CONCERNS	*julepynt/jul* (christmas decoration/ christmas)	646	7%
INVOLVED_AGENT	*guitar/guitarist* (guitar/guitarist)	625	7%
MADE_BY	*tøj/at sy* (clothes/to sew)	363	4%
NEAR_SYNONYM	*bog/hæfte* (book/pamphlet)	211	2%
other relations (EXTRA HYPERNYM (206), ANTONYMY (9), INVOLVED PATIENT (57), INVOLVED INSTRUMENT (8), ENGLISH EQUIVALENT (53))		333	4%

Fig. 15 The distribution of the manually assigned relations in 8,836 artefact synsets

Semantic relation	Example of inherited relation	Number of times the relation is inherited	Percentage of the 38,859 inherited relations
MERONYMY (HAS_HOLO_PART, HAS_MERO_PART, HAS_HOLO_MEMBER, HAS_MERO_MEMBER, HAS_HOLO_MADE_OF, HAS_MERO_MADE_OF, HAS_MERO_LOCATION, HAS_MERO_LOCATION))	*sandwichbrød/mel* (sandwich bread/flour) *børnebog/side* (children's book/page)	14,027	36%
USED_FOR	*børnebog/at læse* (children's book/to read)	10,176	26%
MADE_BY	*undertøj/at sy* (underwear/to sew)	6,497	17%
INVOLVED_AGENT	*strengeinstrument/musiker* (stringed instrument/musician)	4,663	12%
USED_FOR_OBJECT	*kasserolle/fødemiddel* (saucepan/food)	2,094	5%
CONCERNS	*salmebog/sang* hymn book/singing	1,037	3%
other relations		365	1%

Fig. 16 The distribution of inherited relations in 8,836 artefact synsets

complex objects (e.g. books having pages, backs and titles) was not always found in DDO and has therefore been explicitly added in several cases, as already mentioned in Sect. 2.

The numbers of manual assignments of relations also indicate how often we find lexical connections between the different semantic relations of a synset. For example we find information on the INVOLVED_AGENT, the user of the object, in one out of four cases (7/27) of a manual assignment of the USED_FOR relation, and thereby the synset 'pilot licence' indirectly relates 'to fly' and 'pilot'. It is important to underline that other ontological types expose very different distribution patterns. For the ontological type HUMAN, for example, the relations ROLE_AGENT and ROLE_PATIENT play a crucial role, as well as features indicating negative/positive connotation and gender.

6.2 Data validation on DanNet samples

The DanNet data have been evaluated twice; the first evaluation was carried out after the completion of the hyponymy hierarchy and the second one after the addition of other relations. It was never the intention to evaluate all of the data, but in order to obtain a general picture it was decided to look into specific parts of the wordnet, e.g. instruments, physical substances, food, furniture. Thus, the evaluation could be characterised as a spot test in which 2% of the synsets have been checked.

The test was done by two persons who have not taken part in the DanNet encoding but who on the other hand have a large knowledge of the DDO dictionary data.

The overall impression of the spot test is that the established hyponymy hierarchy is relatively good and that the other relations add valuable information. As for coverage it could be noted that a lot of the synsets that were missing in the first evaluation had been added when the second evaluation took place. Within rapidly changing fields such as computing it is interesting (but not surprising) to observe that new terms are missing whereas already half-obsolete terms have been included. This is mostly due to the age of the corpus data from which the dictionary was compiled, dating from the nineteen-eighties and early nineties.

Another observation is that the coding is somewhat uneven, meaning that some synsets have very few relations and others have many. This can partly be explained by the fact that the synsets have reached different stages in the encoding process but also by differences in the encoders' practice of expanding features from the dictionary definition (cf. Sect. 5.3). The distribution of definition features on DanNet relations is not always transparent to an outside observer; especially in cases where the encoder has had several possibilities regarding choice of relation. For instance, the associative relation CONCERNS is sometimes used in cases where a more specific relation would seem more appropriate: for example, the synset *udgiver* (publisher) has the relation CONCERNS: *publikation* (publication). Since the synset also has the relation ROLE_AGENT: *udgive* (to publish) assigned to it, another probably more precise solution would have been to use the relation USED_FOR_OBJECT since *publikation* is the actual object of the verb *udgive*. Still, the CONCERNS relation adds a broader semantic content to the synset than the USED_FOR_OBJECT does.

