DanNet2: Extending the coverage of adjectives in DanNet based on thesaurus data

Sanni Nimb
Society for Danish Language and Literature
sn@dsl.dk

Bolette S. Pedersen
University of Copenhagen, CST
bspedersen@hum.ku.dk

Sussi Olsen
University of Copenhagen CST
saolsen@hum.ku.dk

Abstract

The paper describes work in progress in the DanNet2 project financed by the Carlsberg Foundation. The project aim is to extend the original Danish wordnet, DanNet, in several ways. Main focus is on extension of the coverage and description of the adjectives, a part of speech that was rather sparsely described in the original wordnet. We describe the methodology and initial work of semi-automatically transferring adjectives from the Danish Thesaurus to the wordnet with the aim of easily enlarging the coverage from 3,000 to approx. 13,000 adjectival synsets. Transfer is performed by manually encoding all missing adjectival subsection headwords from the thesaurus and thereafter employing a semi-automatic procedure where adjectives from the same subsection are transferred to the wordnet as either 1) near synonyms to the section's headword, 2) hyponyms to the section's headword, or 3) as members of the same synset as the headword. We also discuss how to deal with the problem of multiple representations of the same sense in the thesaurus, and present other types of information from the thesaurus that we plan to integrate, such as thematic and sentiment information.

1. Introduction to the project

In this paper, we provide a project description of the recently initiated 'DanNet2' project financed by the Carlsberg Foundation. The project runs from 2019-2022 and aims at investigating to which degree a recently compiled Danish Thesaurus, DDB (Nimb et al. 2014a; Nimb et al. 2014b) can be used to facilitate the extension of the lexical coverage of the Danish wordnet DanNet (cf. Pedersen 2009). Where the first version of DanNet was semi-automatically compiled on the basis of the isolated information on genus proximum in the manuscript of The Danish Dictionary (Hjorth & Kristensen 2003-2005, henceforth DDO) and covers 50% of its senses, we now want to exploit that 90% of the DDO vocabulary is thematically and semantically grouped in a newly compiled thesaurus. The three lexical resources share id numbers at sense level, making it possible to develop methods where data from one are transferred to the other. This was already exploited in the compilation of many thesaurus sections. DanNet data constituted for example the basis of the sections on diseases, garment and furniture based on information on ontological type in the wordnet. Like the current version of DanNet, also the extended version compiled in the DanNet2 project will be open source and downloadable via CLARIN-DK and github.

The main focus in DanNet2 is on the upgrade of the coverage and description of adjectives in DanNet, which in the thesaurus are richly represented, but rather sparsely described in the wordnet with a quite limited coverage of approx 3,000 adjective synsets. Our goal is by the end of the project to reach a more or less complete coverage of approx. 13,000 adjectival synsets.

We start out in Section 2 with related work on the treatment of adjectives in wordnets and similar resources, and move on to the way they are currently described in DanNet.

In Section 3 we describe how adjectives in the thesaurus are presented along topical chapters, sections and subsections, a structure that we want to use as source for the semi-automatic extension of adjectives in DanNet.

Section 4 presents the semi-automatic transfer method, where we employ a multistep procedure, first manually encoding the headwords of each section into DanNet, and thereafter automatically enlarging the DanNet vocabulary by encoding the semantic similarity of the other adjectives in the subsection with a default relation to the headword.

Section 5 addresses additional information to be transferred from DDB to DanNet such as thematic and sentiment information. Finally in Section 6 we conclude.

2. Adjectives in wordnets and similar resources

Adjectives are generally recognized as being indeed very challenging to categorize from a lexical-semantic perspective, mainly because of their plasticity in the sense that they have an extreme ability to take colour from their surroundings. In other words, a core semantic description which is somewhat stable across a certain number of contexts seems even more difficult to provide for adjectives than for other content words (cf. Cruse 1986; Pustejovsky 1995; Bick 2019; Peters & Peters, 2000; and others).

