Towards a richer wordnet representation of properties

Sanni Nimb, Bolette Sandford Pedersen

Society for Danish Language and Literature & University of Copenhagen, Denmark

E-mail: sn@dsl.dk, bspedersen@hum.ku.dk

Abstract

This paper discusses how information on properties in a currently developed Danish thesaurus can be transferred to the Danish wordnet, DanNet, and in this way enrich the wordnet with the highly relevant links between properties and their external arguments (i.e. *tasty – food*). In spite of the fact that the thesaurus is still under development (two thirds still to be compiled) we perform an automatic transfer of relations from the thesaurus to the wordnet which shows promising results. In all, 2,362 property relations are automatically transferred to DanNet and 2% of the transferred material is manually validated. The pilot validation indicates that approx. 90 % of the transferred relations are correctly assigned whereas around 10% are either erroneous or just not very informative, a fact which, however, can partly be explained by the incompleteness of the material at its current stage. As a further consequence, the experiment has led to a richer specification of the editor guidelines to be used in the last compilation phase of the thesaurus.

Keywords: wordnet, properties, thesaurus

1. Introduction

Where wordnets most often represent properties along the lines of hyponymy and antonymy, thesauri have the tradition of providing a thematic related representation of properties with implicit reference to the arguments that they modify. To exemplify, most thesauri would group taste denoting properties such as *hot*, *tasty* and *insipid* under the food chapter and thereby provide not only a unique reference to the food concepts that they are properties of, but also a thematic grouping of the properties themselves. In this paper we argue that information on arguments and the thematic information covered so richly in thesauri is highly relevant also for lexical semantic networks meant for natural language processing, such as wordnets. To this end, we describe how we are automatically enriching properties in the Danish wordnet, DanNet, by extracting explicit data on semantic relations as well as information on thematic relatedness from a new Danish thesaurus.

The presented research is joint work between the Society for Danish Language and Literature and the University of Copenhagen; the team who also built the original version of DanNet (cf. Pedersen et al., 2009 as well as wordnet.dk where the wordnet can be downloaded as open source) on the basis of a monolingual dictionary of Danish, The Danish Dictionary (DDO).

The transfer of data is performed in a triangle-like fashion, illustrated by Figure 1, where information is recompiled from DDO to DanNet, then from DanNet to The Danish Thesaurus (DT) where it is supplied with the approx. 50,000 DDO senses not yet represented in DanNet, and finally – as described in this paper – from DT to DanNet. At each level, formalized linguistic information on semantic relations, domain etc. is added to the original DDO senses.

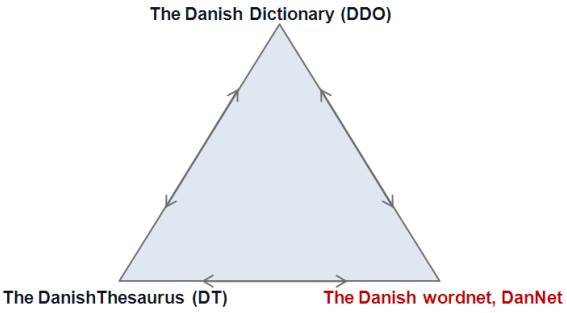

Figure 1: The re-compiling triangle DDO-DanNet-DT

In the following we will go through related work (Section 2), how properties where initially encoded in DanNet (section 3), how they are represented in DT (section 4), and finally how DT information on properties is transferred into DanNet (Section 5).

2. Related work

Properties are most typically expressed as adjectives, a word class which has been treated rather heterogeneously in wordnets over the years. Princeton WordNet originally operates with a number of adjective classes and focuses on clusters of synonymous adjectives mutually related by antonymy relations (Miller, 1998). EuroWordNet (Vossen, 1998 and 1999) allows for adjectives to be further described by relations across word classes, such as cause relations (*dead* is_caused_by *to kill*) and synonymy relations (*poor* xpos_near_synonym *poverty*). GermaNet (Hamp & Feldweg, 1997) groups adjectives into semantic classes. They 'avoid the rather fuzzy concept of indirect

antonyms introduced by WordNet' by considering antonymy a lexical relation. Instead they model adjective synsets following the same taxonomic approach as with nouns and verb. In order to avoid the otherwise very flat taxonomies, they introduce artificial concepts (Kunze, 2000). Adjectives are linked to their nominal base by the relation 'pertains_to' (i.e. *financial*/*finances*).

Also the Portuguese wordnet (Mendes, 2006) introduces another treatment of adjectives than Princeton WordNet. Adjectives are not linked to one another by antonymy relations but instead linked to a noun denoting a related property by the relation 'characterises with regard to'. I.e. adjectives like *short*, *tall* and *high* are linked to the synset {*height*}. As in GermaNet, antonymy is considered to hold between members of synsets, and not between the synsets themselves. Since many nouns are described by inheritable adjectival synsets (for example *shark* has_as_a_characteristic *carnivorous*), information of the adjective synsets is furthermore obtained via the reverse relation (*carnivorous* is_a_characteristic_of *shark*).

3. Properties in DanNet

DanNet applies the EuroWordNet Top Ontology and thereby the ontological type PROPERTY to characterize properties. 12 different subtypes of PROPERTY have been established by combining PROPERTY with the EuroWordNet's so-called situation components (TIME, MENTAL etc.) as seen in Figure 2.

Property
Property+Existence
Property+LanguageRepresentation
Property+Location
Property+Mental
Property+Physical
Property+Physical+Colour
Property+Physical+Condition
Property+Physical+Form
Property+Social
Property+Stimulating+Physical
Property+Time

Figure 2**:** Subtypes of PROPERTY in DanNet

The most frequently used ontological types are PROPERTY+PHYSICAL, PROPERTY+MENTAL as well as the lesser specified PROPERTY. The most specific ones are the subtypes of PROPERTY+PHYSICAL, including CONDITION (typically nouns specifying illnesses), FORM and COLOUR. Positive and negative connotation has been ascribed to a subset of adjectives, but apart from this distinguishing feature, the taxonomical structure is predominantly flat (as we saw in GermaNet), most properties simply being hyponyms of the synset {*egenskab, beskaffenhed, side*} ('property', 'feature', 'characteristics') since DanNet does not allow artificial concepts. This leaves us with a rather underspecified resource where the following semantically very diffent synsets are all co-hyponyms of the top property synset : *arbejdsløs* ('unemployed'), *asbestfri* ('asbestos-free'), *automatisk* ('automatic'), *begivenhedsrig* ('eventful'), *behovsorienteret* ('needs-oriented'), and *bladløs* ('leafless'). In some cases, small taxonomies and a more precise ontological type assures a higher level of information, but still quite different adjectives like *filminteresseret* ('interested in movies'), *fodbold-interesseret* ('interested in football'), *kompromissøgende* (seeking a compromise), *kærlighedssøgende* ('love seeking') and *centrumsøgende* ('centre-orientated' - about politicians seeking a middle way) are not formally separable since they are all hyponyms of the synset {*interesseret*} ('interested') with the ontological type PROPERTY+MENTAL. In this case the ontological type gives us the information that the synsets have to do with a person's mental state, i.e. either a way of thinking or feeling. But no features distinguish the co-hyponyms from one another and links between characterizing adjectives and their corresponding nouns have not been established in previous versions of DanNet. Furthermore adjectival synsets have only been used in very few cases to characterize other synsets in the wordnet leaving us with almost no reversed information.

4. Properties in DT

The Danish Thesaurus (DT) is divided into 22 thematic main chapters (with titles like 'Food and drinks', 'Feellings', 'Sports' etc.), which again are divided into 970 numbered thematic sections, i.e. 'Drinking alcohol'), 'Anger' and 'Football'. The chapter and section structure is translated from a German thesaurus (Dornseiff, 2004). In each section, senses associated with the theme are combined in clusters of near synonyms (unless synonymy or antonymy is explicit marked), which are again combined in different types of larger semantic groups established according to a set of principles based on the type of semantic relation that holds between the members (co-hyponyms, co-meronyms, properties of the same type of argument etc.). For each sense, a shared ID number guarantees the exchange of data between DT, DanNet and DDO (see Figure 3).

DT	chapter 16 ('Food and drinks') section 16.09 ('Drinking alcohol') group 16.09.01 cluster 16.09.01.01 *fuld* **ID 21025001**
DanNet	Synset 1405 {*fuld* **ID 21025001**, word 2..}
DDO	*fuld* sense 4.1 **ID 21025001**

Figure 3: Structure of DT with shared ID-numbers in DT, DanNet and DDO, example *fuld* ('drunk')

Each of these groups contains formalized data on the relations that hold between the members or between the

members and a specific wordsense (holonym, involved agent, involved patient etc.), corresponding to some of the already existing relation types in DanNet. In the case of the external argument of properties though, a relation is introduced named 'property_of' (*fuld* ('drunk') - *fuldskab* ('drunkenness') - *have en lille fjer på* ('have had a drop too many') property_of *person*). The relation is used to express information which is in fact also described in traditional (monolingual as well as bilingual) dictionaries, i.e. when adjectives (but also other word classes) expose a very specific lexical choice. For illustration, Oxford English Dictionary says about the adjective *mature* that it is about fruits (sense 1a: "Of fruit, etc.: ripe"). In DT the information of the type of external argument is also given in the many cases where it is implicit in dictionaries, since it is often the unifying feature of a group of words within a section. Also other thesauri tend to group adjectives according to the type of external argument but without informing on its type in the text (Roget 2002, Dornseiff 2004, Andersen 1945). In opposition DT formalizes this information explicitly with the purpose of being able to transfer it to DanNet.

This has as consequence that properties within the same domain, but having different external arguments are kept apart. For illustration, words characterizing alcoholic drinks (*alkoholisk*, *stærk*, *svag* ('alcoholic', 'strong', 'weak')) are found in the same section as *fuld* ('drunk') but in a different group. In those cases where the properties also can be assigned a common hypernym (including hyponymy across word classes) the groups also contain formalized information on this (i.e. in the case above where *fuldskab* ('drunkenness') is the hypernym of the nouns and the xpos-hyperonym of the verbal phrase *have en lille fjer på* ('have had a drop too many') and the adjective *fuld* ('drunk'). We exemplify this by presenting data from a semantic group in section 2.24 'Appearance' covering the concept of "short height", cf. Figure 4. In Figure 5 the resulting data of the formalized information in Figure 4 is presented.

Figure 4: Section in DT describing the concept of short height (of person: 'short' (5 synonyms), 'dwarfishness', 'be a head shorter than..')

Groups describing properties are very frequent in DT. In the completed third of the dictionary, 60 % of the sections contain at least one property, and 16 % of the words or expressions (more than 8,000) belong to a group of this type.

header information	type of group ='properties' (*04_Egenskaber*) number of group: *2.24.1* has_hyperonym *højde* ('height') property_of *person* ('person')
clustered members	number of cluster: *2.24.1.05* *lav · lille · lille af vækst · lille af højde · af lav højde* (expressions for 'short')
not clustered members	*dværgvækst* ('dwarfishness') *være et hoved mindre end ..* ('to be a head shorter than..')

Figure 5: The formalized information for property words of persons associated with the concept 'short height'.

This indicates that we will end up with a huge amount of new information on these types of senses compared to what we find in the current version of DanNet when the DT is fully completed in 2013.

5. Information on properties transferred from DT to DanNet

An experiment with the transfer of data from DT to DanNet was carried out at an early stage in order to test whether the formalized descriptions function as planned and whether the transfer can be carried out automatically or should be semi-automatically inserted.

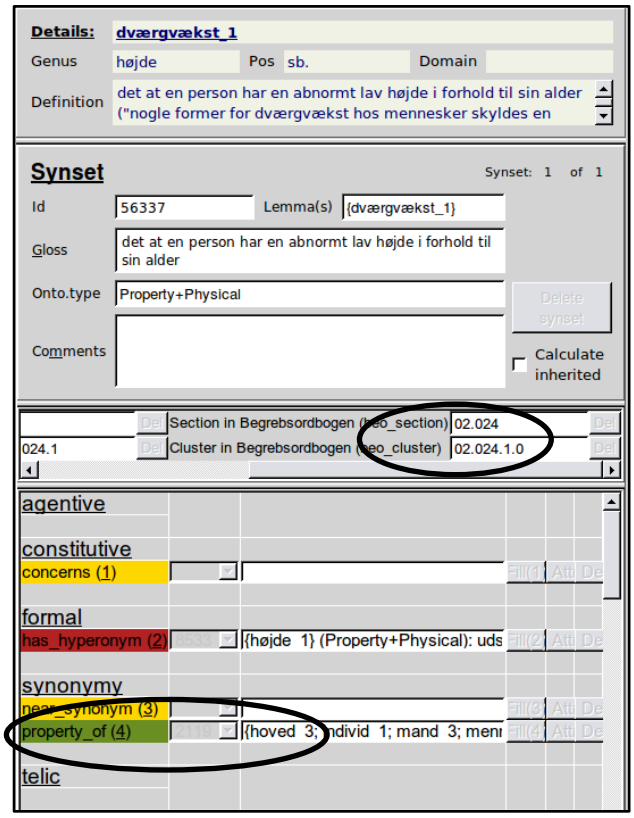

Figure 6: The DanNet interface where {*dværgvækst*} ('dwarfishness') is assigned property_of {*individ, mand, menneske*} ('person') as well as section numbers and cluster numbers from DT.

Initially, the header information on the property_of relation was assigned to each sense represented in the property group. Collocational expressions in DT were discarded from the transfer set (25 %). The transfer of the remaining data was easily carried out via the shared ID numbers of word senses in DanNet and DT. Each sense should be member of a synset in DanNet as well as should the sense constituing the value of the external argument in the property_of relation (i.e. *person* in the case of *fuld* (‘drunk’)). In the case of a clash of data, i.e. where two senses from the same synset were assigned two different values, no value was inserted. The transfer resulted in 2,362 inserted property_of relations in DanNet. For instance the noun synset {*dværgvækst*} ('dwarfishness') which is already represented in DanNet having the hypernym {*højde*} ('height') was extended with the relation property_of *person* (‘person’) as seen in Figure 6.

To form a pilot testing, 2 % of the extended synsets were validated, leading to the conclusion that approx. 10 % of the assigned property_of values were either wrong or not very informative. The wrong cases were studied and we concluded that in these cases the relation had been wrongly assigned in DT, which leads to an adjustment of the editor guidelines. In cases where the assigned relation was not considered to be very informative, it was due to the fact that the external argument of the property word seen in isolation was not restricted to only the one type which was transferred from the DT data, but should be supplemented with more information. Figure 7 presents more examples of the obtained results.

property_of relations transferred from DT to DanNet		
{*særegen*} ‘peculiar’	property_of	{*genstand,ting*} (‘object’,’thing’)
{*forskrækkelig*} (‘awful’, ‘dreadful’)	property_of	{*genstand,ting*} (‘object’,’thing’)
{*komisk*} (‘comical’, ‘comic’)	property_of	{*situation*} (‘situation’)
{*bedrøvelig,sørgelig, trist*}(‘sad’)	property_of	{*situation*} (‘situation’)
{*komplementær*} (‘complementary’)	property_of	{*farve*}(‘colour’)
{*stærk*} ‘intense’	property_of	{*følelse* }(‘feeling’)
{*smartness*} (‘smartness’)	property_of	{*person*}(‘person’)
{*tilbøjelighed*} (‘proclivity’,‘propensity’)	property_of	{*person*}(‘person’)
{*berøringsangst*} (‘reluctance’)	property_of	{*person*}(‘person’)
{*ustyrlig*}(‘unruly’)	property_of	{*person*}(‘person’)
{*smage igennem*} (‘have a dominant taste’)	property_of	{*mad*} (‘food’)
{*tung*} (‘heavy’, ‘stodgy’)	property_of	{*mad*} (‘food’)
{*groftskåren*}(‘coarse-cut’)	property_of	{*mad*} (‘food’)

Figure 7: Examples of property_of relations obtained in DanNet via transfer of information from DT

The ideal scenario when DT is completed is that all properties will be described with information on all their prototypical types of external arguments. However, we estimate that it will be necessary to validate the data transfer, i.e. perform it in a semi-automatic way in order to supplement the information with manually inserted property_of relations when needed.

In future work, we plan to experiment with the formalized information on hypernyms in DT in order to create new synsets in DanNet since many DT senses are not yet included in the wordnet. Also other relations can be deduced from the data, i.e. xpos_has_hyperonym relations (covering the relation between a synset and a hypernym from a different word class). An example is given in Figure 4 and 5, where all the synsets based on adjectives, adverbials and verbs can be assigned the xpos_hyperonym value *højde* ('height'). We also plan to link the synsets with the relation near_synonym when POS values are identic, otherwise via the relation xpos_near_synonym. If we compare to the treatment of adjectives in the portuguese wordnet (Mendes, 2006), their relation ‘characterises with regard to’ (*short* ‘characterises with regard to’ *height*) corresponds to an xpos_hyperonymy relation in our model. Their deduced relation 'is_a_characteristic_of' (*carnivorous* is_a_ characteristic_of *shark*), is described in more general terms in our model via the transfer of the property_of relation from DT. The external argument is in DT often the most general word having this characteristic (for *short* property_of *person*, for *carnivourous* it would be *carnivore*).

The domain structure itself in DT is also relevant in DanNet, creating links between members of the same cluster, the same group, the same section and the same chapter. I.e all words in Figure 4 are closely related to the group of words denoting ‘short persons’ in the same section, resulting in information like ‘short’ related to ‘pygmy’; ‘dwarfishness’ related to ‘pygmy’, ‘short’ related to ‘manikin’, etc. In the cases of senses sharing the same hypernym in DanNet, this relatedness can be used as subdividing information. For instance, if we return to the words mentioned in Section 3, which are co-hyponyms in DanNet with no distinctive semantic features, *centrumsøgende* ('centre-orientated') belongs to one section while *kærlighedssøgende* ('love seeking') belongs to another. In Figure 6, the added domain information for section and clusters is shown in the synset {*dværgvækst*} ('dwarfishness').

6. Conclusion

As we have shown, the formalized description of the external argument of properties in a Danish thesaurus, combined with the re-compiling triangle it establishes with DDO and DanNet, allows us to transfer a new semantic relation to DanNet with an estimated correctness of around 90 %. However, we plan to perform the final

transfer semi-automatically in order to guarantee the insertion of information on the whole set of prototypical arguments for property synsets in the wordnet. Furthermore, the detailed coverage of the Danish vocabulary in DT, where a high number of DDO senses are divided into more than 900 domains and described with different types of formalized information, allows us to continue with more experiments on data transfer from the thesaurus to the wordnet, i.e. regarding the extension of the number of synsets as well as relations and the use of domain information as a subdiving factor.

7. Acknowledgements

Nicolai H. Sørensen and Thomas Troelsgård, Society for Danish language and Literature, carried out the transfer of data from DT to DanNet. The Carlsberg Foundation is funding the DT project from 2010-2013.

8. References

DDO: Hjorth, Ebba & Kjeld Kristensen (eds.) (2005). *Den Danske Ordbog*. Copenhagen: Gyldendal & Det Danske Sprog- og Litteraturselskab. Online version 2012: http://ordnet.dk/ddo.

Dornseiff, Franz (2004). *Der deutsche Wortschatz nach Sachgruppen*, 8. Auflage. Berlin/New York: Walter de Gruyter.

Hamp, Birgit and Helmut Feldweg (1997). GermaNet - a lexical-semantic Net for German. In P. Vossen et al. (eds.). *Proceedings of the ACL/EACL-97 workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, Madrid, pp. 9-15.

Kunze, Claudia (2000). Extension and use of GermaNet, a lexical-semantic database. In *Proceedings of LREC 2000*, Athens, Greece.

Mendes, Sara (2006) Adjectives in WordNet.PT. In Soika, P., K.-S. Choi, C. Fellbaum & P. Vossen (eds.) *Proceedings of The Third Global WordNet Association Conference*, Jeju Island, Korea, pp. 225-230.

Miller, Katherine J. (1998). Modifiers in WordNet. In Fellbaum, Christiane (ed.): *WordNet: an electronic lexical database*, Cambridge MA: The MIT Press pp. 47-68.

Oxford English Dictionary (2012), online version oed.com, Oxford University Press.

Pedersen, Bolette .S, Sanni Nimb, Jørg Asmussen, Nicolai Sørensen, Lars Trap-Jensen & Henrik Lorentzen (2009). DanNet: The challenge of compiling a WordNet for Danish by reusing a monolingual dictionary. In *Language Resources and Evaluation, Computational Linguistics Series 43 (3),* 269-299, doi:10.1007/s10579-009- 9092-1.

Vossen, Piek (ed.) (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers.

Vossen, P. (ed.) (1999). *EuroWordNet General Document*. http://vossen.info/docs/2002/EWNGeneral.pdf (accessed 12.March 2012).