A final conclusion emerging from the evaluation is the conflict that is likely to arise between a generalist taxonomy and a specialist taxonomy. In DanNet specific areas such as musical instruments have received special treatment which may pull them away from the general point of view in order to satisfy the specialist's needs. On the other hand, many other fields have been described according to a general approach originating in the dictionary description, intentionally written for non-specialists. A possible further development may consist in merging those two approaches, for instance by collaborating with terminologists within particular subject fields.

7 Conclusions

DanNet is compiled on a strongly monolingual basis in contrast to the approaches used in a large number of recently compiled wordnets of other languages [cf. among others Fernández-Montraveta et al. (2008) on the Spanish wordnet, Rodríguez et al. (2008) on the Arabic wordnet and Márton et al. (2008) on the Hungarian wordnet]. Our arguments for applying a monolingual approach to the Danish wordnet are partly linguistic, partly pragmatic, namely:

1. that we believe that a wordnet should ideally reflect the *inherent* characteristics of the general vocabulary of the language described, in this case Danish, in

order to constitute a really strong resource for NLP tools for that particular language,
2. that DDO constitutes an excellent source for our approach since it is corpus-based, i.e. it reflects contemporary Danish language use,
3. that access to DDO and the information in it is fairly straightforward, influenced by the fact that part of the DanNet group participated in the compilation of it.

This said, the use of a dictionary as the primary source for compiling a wordnet is not a widely used approach so there are only sparse experiences from other wordnet projects, the Polish wordnet (Derwojedowa et al. 2008) being one example of a wordnet project applying a monolingual approach similar to ours. Nevertheless, the clear conclusion from the work presented in this article is that extracting knowledge from a monolingual dictionary and reusing it in a wordnet is definitely worthwhile. In contrast to what was concluded more than a decade ago by Ide and Véronis (1995), DanNet shows that it is feasible to extract knowledge from a dictionary provided that its semantic information is well-structured and that automatic means of exploiting this information are applied with care. DanNet's approach of transferring, adjusting and supplementing lexical knowledge from a conventional dictionary seems much less labour-intensive than initiating a wordnet project from scratch. The fact that the sense inventory was established beforehand (and simplified in a systematic way in the wordnet framework), and that the genus proximum information in DDO definitions could be applied as the principal driving factor in the initial phase, helped speed up the compilation process radically. However, some word classes proved easier to transfer from dictionary to wordnet than others. As was shown, verbs turned out to be much more difficult to transfer than initially assumed and the amount of hierarchical reorganization of synsets proved much larger. Furthermore, it has been necessary to reorganize or make explicit a certain amount of underspecified semantic information in DDO when transferring it to DanNet. For instance, the so-called ISA overload has been handled by choosing a basic taxonomic scheme for each sub-hierarchy combined with an orthogonal feature on other hyponyms, and the human reader's implicit lexical knowledge about many artefacts has been made explicit by means of semantic relations.

A subset of DanNet is currently being tested as a means for text indexation in a text retrieval environment. The enhanced and more explicit semantic description introduced in DanNet during the transferral from dictionary to wordnet will also be utilized in an online version of DDO which will provide onomasiological access to the vocabulary and allow for more systematic navigation and querying. This will hopefully bring new insights into the synergy between wordnets, dictionaries and electronic publishing.

References

Agirre, E., Ansa, O., Arregi, X., Artola, X., Díaz de Ilarraza, A., Lersundi, M., et al. (2000). Extraction of semantic relations from a Basque monolingual dictionary using constraint grammar. In *Proceedings from the ninth Euralex international congress* (pp. 639–640). Universität Stuttgart.