While the structuring feature of wordnets is basically the hyponymy relation between synsets, it has been argued that adjectives are maybe better characterized by their polarity and antonymy relations, their scalarity, their connotation (positive, negative), or simply by the semantics of the external argument (typically a noun) that they prototypically affiliate to.

Consequently, in many wordnets adjectives are to some extent only rudimentarily described and with a not too specific taxonomic labeling. This can be seen as a pragmatic approach in order to be able to cope with their extensive semantic variability.

Exceptions to this are wordnets that have developed their own very elaborate feature scheme for adjectives after thorough analysis, such as GermaNet (Hamp & Feldweg, 1997) with a specific class hierarchy for adjectives of around 100 types relating basically to the semantics of the prototypical external argument of the adjective. Maziarz et al. (2015) describes a set of adjective relations in the Polish WordNet 2.0 based on the principles of especially PWN and EuroWordNet combined with specific lexico-semantic features of the Polish language. Bick (2019) also suggests an annotation scheme of approx 100 taxonomically structured tags partly based on the semantics of the external argument (such as Human, Action, and Semiotic product etc.).

In comparison, Peters & Peters (2000) provides a slightly different description model for adjectives with a primary distinction between Intentional (as in *former president*) and Extensional (as in *American president*), respectively, and a further subdivision according to meaning components such as social, physical, temporal, intensifying etc. The model was developed for the computational lexicon project SIMPLE (Lenci et al. 2001), but was to our knowledge never implemented at a larger scale - maybe due to its complexity.

Previous pilot studies on the Danish adjectival data (Nimb & Pedersen 2012) support the idea that the semantics of the external argument of the adjective can actually function as an appropriate classification scheme, indicating for instance that *bekymret* ('worried') is a prototypical property of human beings whereas for instance *groftskåren* ('coarsely cut') prototypically relates to food items. For Danish, these features can for some of the adjectives be derived from the DDB and might be considered in future transfer of lexical information from the thesaurus to the wordnet.

In DanNet as it stands, the adjectives, like nouns and verbs, are mainly structured according to the EuroWordNet Topontology (Vossen et al. 1998). They are encoded primarily in terms of the ontological type Property combined with a limited set of meaning components, such as Mental and Physical as seen in table 1.

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

Table 1: Ontological types assigned to adjectives in DanNet

In fact, these meaning components can also be interpreted as referring indirectly (and coarsely) to the type of the external argument of the adjective in context. In other words, an adjective of the type Property + Mental will relate to humans, as in *en bekymret politimand* ('a worried policeman'), whereas an adjective of the type Property + Time will have a temporal entity as its external argument, as in *en lang* _uge_ ('a long week').

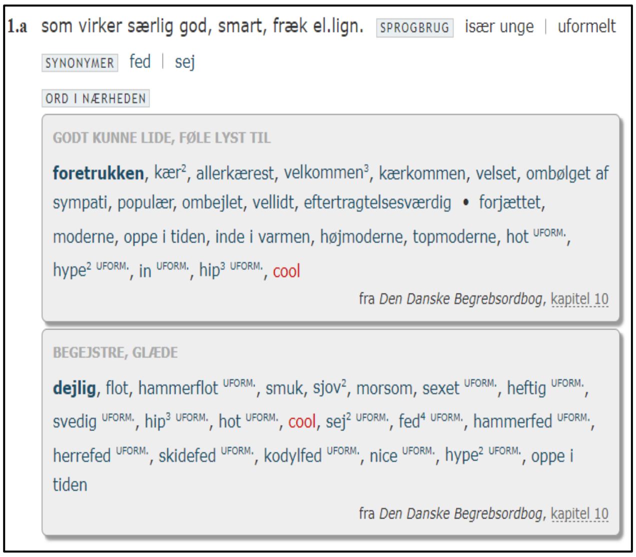
In addition, some adjectives are encoded wrt. their positive or negative connotation.

