

Use, Re-use and Synergetic Benefit: The Interplay between Wordnet and Dictionary Data

1 Introduction

The first version of a Danish wordnet, DanNet, was released in March 2009. In this paper we will discuss some of the methodological issues of compiling a wordnet on the basis of a large, corpus-based printed dictionary of modern Danish (Den Danske Ordbog, henceforth DDO). Furthermore, we will focus on the synergetic effects between dictionary and wordnet, the first use of DanNet being an onomasiological search engine in the online version of the very same dictionary which was used as the basis for the establishment of the wordnet. The online version of DDO was also published in summer 2009 and can be seen at www.ordnet.dk/ddo.

After a short presentation of some basic details about the DanNet project we will describe how the definitions as well as the information on hypernymy in DDO were translated into wordnet relations in DanNet. We will present some typical cases where we decided to encode a semantic feature or relation in DanNet even though, for various reasons, it was underspecified in DDO. Finally, we will show how the wordnet data have contributed to new search possibilities in the online version of DDO.

2 DanNet – background and structure

The Danish wordnet project DanNet is a joint work between two institutions, the Centre for Language Technology (CST) at Copenhagen University, which compiled a pilot version of the computational semantic lexicon SIMPLE-DK for Danish, (Pedersen and Paggio (2004)), and the Society for Danish Language and Literature (DSL), which was responsible for DDO, the dictionary used as the starting point for DanNet (DDO (2005)).¹

The first version of DanNet contains approximately 40,000 synsets described with hyponymy relations. A large subset of these, approx. 12,000 synsets describing concrete objects or human beings are fully described with a richer set of semantic relations such as meronymy, near-synonymy and antonymy, and in the case of artefacts also relations describing origin, purpose as well as agents and instruments involved in the use of the artefact.

¹The initial four year phase (2005-2008), which was concluded with the launch of the first version of DanNet, was funded by the Danish Research Council (approx. €400,000), and an additional funding by the same council of approx. €130,000 has secured an extension of the wordnet of more than 25,000 synsets as well as links of some 8,000 base concepts to Princeton WordNet, to be ready by the end of 2010.

The wordnet was established on purely monolingual grounds and not, as is the case for many other wordnets, by translating synonym sets from Princeton WordNet to Danish. An important factor in the choice of method – the so-called merge approach – was the fact that a corpus-based dictionary of Danish had been completed in 2005 and was available in a machine-readable version with hypernymy information explicitly specified for each of the approximately 100,000 sense definitions. Thereby, a wordnet for Danish could be semi-automatically built on a well-consolidated sense distinction basis, with sense division and definitions based on corpus evidence, ensuring both a more loyal picture of linguistic conceptualization in Danish and a better sense coverage when it comes to the future computational treatment of Danish text material by the use of DanNet.

3 Use: Danish monolingual dictionary → wordnet for Danish

Asmussen et al. (2007) gives a detailed description of the microstructure of DDO and of the information extraction from the dictionary to DanNet. The article focuses on the semi-automatic generation of the wordnet's hyponymy structure and on the exploration of automatic procedures for assigning other semantic relations, such as meronymy and holonymy, mainly based on semantic patterns extracted from the definitions.

Centrally involved in the generation of a hyponymy structure built on dictionary data was the general and sometimes challenging task of harmonising the raw, automatically extracted hyponymy structure that was derived directly from the genus proximum information contained in DDO. Among the problematic cases was the arbitrary choice of genus proximum in DDO in case of synonymous alternatives. The task of disambiguating cases of homonymy and polysemy was described in Asmussen et al. (2007), but the conclusion was that the compilation of the hyponym hierarchy was indeed facilitated by the utilization of genus proximum information available in the DDO. Finally, it was argued that in order to ensure the practical use of wordnets as resources in formal ontologies, one should be able to separate the so-called taxonomical hyponyms, e.g. for the concept 'tree' the different kinds of trees such as *birch*, *oak* and *cherry tree*, from those hyponyms that are not taxonomical kinds but instead denote, say, a functional aspect as in the case of *roadside tree* or *climbing tree*. This means that an *orthogonal*, i.e. non-taxonomical, feature is added to the hyponyms in DanNet. This information is not directly deducible from the data in DDO which does not, as it is customary in a semasiological dictionary, include information about different groups of hyponyms. Apart from having a more precise set of hyponymy relations than DDO, DanNet also contains information about hyponym categories. See also Pedersen and Sørensen (2006), Pedersen and Nimb (2008) and Pedersen et al. (2009) for further discussion on the subject.

With respect to the automatic extraction of semantic relations, Asmussen et al. (2007) concluded that due to the lack of obligatory structural templates for definitions in DDO regulating both the defining vocabulary and the grammatical expressions, it was not possible to automatically extract semantic patterns on the basis of definitions. In order

to lend itself to automatic extraction, definitions would have to be constructed in a more consistent and predictable way than it was done in DDO, with an explicitly defined syntax connecting certain syntactic patterns with semantic relations, it was claimed. Furthermore, it added to the problem of automatic extraction that no specifications had been prepared to determine which types of features or relations should obligatorily apply to which types of words in the dictionary.

We will discuss how these factors may also cause problems when the DDO definitions are manually translated into wordnet relations, as well as other problems deriving from the use of a dictionary as a lexical knowledge basis for a wordnet.

4 Semantic Relations in DanNet

As described in Asmussen et al. (2007), the definitions of some groups of words in DDO, e.g. in the cases of meals, cakes etc., cover the semantics that we estimate to be relevant in a wordnet and the translation from dictionary definition to semantic relations in DanNet is fairly straightforward in these cases. But as the wordnet compilation proceeded, it became clear that DDO sometimes falls short of the semantic requirements needed in a wordnet to be used for information retrieval, text understanding etc. As a starting point, we will present an overview of the set of semantic relations used in DanNet.