Asmussen, J. (2007). Korpuslinguistische Verfahren zur Optimierung lexikalisch-semantischer Beschreibungen. In W. Kallmeyer & G. Zifonun (Eds.), *Jahrbuch des Instituts für Deutsche Sprache 2006* (pp. 123–151). Berlin and New York: Walter de Gruyter.

Boguraev, B., & Briscoe, T. (Eds.). (1989). *Computational lexicography for natural language processing*. London and New York: Longman.

Church, K., & Hanks, P. (1989). Word association norms, mutual information and lexicography. In *ACL proceedings, 27th annual meeting*, Vancouver.

Cruse, D. A. (1991). *Lexical semantics*. Cambridge: Cambridge University Press.

Cruse, D. A. (2002). Hyponymy and its varieties. In R. Green, C. A. Bean, & S. H. Myaeng (Eds.), *The semantics of relationships: An interdisciplinary perspective, information science and knowledge management* (pp. 2–21). Springer.

DDO = Hjorth, E., Kristensen, K., et al. (Eds.). (2003–2005). *Den Danske Ordbog 1–6* (*'The Danish dictionary 1–6'*). Copenhagen: Gyldendal and Society for Danish Language and Literature.

Derwojedowa, M., Piasecki, M., Szpakowicz, S., Zawislawska, M., & Broda, B. (2008). Words, concepts and relations in the construction of the polish WordNet. In *Global WordNet Conference 2008* (pp. 162–177). Szeged, Hungary.

Dirven, R., & Verspoor, M. (Eds.). (1998). *Cognititive exploration of language and linguistics*. Amsterdam/Philadelphia: John Benjamins.

Dunning, T. (1994). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19*(1), 61–74.

Fellbaum, C. (2002). Parallel hierarchies in the verb lexicon. In *Proceedings of the OntoLex workshop, LREC* (pp. 27–31). Las Palmas, Spain.

Fernández-Montraveta, A., Vázquez, G., & Fellbaum, C. (2008). The Spanish version of WordNet 3.0. Text resources and lexical knowledge. In *Text, translation, computational processing* (pp. 175–182). Berlin and New York: Mouton de Gruyter.

Fillmore, C. J., Johnson, C. R., & Petruck, M. R. L. (2003). Background to FrameNet. *International Journal of Lexicography, 16*(3), 235–250 (Oxford: Oxford University Press).

Fontenelle, T. (1997). Using a bilingual dictionary to create semantic networks. *International Journal of Lexicography, 10*(4), 275–303.

Guarino, N. (1998). Some ontological principles for designing upper level lexical resources. In *Proceedings from the first international conference on language resources and evaluation* (pp. 527–534). Granada.

Guarino, N., & Welty, C. (2002). Identity and subsumption. In R. Green, C. A. Bean & S. H. Myaeng (Eds.), *The semantics of relationships: An interdisciplinary perspective, information science and knowledge management*. Springer.

Huang, C., Hsiao, P., Su, I., & Ke, X. (2008). Paranymy: Enriching ontological knowledge in WordNets. In *Proceedings of the fourth global WordNet conference* (pp. 221–228). Szeged, Hungary.

Ide, N., & Véronis, J. (1995). Knowledge extraction from machine-readable dictionaries: An evaluation. In P. Steffens (Ed.), *Machine translation and the lexicon*, *third international EAMT workshop, Heidelberg, April 26–28, 1993, proceedings*. Lecture Notes in Computer Science 898, Springer.

Ide, N., & Wilks, Y. (2007). Making sense about senses. In E. Agirre & P. Edmonds (Eds.), *Word sense disambiguation—Algorithms and applications.* Springer.

Jackson, H. (2002). *Lexicography: An introduction*. London: Routledge.

Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities, 31*(2), 91–113.

Kokkinakis, D., Toporowska Gronostaj, M., & Warmenius, K. (2000). Annotating, disambiguating & automatically extending the coverage of the Swedish SIMPLE lexicon. In *Proceedings from the second international conference on language resources and evaluation* (pp. 1397–1405). Athens.

Lenci, A., Bel, N., Busa, F., Calzolari, N., Gola, E., Monachini, M., et al. (2000). SIMPLE—A general framework for the development of multilingual lexicons. *International Journal of Lexicography, 13*(4), 249–263.

Levin, B. (1993). *English verb classes and alternations—A preliminary investigation*. Chicago: The University of Chicago Press.

Lorentzen, H. (2004). The Danish dictionary at large: Presentation, problems and perspectives. In G. Williams & S. Vessier (Eds.), *Proceedings of the eleventh EURALEX international congress* (pp. 285–294). Lorient, France.

Lyons, J. (1977). *Semantics*. Cambridge: Press Syndicate of the University of Cambridge.

Márton, M., Hatvani, C., Kuti, J., Szarvas, G., Csirik, J., Prószéky, G., et al. (2008). Methods and results of the Hungarian WordNet project. In *Proceedings of the fourth global WordNet conference* (pp. 311–320). Szeged, Hungary.

Miller, G. A. (1998). Nouns in WordNet. In C. Fellbaum (Ed.), *WordNet—An electronic lexical database* (pp. 23–47). Cambridge, London: The MIT Press.

Norling-Christensen, O., & Asmussen, J. (1998). The corpus of the Danish dictionary. *Lexikos. Afrilex Series, 8*, 223–242.

Pedersen, B. S., & Nimb, S. (2000). Semantic encoding of Danish verbs in SIMPLE—Adapting a verb-framed model to a satellite-framed language. In *Proceedings from the second international conference on language resources and evaluation* (pp. 1405–1412), Language resources and evaluation—LREC 2000, Athens.

Pedersen, B. S., & Paggio, P. (2004). The Danish SIMPLE lexicon and its application in content-based querying. *Nordic Journal of Linguistics, 27*(1), 97–127.

Pedersen, B. S., & Sørensen, N. H. (2006). Towards sounder taxonomies in wordnets. In A. Oltramari Chu-Ren Huang, A. Lenci, P. Buuitelaar, & C. Fellbaum (Eds.). *Ontolex 2006 at 5th international conference on language resources and evaluation* (pp. 9–16), Genova, Italy.

Pedersen, B. S., Braasch, A., Henriksen, L., Olsen, S., & Povlsen, C. (2008). Merging a syntactic resource with a WordNet: A feasibility study of a merge between STO and DanNet. In *Proceedings from the sixth international conference on language resources and evaluation*, Marrakech, Morocco.

Pustejovsky, J. (1995). *The generative lexicon*. Cambridge, MA: The MIT Press.

Rigau, G., & Agirre, E. (2002). Semi-automatic methods for WordNet construction. In *Tutorial at 2002 international WordNet conference*, Mysore, India.

Rodríguez, H., Farwell, D., Farreres, J., Bertran, M., Alkhalifa, M., Martí, M. A., et al. (2008). Arabic WordNet: Current state and future extension. In *Proceedings of the fourth global WordNet conference* (pp. 387–405). Szeged, Hungary.

Ruus, H. (1995). *Danske kerneord*. Copenhagen: Museum Tusculanums Forlag.

Svensén, B. (1993). *Practical lexicography. Principles and methods of dictionary-making*. Oxford: Oxford University Press [translated from the Swedish Handbok i lexikografi (1987) by Sykes, J. & Schofield, K.].

Veale, T., & Hao, Y. (2008). Enriching WordNet with folk knowledge and stereotypes. In *Proceedings of the fourth global WordNet conference*, Szeged, Hungary.

Vossen, P. (Ed.). (1999). *EuroWordNet, a multilingual database with lexical semantic networks*. The Netherlands: Kluwer.

Vossen, P., Maks, I., Segers, R., & van der Vliet, H. (2008). Integrating lexical units, synsets and ontology in the Cornetto database. In *Proceedings from the 6th international conference on language resources and evaluation, language resources and evaluation—LREC 2008*, Marrakech, Morocco.

Zgusta, L. (1988). Pragmatics, lexicography and dictionaries of English. *World Englishes, 7*(3), 243–253.