3. Adjectives in the DDB

In contrast, the thesaurus DDB presents adjectives from a thematically point of view in 22 named chapters (e.g. *Følelser* ('Feelings, emotions')), 888 named sections (e.g. *Vrede* ('Anger') and *Tristhed* ('Sadness')), which are furthermore divided into subsections, initiated with a headword. All the other words in the same subsection are closely semantically related to the headword. The grade of similarity ranges from full synonymy over near synonymy to weaker similarity like hyponymy or just relatedness.

The adjectives in DDB are linked to the sense inventory of the DDO dictionary. The sense links between the two resources and the keyword information in DDB have already shown very useful for the automatic presentation of near synonyms to senses in the online DDO (ordnet.dk/ddo), see Nimb et al. (2018). Exactly which adjectives to extract and present is based on the automatic calculation of the scope of the headword as well as on the further division of the headwords' subsection into even smaller groups of very related words, expressed in terms of dots in the boxes in figure 1. The figure illustrates the near synonyms of the adjective *cool* ('cool; smart') in the online DDO. The focus of the DanNet2 project is to investigate to which degree these principles can be reapplied in the semi-automatic extension of the number of adjectives in DanNet.

The thesaurus contains most of the approx. 13,000 DDO adjective lemmas and represents 90% of the 17.000 adjective senses of the dictionary. Most of them, also the headwords, are not yet included in DanNet where only 17% of the senses are represented. To illustrate this, consider the subsection headword *smittefarlig* ('contagious') in figure 2 where neither the headword, nor any of the semantically related adjectives in its subsection are presently in DanNet, these being *smittebærende* ('contagious'), *virulent* ('virolent'), *patogen* ('pathogen'), *smitsom* ('contagious'), *kontaminøs* ('contaminated'), and *epidemisk* ('epidemic'). However, the noun *smittefare* ('risk of infection') is.



1.a som virker særlig god, smart, fræk el.lign. SPROGBRUG især unge | uformelt

SYNONYMER fed | sej

ORD I NÆRHEDEN

GODT KUNNE LIDE, FØLE LYST TIL

foretrukken, kær^{2}, allerkærest, velkommen^{3}, kærkommen, velset, ombølget af sympati, populær, ombejlet, vellidt, eftertragtelsesværdig • forjættet, moderne, oppe i tiden, inde i varmen, højmoderne, topmoderne, hot $^{\text{UFORM.}}$, hype^{2} $^{\text{UFORM.}}$, in $^{\text{UFORM.}}$, hip^{3} $^{\text{UFORM.}}$, cool

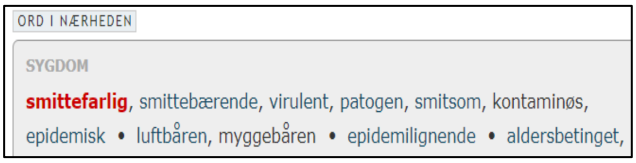
fra *Den Danske Begrebsordbog*, kapitel 10

BEGEJSTRE, GLÆDE

dejlig, flot, hammerflot $^{\text{UFORM.}}$, smuk, sjov^{2}, morsom, sexet $^{\text{UFORM.}}$, heftig $^{\text{UFORM.}}$, svedig $^{\text{UFORM.}}$, hip^{3} $^{\text{UFORM.}}$, hot $^{\text{UFORM.}}$, cool, sej^{2} $^{\text{UFORM.}}$, fed^{4} $^{\text{UFORM.}}$, hammerfed $^{\text{UFORM.}}$, herrefed $^{\text{UFORM.}}$, skidefed $^{\text{UFORM.}}$, kodylfed $^{\text{UFORM.}}$, nice $^{\text{UFORM.}}$, hype^{2} $^{\text{UFORM.}}$, oppe i tiden

fra *Den Danske Begrebsordbog*, kapitel 10

Figure 1. The adjective *cool* ('cool, smart') in the online DDO, with thesaurus data presented in boxes. The first box is extracted from the section *Godt kunne lide; føle lyst til* ('to like, to be fond of, fancy') initiated by the headword *foretrukken* ('preferred'). The second box is extracted from the section *Begejstre; glæde* ('please, make happy') initiated by the headword *dejlig* ('nice').



ORD I NÆRHEDEN

SYGDOM

smittefarlig, smittebærende, virulent, patogen, smitsom, kontaminøs, epidemisk • luftbåren, myggebåren • epidemilignende • aldersbetinget,

Figure 2. The headword *smittefarlig* ('contagious') in DDB. None of the 7 adjectives in the subsection limited by the first dot are presently part of DanNet.

Four out of five sections in the thesaurus contain adjectives (710 of the 888 sections). In particular, the chapters regarding human thinking, behavior and appearances do. There are many adjectives describing feelings and emotions (chapter 10), as well as volition and action (chapter 9), and also many describing social life (chapter 15). This also goes for 'physical' life (chapter 2) where we find the many adjectives for looks and physical conditions, e.g. diseases. Also thesaurus sections describing understanding, knowledge and opinions (chapter 11) contain quite a lot of adjectives. We find lesser adjectives in chapters on e.g. artifacts and food. See table 2.

DDB chapter number and name (in English)	%	Examples of adjectives (in English)
10.Feelings, emotions	11	‘angry’, ‘happy’
15.Social life	9	‘famous’, ‘hostile’, ‘married’, ‘foreign’
9.Will, volition, act, action	9	‘lazy’, ‘active’, ‘stubborn’
2.Life	8	‘young’, ‘blond’, ‘ill’
11.Cognition, thinking, reflection, reasoning	7	‘wise’,’ clever’, ‘thought out’
7.Sense, impression, sensation, state of matter	6	‘cold’, ‘warm’, ‘fluid’, ‘gaseous’
5.Condition,characteristics	6	‘possible’, ‘optional’, ‘sudden’
4.Size, amount, number, degree	5	‘big’, ‘small’, ‘huge’, ‘numerous’
12.Sign, communication, language	5	‘French’, ‘open-mouthed’, ‘clear’
6.Time	4.6	‘late’,’ early’, ‘simultaneous’
20.Economy, finance	4	‘economical’,’ rich’, ‘poor’
18.Society	3.5	‘political’, ‘conservative’, ‘ministerial’
13. Science	3	‘scientific’, ‘mathematical’
3.Space, shape	3	‘round’, ‘triangular’
19.Equipment, machinery, devices, artifacts	2.7	‘wowen’, ‘patterned’, ‘computer-based’
21.Court, legal system, ethics	2.4	‘legal’, ‘illegal’, ‘immoral’
1.Nature, environment	2	‘polar’, ‘rainy’, ‘ecological’
16.Food and drink	2	‘hungry’, ‘spicy’, ‘hard boiled
8.Place, motion	1.8	‘fast’, ‘slow’, ‘trafficked’
14.The arts and culture	1.7	‘artistic’, ‘cultural’, ‘poetic’
22.Religion, supernatural	1.3	‘religious’,’ islamic’, ‘Christian’
17.Sport and leisure	1	‘well-trained’, ‘football-wise’

Table 2. The 22 chapters and their share of the total number of adjectives in DDB, ranged from the highest share (11%) to the lowest (1%) (average 4.5%).

4. Transfer method and data

The transfer is carried out in three steps:

Initially the 766 adjectives which are headwords in the thesaurus (some of which in more than one section), are manually inserted into the wordnet hierarchy representing properties. This includes manual assignment of the appropriate ontological type and is a time-consuming task. The hypothesis is that the headword senses are probably also good candidates for central concepts in the wordnet. The lexicographer carefully studies the headword and its surrounding words in the thesaurus, as well as the existing wordnet hierarchies and the ontological values of the already encoded adjective synsets before the new adjective synset is created and linked to a hypernym, preferably at the very top level of the taxonomy. Already existing ‘top’ adjective synsets in DanNet sometimes also have to be adjusted according to the new adjectival taxonomy.

Secondly, all other adjectives from the headword group in the thesaurus are extracted into synsets in DanNet. They are selected automatically by applying the same method as illustrated in figure 1 (see Nimb et al. 2018), and assigned a) the ontological type of the headword and b) the default relation ‘near synonym’ to the headword.

As a third and final step, the automatically transferred synsets are manually validated. When appropriate, they are changed into co-synset members or hyponyms of the headword instead of the default value ‘near synonym’. This step will be combined with extracted information on synonyms in DDO in order to insert some of the adjectives as an extra synset member instead of the default value ‘near-synonym’.

Most of the adjective senses in DDB (64 % 11,000) only have one representation, making the method straightforward to follow in these cases. However, 22 % of them are part of two sections, 10 % of three, and 4% of even four or more sections. These cases of multiple representations of the same adjectival sense in the thesaurus are a challenge. We have chosen to let headword representations overrule non-headword representations. In the case where the adjective sense is never a headword but represented more than once, we relate it to the headword having the largest number of words in its scope. According to this rule, *cool* in figure 2 would be inserted as a near-synonym to *dejlig* (‘nice’) in

DanNet, and not to *foretrukken* (‘preferred’) in the second box.

The method also allows us to improve the thesaurus. We plan to look closer into the approx. 200 adjectives which are represented in five or more sections. Especially the 20 adjective senses which are represented in 6 up to 9 sections will be checked in order to see whether they are in fact overrepresented and should rather be removed from some sections.

5. Additional information on adjectives that can be transferred

In the initial phase of DanNet2, we also compile a sentiment list with a high lexical coverage based on the polarity values of DDB thesaurus sections. We plan to transfer also this *polarity information* to the wordnet (which already contains this information for a small part of the vocabulary as previously mentioned) relying again on the shared id numbers across our resources. By doing so, we enable DanNet to be used for sentiment analysis. The DanNet2 sentiment list is compiled in a rather efficient way due to the fact that many thesaurus sections contain almost only positive or negative words, respectively. The manual annotation of the 888 DDB sections was the starting point. ¼ of the 888 sections were estimated to contain polarity words based on the section name – 122 annotated to be negative (e.g. ‘Unimportant’ and ‘Sadness’), 80 to be positive (e.g. ‘Important’, ‘Admire’ and ‘Friendship, amity’), and 12 to be more unclear cases, however estimated to be relevant to include in a sentiment lexicon (e.g. ‘Reputation’ and ‘Protest, uprising’). The annotated values were transferred to all the words in the section and manually checked, and words that did not convey polarity of any kind were assigned a zero value.

The more challenging part of this task is to find an objective way of including scalable values to the default polarity annotation. We study the polarity degree of the words in existing sentiment lexica for Danish (Nielsen 2011) with a much smaller lexical coverage. The high negative or positive degree is expanded manually to the near-synonyms in the thesaurus sections when appropriate. Following this line further, also the section and chapter numbers and names from DDB (all translated into English) might be valuable information to include in DanNet. It allows for the identification of *thematically related vocabulary* in the wordnet addressing what is sometimes labelled ‘the tennis problem’ of wordnets (meaning that wordnets generally do not resemble thematic relatedness well). This could be useful especially when it comes to adjectives that are difficult to categorize from a taxonomical point of view.

6. Conclusions

In this paper we have accounted for the aims and initial steps of the DanNet2 project. The first phase of the project has focused on examining the DDB adjective data, and establishing a qualified procedure for semi-automatic transfer of the adjective vocabulary from DDB into DanNet based on the same principles that have already proved useful in the automatic presentation of selected thesaurus data in the online dictionary DDO. In the case of the transfer of thesaurus data to a wordnet, a major challenge is the possible multiple representation of the same word sense in the thesaurus, reflecting again the previously discussed feature of variability which is so characteristic for adjectives This is the case for 1/3 of the adjective senses we plan to transfer. We have discussed different ways of dealing with this problem and described a method which combines the manual encoding of a rather small part of the adjectives, namely those that are headwords in the thesaurus, with the semi-automatic transfer of the rest and much larger part of the adjective vocabulary.

We intend to do a validation of the manually inserted headwords along with the validation of the automatically transferred synsets in order to ensure consistency. Since the method is based on carefully edited and already validated data in the published DDB, we expect to end up with high quality data. Another issue not quite clear yet is how much time and resources the transfer task will require.

Last but not least, we have looked into how sentiment information from a sentiment word list which we simultaneously compile in the DanNet2 project, and which is also based on the thesaurus, could be fruitfully integrated into DanNet. Furthermore we have discussed some future ideas on how to transfer thematic information from the thesaurus into the wordnet.

References

Bick, Eckhard (2019). A Semantic Ontology of Danish Adjectives. In *Proceedings of the 13th* *nternational Conference on Computational Semantics - Long Papers*. Gothenburg, Sweden, 2019.

Cruse, D.A. (1986). *Lexical Semantics*. Cambridge University Press.

Hamp, Birgit and Helmut Feldweg (1997). GermaNet - a Lexical-Semantic Net for German. In *Automatic Information Extraction and Building of Lexical Semantic Resources for {NLP} Applications.* https://www.aclweb.org/anthology/W97-0802.

Hjorth, Ebba & Kristensen, Kjeld (eds.) (2003-2005). *Den Danske Ordbog, volume 1-6*, Det danske Sprog- og Litteraturselskab/Gyldendal, Copenhagen, Denmark. Online: ordnet.dk/ddo

Lenci, A.; Bel, N.; Busa, F.; Calzolari, N.; Gola, E.; Monachini, M.; Ogonowski, A.; Peters, I.; Peters, W.; Ruimy, N.; Villegas, M. & Zampolli, A. (2000). 'SIMPLE – A General Framework for the Development of Multilingual Lexicons'. In: *International Journal of Lexicography 13*. 249–263.

Maziarz, Marek, Stanislaw Szpakovicz, Maciej Piasecki (2015). Semantic Relations among Adjectives in Polish WordNet 2.0. A New Relation Set, Discussion and Evaluation. In Cognitive *Studies | Études cognitives*. 149-179. https://doi.org/10.11649/cs.2012.011.

Nielsen, Finn Årup (2011). A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*. Volume 718 in CEUR Workshop Proceedings: 93-98.

Nimb, Sanni, Nicolai Hartvig Sørensen & Thomas Troelsgård (2018). From standalone thesaurus to integrated related words in the Danish Dictionary. In: *Proceedings from Euralex 2018*, Ljubliana, Slovenia.

Nimb, Sanni, Henrik Lorentzen, Liisa Theilgaard, Thomas Troelsgård (2014). *Den Danske Begrebsordbog*, Det Danske Sprog- og Litteraturselskab & Syddansk Universitetsforlag.

Nimb Sanni, Lars Trap-Jensen, Henrik Lorentzen (2014). The Danish Thesaurus: Problems and Perspectives. In: Andrea Abel, Chiara Vettori & Natascia Ralli (eds.), *Proceedings of the XVI EURALEX International Congress: The User in Focus*. 15-19 July 2014. Bolzano/Bozen: EURAC Research, pp. 191-199.

Nimb, Sanni, and Bolette S. Pedersen (2012). Towards a richer wordnet representation of properties – exploiting semantic and thematic information from thesauri." In *LREC 2012 Proceedings*. Istanbul, Turkey, 2012.

Pedersen, Bolette Sandford, Sanni Nimb, Jørg Asmussen, Nicolai Hartvig, Sørensen, Lars Trap-Jensen & Henrik Lorentzen. 2009. DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation* 43(3). 269–299. https://doi.org/10.1007/s10579-009-9092-1.

Peters, Ivonne & Wim Peters (2000). The treatment of adjectives in SIMPLE: Theoretical observations. In *LREC 2000 Proceedings*. Athen, Greece, 2000.

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

Vossen, Piek & Bloksma, Laura & Calzolari, Nicoletta & Roventini, Adriana & Bertagna, Francesca & Alonge, