The Semantic Relations of Artifacts in DanNet

Sanni Nimb
Det Danske Sprog- og Litteraturselskab (Society for Danish Language and Literature)
Christians Brygge 1, DK-1219, Copenhagen, Denmark
sn@dsl.dk

Abstract

This paper presents the newly released first version of the Danish wordnet, DanNet, focusing on the lexicon model and on the semantic description of artifacts. Apart from being the necessary resource for computational processing of Danish text material such as automatic indexing and information retrieval, the first version of DanNet also makes it possible to carry out linguistic investigations on parts of the Danish lexicon, due to the large number of well structured and consistent lexical data. One example is an investigation on the hyponymy relation at different levels of conceptual domains in Danish, showing a tendency of far more non-taxonomical sister concepts at the general language level than at the basic and specific levels of the language. Another example is an investigation of the distribution of the manually assigned relations in the synsets having an artifact sense in DanNet, showing that the telic relation 'used_for' describing the purpose of the artifact is by far the most frequently applied relation for this group of words. The paper also discusses the differences between the information found in dictionaries and the information to be included in a wordnet.

1 Introduction

The first version of the Danish wordnet, DanNet, was released in March 2009 as an open-source resource (see http://wordnet.dk). DanNet is the product of a joint project between two institutions, The University of Copenhagen, Center for Language Technology (CST), previously having compiled a pilot version of a computational semantic lexicon for Danish, SIMPLE-DK (Pedersen and Paggio, 2004), and the Society for Danish Language and Literature (DSL) that compiled the Danish dictionary which was used as the basis for the wordnet ((Den Danske Ordbog (henceforth DDO (DDO, 2003-2005)).

The 4 years (2005-2009), resulting in the first version, were funded by the Danish Research Council (3,000,000 DKK). In 2008 an additional 3-year funding of 1,000,000 DKK within the DK-Clarin project ensures that the wordnet will be extended by 25,000 synsets.

The first version of DanNet contains approx. 41,000 synsets (34,000 noun synsets, 6,000 verb synsets and 1,000 adjectival synsets). A synset is a set of synonymous lemmas referring to the same concept. e.g. {lys; stearinlys} (candle), {raritet; sjældenhed}(rarity); {humorist; humørbombe; humørspreder} (humorist) and {hoppe} (to jump). Often a synset contains just one single lemma. 26,458 noun lemmas, 3,094 verb lemmas and 809 adjective lemmas are described in the first version. Many of them are polysemous and we have focused on describing at least the main senses of the lemmas.

All synsets in the first version of DanNet are described with hyponymy relations as well as ontological type such as [Living+Object], [Artifact +Object+Part], [Human+Occupation], [Property] etc. 27,000 of the 41,000 synsets in the first version describe nouns having a concrete sense. Of these, approx 12,000 synsets, those referring to objects or human beings, are fully described with information on meronymy, near synonymy, connotation etc., in the case of humans the typical role of the person (e.g. humorist: entertain) and in the case of artifacts also information on origin (how it was made), purpose (what it is used for) as well as agents and instruments involved in the use of the artifact.

A small subset of the synsets in DanNet is linked to Princeton WordNet, and the aim is that 8,000 have been linked 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 i.e. Princeton WordNet to the language in question, in this case Danish. This method – the so-called merge approach – was chosen due to the fact that a corpus-based dictionary of Danish was completed in 2005 and accessible in a machine-readable version with hyperonymy information explicitly specified for each of the approx. 100,000 sense definitions. First of all, this made

it possible to build a Danish wordnet using semi-automatic methods, and we estimate that approx. 50% of the data in DanNet has been semi-automatically produced without further adding of data than what is found in DDO. But not less important, it guaranteed that the senses included in the wordnet were actually frequent in general language texts, as the aim of DanNet was to establish a linguistic resource for computational processing of Danish text material, for example automatic indexing, information retrieval, and automatic sense annotation.

Apart from offering linguistic data to developers within the language technology community, DanNet also makes it possible to carry out a wide range of lexical investigations on the Danish lexicon which have not been possible before, due to the systematic organization of the semantics we find in the definitions in DDO as well as completely new data on certain semantic relations not deducible from DDO.