The set of semantic relations in DanNet are identical with the standard WordNet relations, with a few but, in our view, important extensions taken from the Danish SIMPLE project. As opposed to other wordnet models, the SIMPLE model (Lenci et al. (2000)), organized the semantic relations according to the four so-called qualia roles (Pustejovsky (1995)), which relate to inheritance structure, origin, composition and purpose, respectively. None of the standard WordNet relations cover the origin dimension, and the same is true of the purpose dimension of a concept. Our experience from the work on the Danish SIMPLE lexicon (Pedersen and Paggio (2004)) was that all four dimensions of the qualia structure were needed to describe a concept and in fact it was the only effective way to ensure coverage of a maximum number of word sense nuances in the encoding process. Therefore, it was decided to organise the set of relations in the same way in DanNet. So in DanNet, the SIMPLE relations `MADE_BY` and `USED_FOR` (telic relations) and the more flexible relation `CONCERNS` (constitutive relation) were added to the sets traditionally used by wordnets. Finally, we added a `USED_FOR_OBJECT` relation, used to describe the object of the `USED_FOR` relation. An example is *brødkniv* 'bread knife' which has the `USED_FOR` relation *skære* 'cut' and the `USED_FOR_OBJECT` relation *brød* 'bread'. See Table 1.

The four-sided organisation of the relations was very useful in the establishment of templates for each ontological type in DanNet. E.g. the template type `[ARTIFACT+OBJECT]` would contain the `MADE_BY` relation as well as the `USED_FOR` relation since these will always be relevant in the semantic description of artefacts: they are always man-made and they are always made to be used for a purpose. The interface editor used in the encoding of DanNet produces a synset on the basis of templates. Once the editor has

Formal Role (INHERITANCE)	Agentive Role (ORIGIN)	Constitutive Role (COMPOSITION)	Telic Role (PURPOSE)
HAS_HYPERNYM	MADE_BY (SIMPLE)	HAS_HOLO_MADE_OF	USED_FOR (SIMPLE)
HAS_HYPONYMS		HAS_HOLO_PART	USED_FOR_OBJECT (DANNET)
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 (SIMPLE)	
		INVOLVED_AGENT	
		INVOLVED_PATIENT	
		INVOLVED_INSTRUMENT	

Table 1: Semantic relations in DanNet

decided on the ontological type of the concept to be encoded, the tool establishes a synset containing the relations that are relevant for the ontological type in question. In this way, the initial creation of each template encompasses the specifications for those relations that are obligatory for a certain ontological type of words in DanNet. This is the type of specifications that were missing in the dictionary-making process of DDO and one of the main reasons why an automatic extraction of relations was difficult to carry out (Asmussen et al. (2007)).

Furthermore, the encoding process in DanNet begins with the linguistic top hypernyms, e.g. concepts like *bog* ('book'), *legetøj* ('toy') or *beklædningsgenstand* ('garment') with a maximum of semantic relations (see Table 2). In that way, the top hypernym also comes to function as a kind of specification for all types of books, toys and garment, the inheritance mechanism of the DanNet interface ensuring that the full set of relations inherited from the top is visible for all hyponym synsets which are to be described. The job of the lexicographer is to adjust the relations according to the hyponym in question, based on the definition in DDO. E.g. for the concept *kogebog* ('cookery book'), in DDO defined as (in translation): "book containing recipes and sometimes instructions for cooking", the relation inherited from *bog* ('book') CONCERNS: *emne* ('subject') is changed to CONCERNS: *mad* ('food'). Furthermore, a CONCERNS relation is added to the inherited one: CONCERNS: *madlavning* ('cooking'), as well as the relation HAS_MERO_PART: *madopskrift* ('recipe').

In DanNet, a systematic top-down working process is employed, beginning with the two sets of specifications for obligatory relations for various types of words: 1. the templates defined for each ontological type, and 2. the complete sets of relations defined at the top hypernym level and subsequently inherited by all the hyponyms.

	bog (book)	legetøj (toy)	beklædningsgenstand (item of clothing)
Ontological type	LANGUAGE REPRESENTATION + ARTIFACT + OBJECT	ARTIFACT + OBJECT	GARMENT + ARTIFACT + OBJECT
Formal role INHERITANCE	HAS_HYPERNYM: genstand ('object')	HAS_HYPERNYM: genstand ('object')	HAS_HYPERNYM: genstand ('object')
Agentive role ORIGIN	MADE_BY: skrive (‘write’), trykke (‘print’)	MADE_BY: fremstille (‘produce’)	MADE_BY: sy (‘to sew’)
Constitutive role COMPOSITION	HAS_MERO_MADE_OF: papir (‘paper’) HAS_MERO_PART: tekst (‘text’), side (‘page’), ryg (‘back’), titel (‘title’) CONCERNS: emne (‘subject’) INVOLVED_AGENT: forfatter (‘writer’), læser (‘reader’)	INVOLVED_AGENT: barn (‘child’) HAS_MERO_MADE_OF: materiale (‘material’)	HAS_HOLO_PART: påklædning (‘dressing’, ‘clothes’) HAS_HOLO_LOCATION: kropsdel (‘bodypart’) HAS_MERO_MADE_OF: stof (‘fabric’) INVOLVED_AGENT: person (‘person’)
Telic role PURPOSE	USED_FOR: læse (‘to read’)	USED_FOR: lege (‘to play’) USED_FOR_OBJECT: leg (‘game’)	USED_FOR: klæde (‘to dress’) USED_FOR_OBJECT: person (‘person’)

Tabelle 2: Relations on top hypernyms *bog* (‘book’), *legetøj* (‘toy’) and *beklædningsgenstand* (‘item of clothing’)

This approach is opposite to the bottom-up process adopted in DDO where the main purpose was to present well-formed definitions with a fairly simple syntax intended for a human reader. In DanNet, on the other hand, the aim is to describe, explicitly and as accurately as possible, the semantics of a linguistic item (a sense) in terms of relations to other concepts, taking into consideration the computer programs' complete lack of knowledge of how to use the dictionary. Veale and Hao (2008) claim that not even the kind of knowledge embodied in dictionaries covers what is needed to make a computer understand everyday language. It is argued that wordnets should be enriched with information on stereotypes and culturally inherited associations, e.g. that snakes are related to treachery and slipperiness and that elephants have a good memory, in order to make this possible. This clearly lies outside the scope of DanNet at its current stage. Our aim in DanNet has been to reach an information level defined as 'the native speaker's lexical knowledge about a concept', focusing on the prototypical semantic aspects. In that respect, the ambition regarding information level does not differ from the goal of DDO. The real difference between the two types of lexical resources lies in the fact that a dictionary definition leans heavily on the reader's ability to make assumptions without any explicit statements in the text (Svensén (1993)). Human readers constantly make assumptions and use their knowledge of the world, and compilers of dictionaries base their definitions on this, whether they are aware of it or not, quite contrary to the case of wordnet compilers who must be careful to avoid inferences of any kind and describe everything explicitly in their encodings. Another difference is the syntactic limits of the type of dictionary definition chosen in DDO: most definitions consist of one well-formed, not too complex phrase aiming to capture the typical usage. This style was deliberately preferred over the definitional style of its predecessor, the Danish monolingual dictionary *Ordbog over det danske Sprog* 1918-1956 (ODS (1956)), whose definitions are often quite complex, with frequent use of subordination, parenthetical elements and interposed reservations, alternatives etc. Thirdly, one should keep in mind that a dictionary's sense description is not necessarily confined to the definition only, but may be distributed over several elements. This means that the user will have to read collocations, valency patterns, citations and other relevant data in order to grasp the semantic description in its entirety.

It is therefore hardly surprising that we sometimes find a discrepancy between the actual number of relations described in the definition of a word, and the number of relations which from a systematic point of view should be described for a given word in DDO in order to reflect the native speaker's lexical knowledge.

Bearing this in mind, we will now turn to discussing some typical instances where DDO does not contain sufficient information to specify all the semantic relations required by a certain type of synset in DanNet.

5 'Missing' lexical information in DDO: reasons and general tendencies

In the case of *bog* ('book'), Table 2 lists a series of semantic relations which are considered relevant for the description of the synset in DanNet. Now, compare the list with (the

translation of) the definition in DDO: “printed or written sheets of paper bound or fastened together so as to form a whole, often a coherent text intended for reading”. Although probably acceptable to most readers, it is nevertheless striking that the definition says nothing about the writer, nor about the reader, the title and subject of the book. There is no mention of a back, and instead of the common word “page”, it has “sheets of paper”. These semantic features have all been added in DanNet as they are considered central semantic aspects in the description of *bog* (‘book’) and its many hyponyms. In fact, the process often turns out to be recursive as relations are added in a series of repeated steps until the final number is fixed at the top level hypernym. The reason is, of course, that it may not be discovered that a relation is relevant until a hyponym is described at a lower, more specific level. A case in point is the CONCERNS relation of *bog* (‘book’). Not until the many hyponyms of *bog* were considered, did it become clear that the subject is a central semantic aspect of the concept even though it might not, initially, seem crucial for the concept ‘bog’ itself: *Koranen* (‘the Koran’) CONCERNS: islam; *fuglebog* (‘bird book’) CONCERNS: bird, *kogebog* (‘cooking book’) CONCERNS: cooking, *kriminalroman* (‘crime novel’) CONCERNS: crime, etc.

The definitional style of using single well-formed sentences is obviously the main reason why DDO has not been able to include all semantic aspects of the quite complex concept of ‘book’. The wordnet model includes much more detailed information than found in DDO and allows many types of relations that may even be used more than once if needed. The same is also true of less complex definitions. For example, for some hyponyms of *bog* we find DDO definitions where central semantic aspects are neglected, probably due to the lack of specifications for the type of words in question: in the definition of *salmebog* (‘hymn book’), “book which contains a selection of hymns”, nothing is said about the typical user (the church goer), or the typical use (to be sung during a church service). Yet much of the information appears in the entry *salme* (‘hymn’). Or consider the example *letlæsningsbog* (‘easy reader’) where the DDO definition (“book with a typography and a language style that are adapted to the user’s low level of reading proficiency”) makes no mention of the typical user: a pupil. In this case the information does appear, but only indirectly, in the citation. Another example is the word *butik* ‘shop’, defined as “room or building where a tradesman displays and sells products”. Next to the definition, and without explanation, are a number of collocations, among others *se på butikker* (literally ‘to look at shops’, i.e. ‘to do window shopping’) and a citation which translates: “We walked down the pedestrian street Strøget. Mona stopped in front of almost every shop. She loved looking at clothes”. The example shows how the dictionary relies on knowledge which is not made explicit in the editorial text. Neither the collocation nor the citation would make sense were it not for the fact that human users know that shops have windows with goods on display. This fact is simply implied in DDO as the definition of ‘butik’ says nothing about it. A similar case is *indlæggelsessedel* (‘referral note’) defined as “document issued by a doctor prescribing hospitalization”, and supplied with the citation “The doctor gave Marie a referral note to the sanatorium”. Here we find nothing in the definition about the person involved,

i.e. the patient. The human dictionary user has no problem in deducing that Marie is a patient, but in DanNet we must add this explicitly, which is why the concept has been encoded with the relations: HAS_HYPERNYM: document, MADE_BY: to issue, INVOLVED_AGENT: doctor, INVOLVED_PATIENT: patient, USED_FOR: hospitalization.

The process of adding full sets of semantic relations to approximately 6,800 object artefacts in DanNet has revealed that the telic role, the USED_FOR relation, is usually described in DDO. This has confirmed us in our decision to add the USED_FOR relation from SIMPLE to the standard WordNet set of relations. This role is centrally involved in the description of the semantics of artefacts as already pointed out in the SIMPLE project. Another general tendency, especially in the case of complex concepts, is that the definitions often lack information about the parts of the object (book: back, page, shop: display window), even when there is a close lexical, or indeed morphological, link between the parts and the whole. On the other hand, we suspect that the definitions of the parts contain information about the whole more often than vice versa, but at this point it remains a hypothesis as this group of words has not yet been supplied with the full set of semantic relations in DanNet. The main reason for not incorporating all parts of an object in DDO is primarily the demand for well-formed dictionary definitions. Especially in the case of complex objects, too many phrases are needed to provide a comprehensive description.

Finally, we have found a strong tendency not to mention the typical user of an artefact object in DDO (e.g. easy reader: pupil, hymn book: church goer), even in cases of morphological relationship. This could be taken to suggest that the typical user – or producer – of an artefact is not central to the understanding of an object. Instead, the user may be more closely connected to the verb describing the act of use. However that may be, it is interesting that the artefact is often morphologically closely related to the user: shop/shopkeeper/shopper, pharmacy/pharmacist, bakery/baker, pilot licence/pilot. Table 3 shows some examples of typical users, added as relations in DanNet. In all cases, this information is missing in DDO.

To sum up, compared to the information in DDO, DanNet has been extended with highly structured data on hyponymy relations as well as on the type of hyponymy relation (taxonomical or non-taxonomical). Furthermore, a large number of semantic relations that are not mentioned in the definitions of DDO have been added in the case of artefacts, especially information about the parts of the artefact and about the typical user. In a long perspective, the enriched wordnet may in turn be utilized to improve search facilities in an online version of DDO. The next chapter will describe how the first version of the online DDO makes use of the DanNet data to present onomasiological information.

6 DanNet data and the dictionary

It is not a new idea to use wordnet data for human users to present onomasiological information. Various visual representations of the Princeton WordNet are available on the net, such as www.visualthesaurus.com, www.thefreedictionary.com and thesaurus.com.

Synset	Added in DanNet as compared to DDO
<i>bog</i> ('book')	INVOLVED_AGENT: forfatter ('writer') INVOLVED_AGENT: læser ('reader')
<i>flyvecertificat</i> ('pilot license')	INVOLVED_AGENT: pilot ('pilot')
<i>briller</i> ('glasses')	INVOLVED_AGENT: person ('person')
<i>forskningsbibliotek</i> ('research library')	INVOLVED_AGENT: forsker ('researcher')
<i>læbestift</i> ('lipstick')	INVOLVED_AGENT: kvinde ('woman')
<i>barberkost</i> ('shaving brush')	INVOLVED_AGENT: mand ('man')
<i>ægteskab</i> ('marriage')	INVOLVED_AGENT: ægtepar ('married couple')
<i>apotek</i> ('pharmacy')	INVOLVED_AGENT: apoteker ('pharmacist')
<i>bageri</i> ('bakery')	INVOLVED_AGENT: bager ('baker')
<i>registreringsattest</i> ('vehicle registration certificate')	INVOLVED_AGENT: motorkontor ('motoring office')

Table 3: Relations added in DanNet

reference.com. Some of these also offer thesaurus information in combination with dictionary data from one or more dictionaries. A case in point is TheFreeDictionary which has thesaurus information from two sources, Princeton WordNet 3.0 and Collins Essential Thesaurus, as well as dictionary information extracted from The American Heritage Dictionary of the English Language, Collins Essential English Dictionary, bilingual learner’s dictionaries from Kernermann Publishing, and various lists of technical terms. The solution is a portal-like presentation where one query is performed in different lexical resources simultaneously and the result is shown as a sequence of adjacent matches.

In the wordnet world, DanNet is unique in that the encoded relations are so closely connected with the dictionary data of DDO. As we have seen, most of the encoding task of the DanNet editors consists of extracting information from the dictionary articles and making the relevant relations explicit, whether they are already expressed directly in the articles, or can be deduced by the human user. In this perspective, the two can be viewed as one combined lexical resource from which both dictionary and wordnet data can be drawn and shown in a user interface. Rather than showing the results of simultaneous queries of two databases, it is our aim to provide one integrated access that offers a choice between a semasiological and an onomasiological presentation of the same underlying data. Let’s turn to see how this works in practice.

Figure 1 and 2 show extracts from a DDO dictionary article from DDO as it looks in a prototype version of the user interface. The use of DanNet information becomes

Den Danske Ordbog

Søg efter:

- Betydninger
- Begreber

Om Den Danske Ordbog

- Hjælp til søgning
- Fakta om DDO

Relaterede søgninger

Begreber

skade¹ subst
Uofficiel stavemåge

Overblik

BØJNING

-set (el. -sen) el.
-sene **i**

Abbildung 1: Section of the interface with semasiological and onomasiological search option

skade¹ substantiv, fælleskøn

Overblik **+**

BØJNING -n, -r, -rne

UDTALE ['sgæ:ðə] **i**

Betydninger **-**

1. fysisk forringelse af eller mangel ved noget, fremkaldt af fx vejrlig, brand, vold eller manglende vedligeholdelse

SYNONYMER defelt | fejl SE OGSÅ beskædigelse | ødelæggelse

BESLÆGTEDE ORD **BEFY** mere generelt: forringelse || mere specifikt: arbejdsskade | brandskade | bræklage | erhvervs sygdom | forbrænding | friskskade | frostskeade | idrætsskade | inseltangreb | krakelering | kvæstelse | men | miljøskade | personsskade | ravage | rørskeade | skadevirkning | skramme | skramme | tingsskade ...vis 22 flere ...skjul **i**

EKSEMPLER ubodelig skade **☒** | uoprettelige skader **☒** | materielle skader **☒** || forvolde skade **☒** | udbedre skaderne **☒**

Branden skyldtes kortslutning og skaden på bygninger er opgjort til 1,6 mio. kroner **Aarh5t87**

Abbildung 2: Section of the interface showing 'related words'

apparent at various places in the article. First, as a quite general possibility, the user has the option of selecting either a semasiological or an onomasiological search mode, displayed in the upper left corner. The traditional (semasiological) view of the dictionary article is selected as default.

Secondly, the dictionary article has been supplied with a new element. In addition to the word relations *SYNONYM*, *ANTONYM* and *CONFER*, already present in the (printed) dictionary, there is a new element *RELATED WORDS FOR*. Here we have automatically extracted and displayed the hypernym, hyponyms, and co-hyponyms, i.e. synsets sharing the same hypernym as the sense described (in this case the hypernym for *damage* is ‘deterioration’), on the assumption that sister senses, although clearly not synonyms, are nevertheless relevant for a user seeking help to produce a text as they often have the same paradigmatic properties as the entry word in a given sense. The information is more helpful for native speakers than for learners of Danish as no explanation is provided to distinguish the words on the list. For practical reasons, only a limited number of co-hyponyms are shown, but the full list unfolds when it is clicked.

Finally, the user can move freely between the two presentational modes. If the user clicks on the icon to the right of the definition (the button with the letter ‘B’ inside – *b* alluding to *begreb* ‘concept’), the interface changes to the onomasiological view of the sense. Clicking on the icon is equivalent to performing an onomasiological search for that word and subsequently selecting the relevant sense from the search result presented in the right column. The outcome of an onomasiological search for *bil* ‘car’ can be seen in Figure 3. Notice that the search result is the entire synset, not just the synset member *bil* alone. The search term is highlighted, however, and after a definition of the synset (taken from the dictionary) the six members of the synset are listed, each member being clickable if the user wants to view the corresponding dictionary entries.

The main focus of attention is on the hyponym hierarchy, partly for the practical reason that this part of DanNet is the most thoroughly developed, but more importantly because we believe it to be very useful to the human user, given that the most relevant other word relations (synonyms, near-synonyms, antonyms) may be imported directly into the dictionary article and shown in the traditional view. The idea behind the visual presentation is to show the hierarchy in relation to the chosen synset. This synset is displayed as the basic level centrally on the screen and at this point only the levels immediately next to the basic level are visible, in upward and downward direction respectively. The interface solution represents one answer to the conflict between two incompatible ambitions: an intention to show all the details of the hyponym hierarchy, and the wish to help the user keep a sense of orientation and overview.

The same kind of compromise between detail and overview is also the cause of some of the other features that are used. For example, some synsets have a large number of co-hyponyms, and for that reason only a limited number is shown, but the user can choose to see them all with a click. Moreover, as default, the subordinate level is not directly visible because one or more of the co-hyponyms shown can have substantial numbers of hyponyms, thereby causing the users to lose track of where they are in the hierarchy. Instead, they are given the option to click on one of the arrow buttons to

?

[automobil 1 | bil 1 | dyt 2.1 | karet 1.2 | slæde 1.3 | vogn 2]

DEFINITION firehjulet motorkøretøj til person- el. godstransport

ORD MED DENNE BETYDNING

- ▶ automobil (gammeldags; især formelt el. spøgende)
- ▶ bil
- ▶ dyt (slang)
- ▶ karet (slang)
- ▶ slæde (slang)
- ▶ vogn (især gammeldags)

Betydningstræk Vis mere +

Begrebshierarki i

...

[motorkøretøj 1] ->

GRUPPER EFTER [FORMÅL](#) | [DELE](#) | [FREMSTILLING](#) | [BEGREBSTYPE](#)

[automobil 1 | bil 1 | dyt 2.1 | karet 1.2 | slæde 1.3 | vogn 2] -->

GRUPPER EFTER [FORMÅL](#) | [DELE](#) | [FREMSTILLING](#) | [BEGREBSTYPE](#)

- [fiskebil 1]
- [flyder 1]
- [forhjulstrækker 1]
- [generatorbil 1]
- [hurtigløber 1.1]
- [højtalerovogn 1]

Abbildung 3: Result of an onomasiological search for *bil* 'car'

unfold the subordinate level for that particular synset. The length of the arrows visually reflects the number of hyponyms: one dash indicates a small number, and two dashes a large number.

Even though only the next immediate level is visible, the user can move up and down in the hierarchy by clicking on a synset. Whenever a synset is selected, the interface updates to display that synset at the basic level and its corresponding immediate levels. To illustrate, we can look at *bil* ‘car’ again. If the user clicks on the hypernym *motorkøretøj* ‘motor vehicle’, the interface displays the synset *motorkøretøj* at the centre as the new basic level and shows the hypernym *køretøj* ‘vehicle’ at the level immediately above. In this way, users can navigate all the way to the topmost level and, similarly, to the most specific item in downward direction. To facilitate navigation and help the users maintain an overview, we have decided to show the entire hierarchy of ancestors in an independent section at the bottom of the central field. This gives the users a quick overview and allows them to jump to non-adjacent levels in the hierarchy should they wish to do so. Because of the frequent branchings at subordinate levels, it should be obvious that a corresponding overview of the hierarchy from the basic level downwards is not feasible.

Another feature, which we have introduced to help with the overview, can be seen immediately before the chosen synset in the hierarchy. Here the user has the option of grouping hyponyms together according to different criteria. This is particularly helpful when the number of hyponyms is large, as is the case for synsets such as *person*, *part* or *place* and *make*, *be* or *get*, which include hundreds or even thousands of synsets as their hyponyms. We have found it necessary to provide the users with some kind of meaningful subgrouping if they are not to become completely lost in the sheer quantity of details. At the same time, the grouping criteria must be of such a nature that they can be employed automatically. The final test deciding which parameters will be implemented has not been concluded at the moment of writing, but the most promising ones are a subset of the DanNet encodings: PURPOSE (the relation USED_FOR), PARTS (the relations HAS_MERO_PART and HAS_HOLO_PART), MANUFACTURE (the relation MADE_BY) and ontological type (ONTOTYPE). The assumption is that synsets which have the same encoding for one of the parameters are likely to have something conceptual in common, and therefore it is meaningful to group them together. However, it cannot be predicted universally which parameter is most meaningful; it varies from synset to synset and must be chosen for each individual lexical unit. Therefore, the grouping feature makes certain demands on the user’s ability to make reasonable judgments as they must select the most appropriate parameter themselves in order to profit from this feature.

Although we find the hyponom hierarchy more interesting for human users, we see no reason to deny the user access to the other types of encoding. Relations other than hypernyms/hyponyms can be seen under the heading *Betydningstræk* ‘features of meaning’ as shown in Figure 4.

It appears that the synset *bil* ‘car’ has the ontological type [MEANS OF TRANSPORTATION + ARTIFACT], and two relations have been encoded in addition to the hypernym.

[automobil | **bil** | dyt² | karet | slæde | vogn]

DEFINITION firehjulet motorkøretøj til person- el. godstransport

ORD MED DENNE BETYDNING

- ▶ automobil (gammeldags; især formelt el. spøgende)
- ▶ bil
- ▶ dyt (slang)
- ▶ karet (slang)
- ▶ slæde (slang)
- ▶ vogn (især gammeldags)

Betydningstræk		Luk -
Begrebstype	Transportmiddel+Artefakt	
HAR SOM OVERBEGREB	[motorkøretøj]	
HAR SOM DEL	[bilmotor motor]	
HAR SOM DEL	[hjul]	
BRUGES TIL	[transportere]	
	--> [gods]	
	--> [individ person]	
BRUGES TIL	[køre]	

Abbildung 4: Features of meaning

The shades of grey of the bars in Figure 4 correspond to different colours on the screen. The first is the holonymic relation PARTS where two parts have been encoded: a car has wheels and it has an engine. The USED_FOR relation also has a dual encoding: it is used for a) transportation or b) driving. The transportation purpose is further subdivided into i) goods and ii) persons or individuals.

In our opinion, information about the other relations is primarily requested by language specialists, and for that reason the panel is hidden in the default view but can easily be unfolded with a click. Notice that the relations can also be used as hints concerning the grouping of hyponyms. In the case of motor vehicles, it seems sensible to group synsets according to their part-whole relation: vehicles that have engines, vehicles with wheels etc. Likewise, the ontological type is often a sensible grouping criterion: The use of ontology as a grouping parameter will separate, for example, plants that are edible from those that are not for all hyponyms of 'plant, vegetable'. However, it may not always be easy for the user to realise it, but as a first step in that direction it helps to present the ontological type [PLANT+OBJECT+COMESTIBLE].

7 Future perspectives

In the first version of the interface, we have given priority to the details of the DanNet encodings and the user's possibility to move about in the hierarchy. Like other websites, however, we are also contemplating a visual presentation of the same data where several relations can be incorporated at a stroke, thereby facilitating the user's overview of a word's relations to other lexical units, as illustrated in Figure 5.

This idea is not new. It has been implemented in other interfaces using information from Princeton WordNet, for example in The Visual Thesaurus (www.visualthesaurus.com) which offers a presentation along similar lines. An example illustrating the same word is given in Figure 6.

Notice that The Visual Thesaurus draws on not only synset based information such as hyponyms, meronyms, synonyms, antonyms etc. (the type of relation becomes visible when you point at the relevant connecting line with the cursor), but also on word based information such as word meaning (*US liquid unit, plant organ*) and synonyms of meanings (of the noun: *cupful, hole, punch, incurvature*, as well as of the verb: *form, shape, transfuse, enclose* etc.). The idea is to help users find the right word, either in a text producing situation or in a learning situation, by showing both words and meanings that are related to the central lexical unit. According to the developers, the interface "works like the brain" and allows users to associate intuitively by using an interactive function: if the user clicks on a lexical unit, it is brought to the centre and new words and sense relations for that word appear. In this way, The Visual Thesaurus represents a development which is interesting for us to pursue in a future version as it meets a genuine user need (i.e. finding or learning the right word) through a combination of wordnet and dictionary data, data which are readily available in our resource.

Another approach which we find appealing is the thesaurus-like presentation used by some learner's dictionaries. Figure 7 shows the result of a thesaurus search in

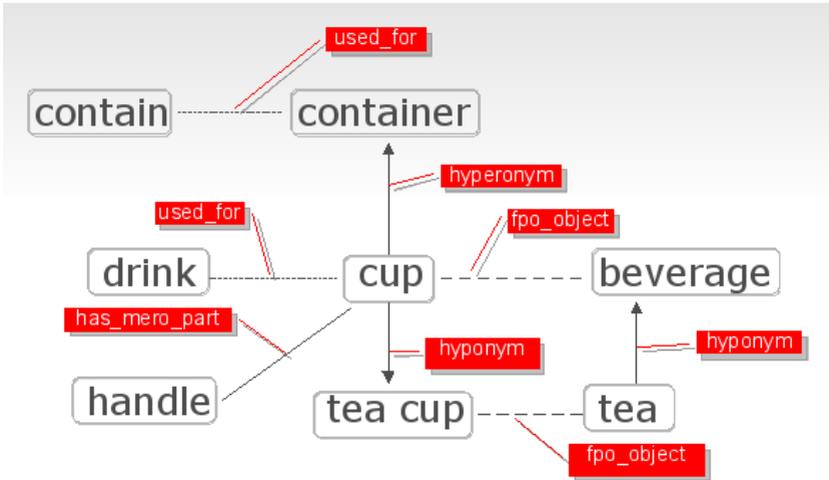


Abbildung 5: A visual presentation of the concept 'cup' showing DanNet relations

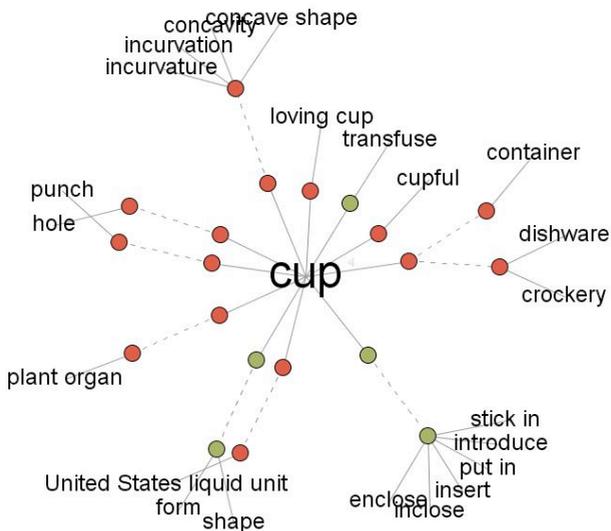


Abbildung 6: The result for 'cup' in The Visual Thesaurus

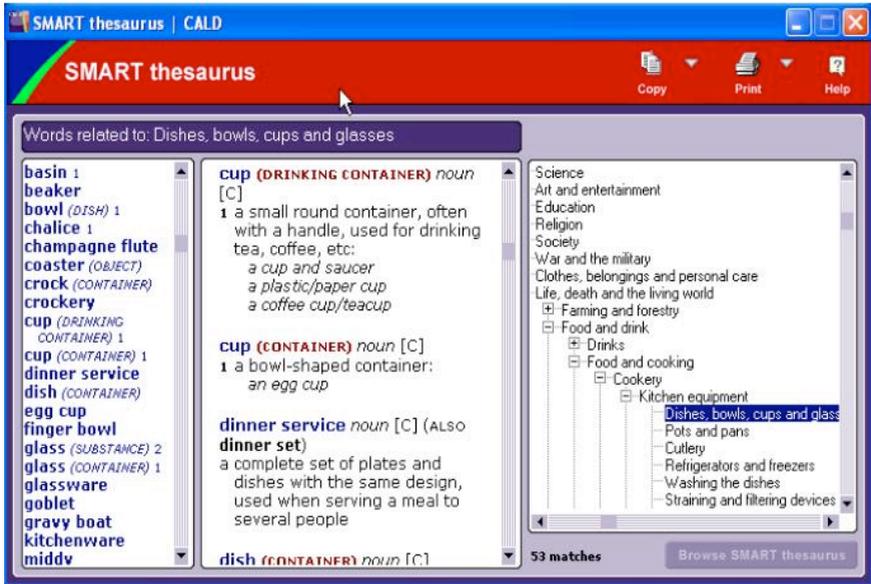


Abbildung 7: The result for 'cup' in the thesaurus of the Cambridge Advanced Learner's Dictionary (CALD)

the Cambridge Advanced Learner's Dictionary (available on CD-ROM or to online subscribers).

Like The Visual Thesaurus, the CALD thesaurus brings together dictionary articles and some kind of conceptual division of the vocabulary, the exact nature of which is not, however, entirely transparent to us. The thesaurus part is shown in the right column and although it might well have used wordnet data either as a point of departure or as a reference point, it seems to have been tailored to a structure that resembles that of Roget's original thesaurus rather than that of a wordnet. The user can navigate through the hierarchy and click to open relevant subcategories until the destination has been found. If the user clicks on a terminal group, the members of that category are displayed in the left column. A click on one of the members opens the corresponding dictionary entry in the central column. In contrast to The Visual Thesaurus, the CALD thesaurus only operates with a single hierarchical dimension based on hypernyms/hyponyms and co-hyponyms. The conceptual categories of the CALD thesaurus seem well suited for human users, in fact often more so than a wordnet hierarchy, in particular in the case where a wordnet category contains a large number of members. Here, the CALD thesaurus categories are more adequately sized and thus easier for human users to grasp. The DanNet hyponym hierarchy could be arranged in a similar way, but not without a substantial amount of manual effort. But as pointed out earlier, a viable semi-automatic

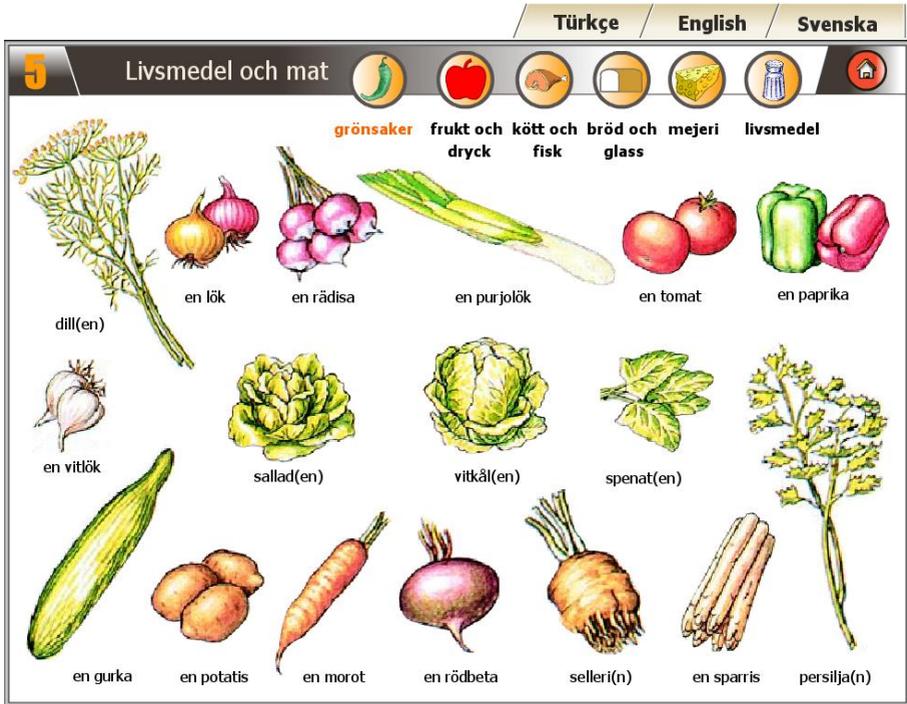


Abbildung 8: Pictures exemplifying the semantic field foodstuffs, taken from the Swedish immigrant dictionary LEXIN

procedure would be to divide large categories according to their ontological type or a relevant relation. Especially in a learning perspective, the presentation of related words belonging to the same domain is useful. For that reason, selected areas are often further developed in learner's dictionaries, and for example supplied with pictures that allow systematic training of vocabulary items (e.g. fruits and vegetables, motor vehicles, kitchen utensils). Figure 8 shows an example taken from the Swedish immigrant's dictionary Lexin.

One final perspective that we would like to mention falls within the area of lexicotainment. Although clearly not among the core functions of an online dictionary, one should not underestimate the role of gadgets and catchy features when it comes to attracting new users or catching the attention of the chance passer-by. Our suggestion is that wordnet data are well suited to improve crossword help in dictionaries. Many online dictionaries have this feature, but all existing helpers that we have come across suffer from the same shortcoming: they ask the user to specify the length of the word

The screenshot shows the 'CROSSWORD HELPER' interface on 'WORDPLAYS.COM'. On the left is a navigation menu with sections for 'Games' (Crossword Challenge, Boggler, Words in a Word, Word Morph) and 'Tool Box' (Anagrams, Crossword Helper, Pig Latin, Words in a Word, Jumble Words, Crypto Cracker, Word Morph, Scrabble Helper). Below that is a 'Dictionary' section with a 'Select Language' dropdown. The main area is titled 'CROSSWORD HELPER' and has 'ENGLISH' and 'Sign in' buttons. It contains instructions: 'Enter the letters that are known in the corresponding boxes. Select the length of the word. Click the Show Completions button. All words of the specified length with matching characters in the corresponding positions will be displayed. Empty fields will match all characters.' Below this is a 20-box grid with the 5th box containing 's'. Below the grid is an input for 'Enter the word length' set to '6' and a checkbox for 'Show Multiple Word Completions'. There are 'Show Completions' and 'Clear Entries' buttons. A message states 'Maximum Limit Reached - 4446 answers found for "-.....s ":'. A list of 25 results follows, each with a number and a word: 1 aaliis, 2 abacas, 3 abacus, 4 abakas, 5 abamps, 6 abases, 7 abates, 8 abatis, 9 abayas, 10 abbess, 11 abbeys, 12 abbots, 13 abeles, 14 abhors, 15 abides, 16 ablins, 17 abmhos, 18 abodes, 19 abohms, 20 abomas, 21 aborts, 22 aboves, 23 abuses, 24 abyms, 25 acarus.

Abbildung 9: A search for six-letter words ending in -s returns 4446 matches

and enter the letters that are known. Subsequently, the input is matched against all words in the database and the results returned. An example is shown in Figure 9 .

The problem is, as anyone who has ever done crosswords will know, that a query based on the string alone overlooks one important factor entirely: the clue. As a result, the string-based crossword helper overgenerates the number of matches and in particular so when the number of known letters are few. Figure 9 shows an example of poor help: the correct matches are lost in the crowd because of the vast number of results. Our suggestion is that dictionary and wordnet data would improve the crossword helper by filtering away undesired matches. This can be done by incorporating information about the clue in at least two respects: using the dictionary’s stock of inflectional information would allow the user to search for a particular inflectional form, e.g. a plural noun or a past tense verb, if the clue is in this form. And using wordnet data would allow the user to specify the semantic field within which the query should be performed. In the example above, if the clue was ‘sport’, a query within this domain in DanNet would quickly provide the answer: *tennis*, *isdans* (‘ice dance’) and *diskos* (‘discus’, regarded as a discipline in athletics) are the only three candidates that match.

Literatur

Asmussen, J., Pedersen, B. S., and Trap-Jensen, L. (2007). DanNet – From Dictionary to Wordnet. In Kunze, C., Lemnitzer, L., and Osswald, R., editors, *GLDV-2007 Workshop*

- on Lexical-Semantic and Ontological Resources*, number 336-3/2007 in Informatik-Berichte, pages 1-9, Hagen. FernUniversität Hagen, Fakultät für Mathematik und Informatik.
- DDO (2003-2005). *Den Danske Ordbog 1-6*. DSL & Gyldendal, København.
- Lenci, A., Bel, N., Busa, F., Calzolari, N., Gola, E., Monachini, M., Ogonowski, A., Peters, I., Peters, W., Ruimy, N., Villages, M., and Zampolli, A. (2000). SIMPLE – A General Framework for the Development of Multilingual Lexicons. *International Journal of Lexicography*, 13:249-263.
- ODS (1956). *Ordbog over det danske Sprog 1-28*. DSL & Gyldendal, København.
- Pedersen, B. S. and Nimb, S. (2008). Event Hierarchies in DanNet. In Tanács, A., Csendes, D., Vineze, V., Fellbaum, C., and Vossen, P., editors, *Proceedings of the Fourth Global WordNet Conference*, pages 339-348, Szeged. University of Szeged, Department of Informatics.
- Pedersen, B. S., Nimb, S., Asmussen, J., Sørensen, N. H., Trap-Jensen, L., and Lorentzen, H. (2009). DanNet – The challenge of compiling a WordNet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation*, to appear.
- Pedersen, B. S. and 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. and Sørensen, N. H. (2006). Towards Sounder Taxonomies in WordNets. In *Proceedings from the OntoLex Workshop in association with LREC 2006*, Genova.
- Pustejovsky, J. (1995). *The Generative Lexicon*. Bradford Book.
- Svensén, B. (1993). *Practical Lexicography: Principles and Methods of Dictionary-Making*. Oxford University Press, Oxford, Translated by John Sykes and Kerstin Schofield edition.
- Veale, T. and Hao, Y. (2008). Enriching WordNet with Folk Knowledge and Stereotypes. In Tanács, A., Csendes, D., Vineze, V., Fellbaum, C., and Vossen, P., editors, *Proceedings of the Fourth Global WordNet Conference*, pages 453-461, Szeged. University of Szeged, Department of Informatics.