2 The hyponymy hierarchy in DanNet

The wordnet was semi-automatically built by extracting all the senses in DDO having the same specified hypernym (genus proximum). The compiler of the wordnet then organized the proposed hyponymy hierarchy by either simply accepting the hypernym from DDO (which also involved a disambiguation in the many cases of polysemous genus expressions) or by manually selecting a new, and from a structural point of view more precise, hypernym, e.g. in the cases where the genus proximum in DDO was chosen arbitrarily among several synonymous possibilities, or in the cases where genus expressions referred to concepts on a higher level in the hierarchy than the nearest one from a structural point of view. One example of the latter case is 'budcykel' (carrier cycle used to bring out goods to customers) which has the genus proximum 'cykel' (bicycle) in DDO although the structurally seen nearest hypernym is 'ladcykel' (carrier cycle) – in DanNet it is therefore inserted as hyponym to 'ladcykel' instead (see Figure 1).

More challenging was the task of choosing between often more than one suitable hypernym. In some of these cases the synset has been linked to two hypernyms in DanNet: an offroader is both a kind of car and a kind of motorcycle, and a 'havestol' (outdoor chair) is both a chair and a piece of garden furniture. But in general only one hypernym was selected, i.e. the one with the highest number of relevant semantic relations to be inherited: 'slips' (a tie) is for this reason in DanNet described as a 'beklædningsgenstand' (a piece of garment) although defined as a piece of fabric in DDO.

In order to facilitate the practical use of the wordnet as a resource in formal ontologies, the so-called taxonomical hyponyms defined by the test: X is a kind of Y (Cruse, 2002) have been separated from the hyponyms for which the test does not hold (Pedersen and Sørensen, 2006, Pedersen et al., forthcoming). E.g. for the concept 'bicycle' the different kinds of bicycle (a mountain bike, a racer bike, a carrier cycle) are taxonomical in contrast to those hyponyms not being kinds of bicycles but instead describing a property transversely to the taxonomical group. Some examples are 'herrecykel' (gentleman's bicycle), and 'jernhest' (old bike). While members of the last group, which in DanNet are considered to be 'orthogonal' and assigned a special feature, are compatible with any hyponym of bicycle (a gentleman's bicycle as well as an old bicycle can at the same time be a racer bike or a mountain bike), members of the taxonomical group are only compatible with the members of the orthogonal group (a racer bike cannot be a mountain bike). In Figure 1, the orthogonal synset 'herrecykel' (gentleman's bicycle) is illustrated by a rhombus, in contrast to the taxonomical hyponyms 'ladcykel' (carrier cycle) and 'klubcykel' (standard bicycle).

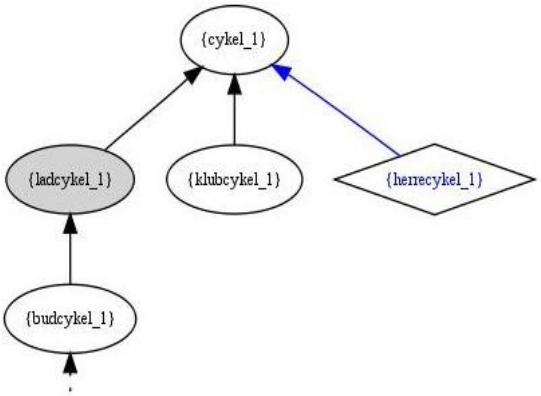

Figure 1. Some hyponyms of 'cykel' (bicycle). The rhombus figure indicates orthogonal hyponymy.

The encoded data on orthogonal versus taxonomical hyponymy in DanNet represents a new description of Danish concepts, the information on different categories of hyponyms not being deducible from the data in a traditional semasiological dictionary like DDO. See (Pedersen and Sørensen, 2006), (Pedersen and Nimb, 2008) and (Pedersen et al., forthcoming), for further discus-

sion of the hyponymy relation in DanNet, also in the case of verbs.

The orthogonal feature makes it possible to carry out linguistic investigations on the nature of the hyponymy relation between Danish words. Concepts can be classified as belonging to three different levels according to Dirven and Verspoor (1998, p. 38): the general level (plant, animal, garment), the basic level (tree, dog, trousers) and the specific level (oak, labrador, jeans). If we consider the hyponymy hierarchy for the approx. 6,800 concrete objects in DanNet, we find a very even distribution between the number of taxonomical and orthogonal co-hyponyms at the general language level. In other words, the direct hyponyms of 'genstand' (object), that is concepts like garment, toy, tool, and vehicle, have many orthogonal sister concepts which, in principle, are compatible with any taxonomical hyponym of 'genstand', such as 'ejendom' (property), 'blikfang' (eye catcher), 'eksemplar' (specimen), 'helligdom' (shrine), 'kopi' (copy), 'nyhed' (novelty), 'opfindelse' (invention), 'original' (original) and 'værdigenstand' (article of value). In Danish we have many words denoting any kind of object which is owned, copied, invented, new, valuable etc. We find a much smaller percentage of orthogonal hyponyms the further down we move in the DanNet hyponymy hierarchy, also when it comes to the generally quite large sets of hyponyms of the basic level concepts (e.g. book: 28 taxonomical and 14 orthogonal hyponyms; shoe: 28 taxonomical and 5 orthogonal hyponyms, trousers: 16 taxonomical and 0 orthogonal hyponyms). The concepts at the specific language level seem to have very few orthogonal sister concepts.

3 The set of semantic relations in DanNet

The set of semantic relations in DanNet is based on the wordnet relations from EuroWordNet (Vossen, 1998), extended by three relations from the SIMPLE lexicon. In the SIMPLE model (Lenci et al., 2000), semantic relations are organized according to the four qualia roles (Pustejovsky, 1995), relating to inheritance structure, origin, composition and purpose. None of the EuroWordNet relations cover the origin dimension and the purpose dimension of a concept. During the compiling of the Danish SIMPLE lexicon (Pedersen and Paggio, 2004), it turned out that the four-dimensional qualia structure in general ensured most semantic aspects of a word sense to be described in the lexicon. Therefore, the two SIMPLE relations 'made_by' and 'used_for' were included in DanNet. Also the relation 'concerns' from the SIMPLE model was added. Furthermore some relations on synonymy are part of the wordnet set of relations. See Table 1.

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

Table 1 Semantic relations in DanNet

4 A concept and its relations in DanNet

The relations assigned to a concept, e.g. the basic-level concept 'bog' (book), see Table 2, is in DanNet mainly based on the DDO sense definitions. In addition to this, an examination of the hyponyms of the concept also proved necessary as the set of hyponyms often reveals a number of central semantic aspects of the hypernym in question which are not mentioned in the DDO definition. Consider for example the many hyponyms of 'bog' (book) which describe the topic of the book thus making it clear that the topic is in fact a central semantic aspect of a book, even though this is not mentioned in the definition of 'bog' itself in DDO. We find 'fuglebog' (bird book, concerns: bird), 'kogebog' (cookery book: concerns: cooking), 'kriminalroman' (detective novel, crime novel: concerns: crime). The semantic relation 'concerns: topic' has therefore

been assigned at the top level of the 'bog'-hierarchy in DanNet and is subsequently restricted to a more precise synset for those hyponyms having a specific topic sense.

Ontological type	[LanguageRepresentation+ Artifact+Object]
Formal role/ INHERITANCE	has_hyperonym: 'genstand' (object)
Agentive role/ ORIGIN	made_by: skrive (write); trykke (print)
Constitutive role/ COMPOSITION	has_mero_made_of: papir (paper) has_mero_part: tekst (text), side (page), ryg (back), titel (title) concerns: emne (topic) involved_agent: forfatter (writer) involved_agent: læser (reader)
Telic role/ PURPOSE	used_for: læse (to read)
Synonymy	near_synonym: hæfte (booklet; pamphlet)

Table 2. The semantic relations of 'bog' (book) in DanNet

In DanNet, the aim has been to describe explicitly as much semantics as possible by giving precise relations to other concepts in order to compensate the likely deficit of world knowledge in NLP software using lexical data like DanNet. Veale and Hao (2008) claim that even the kind of knowledge we normally find in dictionaries does not cover what it takes to make a computer understand everyday language, and that wordnets should be enriched with information on stereotypes and culturally-inherited associations. This is outside the scope of DanNet at its current stage the aim being instead to define the native speaker's lexical knowledge about a concept and focus on the prototypical semantic aspects. From an ideal point of view, DDO would contain exactly this level of information so that the information found here just needed to be translated into semantic relations in DanNet, but due to the fact that dictionary definitions lean on the language-user's ability of making assumptions (Svensén, 1993) this is often far from beeing the case. Also for syntactic reasons DDO does not always bring all the information needed in DanNet. The definition in DDO had to be a well-formed, not too complicated or long phrase, and this is probably the reason why nothing is said about the topic in the case of books, nor about books typically having a title, being written by an author, read by a reader etc. Furthermore, the entries in DDO are meant to be read as a whole, implying that some semantic aspects might emerge from the examples, the list of connotations etc. Finally and maybe most importantly, the DDO definitions were created in a bottom-up way, without schematic specifications for a given group of words in order to ensure all relevant semantic aspects to be covered systematically. Therefore it is not surprising that we often find a discrepancy between the sometimes quite large number of relations which from a systematic point of view should be described for a given sense in order to reflect the native speaker's lexical knowledge, and the ones which are explicitly described in the definition of the word in DDO.

Comparing DDO and DanNet, we can conclude that in the case of artifacts DanNet in general contains more information on the meronymy relations than DDO does, especially in the cases of the basic-level concepts. In DanNet we find information on books having pages, a back and a title, and on shops having display windows, information not found in DDO. DanNet also contains far more information than DDO does on the typical user of an artifact (e.g. easy reader / pupil, hymn book / church goer). In Table 3 we present a range of examples of cases where information on the typical user has been added in DanNet, compared to what is mentioned in DDO. What is interesting in these cases is that the artifact lemma is often morphologically closely related to the user or vice versa, as in the examples of shop, shopkeeper, shopper; pharmacy, pharmacist; bakery, baker; pilot licence, pilot.

Synset	**Added information in DanNet compared to DDO**
flyvecertificat (pilot licence):	involved_agent: pilot (pilot)
briller (glasses)	involved_agent: person (person)
forskningsbibliotek (research library)	involved_agent: forsker (researcher)
læbestift (lipstick)	involved_agent: kvinde (woman)
barberkost (shaving brush)	involved_agent: mand (man)

Synset	Added information in DanNet compared to DDO
ægteskab (marriage)	involved_agent: ægtepar (married couple)
apotek (pharmacy)	involved_agent: apoteker (pharmacist)
bageri (bakery):	involved_agent: bager (baker)
registreringsattest (vehicle registration certificate)	involved_agent 'motorkontor' (motoring office).

Table 3. Examples of added information in DanNet compared to what is described in DDO

A statistical investigation of the manually added relations (i.e., those not automatically inherited from the hypernym of the synset) in the synsets of 6,800 object artifacts gives an idea of the most important relations when describing an artifact by semantic relations. See Table 4.

Semantic relation	Percentage of 6,800 artifact objects described with the relation
used_for	28% (book/to read)
has_mero_part	14% (book/page)
concerns	9% (christmas decorations/christmas)
made_by	6% (clothes/to sew)
involved_agent	6% (guitar/guitarist)
has_holo_part	5% (page/book)
has_mero_madeof	5% (clothes/fabric)
has_holo_location	3% (carpet/floor)
near_synonym	3% (book/pamphlet)
Others relations	1% or less

Table 4. The distribution of the percentage of manually assigned relation types in 6,800 synsets with an artifact object sense (inherited relations not included).

The frequent use of the DanNet relations taken from the SIMPLE lexicon model (used_for, made_by, and concerns) supports the decision of extending the set of standard WordNet relations. It should be remarked that this type of information is often deducible from the DDO definition, in opposition to the information in DanNet on the involved user.

The number of manual assignments of relations also indicates how often we find lexical restrictions between the relations in artifact synsets. In DanNet a general 'used_for' relation is always assigned at the top hypernym of a certain group of artifacts (e.g. tool: used_for: to use; garment: used_for: to dress). Also the involved_agent relation is assigned here with the value 'person' (person), e.g. tool: involved_agent: person. Whenever the inherited relation value is too imprecise and a manual assignment of the two relations is applied for a hyponym, it reflects a lexical relation between the artifact synset itself, the synset describing the kind of use, and the synset describing the kind of user. We find these cases relatively often, since one out of four cases of a manual assignment of the used_for relation, e.g. for shaving brush to shave, and for pilot licence to fly, also has resulted in a restriction on the type of user, e.g. shaving brush: man; and pilot licence: pilot.

5 Conclusion

DanNet contains a high number of well-structured and consistent semantic data on the Danish word senses, and in several cases also more information than what can be found in the definitions in the dictionary on which the wordnet is based, e.g. on different groups of hyponyms and on the involved user of artifacts. The investigations in the DanNet data of 1) the distribution of taxonomical and orthogonal hyponyms at different conceptual levels and 2) the distribution of the different relations used to describe artifact synsets, which have been presented here, shed new light on the semantic relations between a group of concepts in the Danish lexicon and is just a minor example of the types of lexical-semantic studies that can be carried out on a wordnet like DanNet.

References

D.A. Cruse. 2002. Hyponymy and Its Varieties. Green, R., Bean, C.A., Myaeng, S.H. (eds.) *The Semantics of Relationships: An Interdisciplinary Perspective, Information Science and Knowledge Management*. Springer Verlag.

DDO = Hjorth, E., Kristensen, K. et al. (eds.). 2003-2005. *Den Danske Ordbog 1-6* ('The Danish Dictionary 1-6'). Gyldendal and Society for Danish Language and Literature, Denmark.

R. Dirven and M. Verspoor (eds.). 1998. *Cognititive Exploration of Language and Linguistics*, John Benjamin Publishing Company Amsterdam/ Philadelphia.

A. Lenci., N. Bel, F. Busa, N. Calzolari, E. Gola, M. Monachini, A. Ogonowski,, I. Peters, W. Peters, N. Ruimy, M. Villegas. and A. Zampolli. 2000. SIMPLE – A General Framework for the Development of Multilingual Lexicons. T. Fontenelle (ed.) *International Journal of Lexicography* Vol 13, pp. 249-263. Oxford University Press.

B.S. Pedersen and P. Paggio. 2004. The Danish SIMPLE Lexicon and its Application in Content-based Querying. *Nordic Journal of Linguistics* Vol 27(1), pp. 97-127. Cambridge University Press.

B.S. Pedersen and N. Sørensen. 2006. Towards Sounder Taxonomies. A. Oltramari, Chu-Ren Huang, A. Lenci, P. Buuitelaar, C. Fellbaum (eds) *Wordnets*. Ontolex 2006 at 5th International Conference on Language Resources and Evaluation, pp. 9-16. Genova, Italy.

B.S. Pedersen and S. Nimb. 2008. Event Hierarchies in DanNet. A. Tanács, D. Csendes, V. Vincze, C. Fellbaum, and P. Vossen (eds) *Proceedings of Global WordNet Conference,* University of Szeged, Hungary, pp. 339-349. University of Szeged, Juhász Press Ltd., Hungary.

B.S. Pedersen, S. Nimb, J. Asmussen, N. H. Sørensen, L. Trap-Jensen and H. Lorentzen (forthcoming). *DanNet - the challenge of compiling a WordNet for Danish by reusing a monolingual dictionary.* Language Resources and Evaluation. Springer Netherlands.

J. Pustejovsky. 1995. *The Generative Lexicon*. The MIT Press, Cambridge, Massachusetts.

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

T. Veale and Y. Hao. 2008. Enriching WordNet with Folk Knowledge and Stereotypes. *Proceedings of the Fourth Global Wordnet Conference*, University of Szeged, Hungary, pp. 453-461. University of Szeged, Juhász Press Ltd., Hungary.

P. Vossen (ed.). 1999. *EuroWordNet, A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, The Netherlands.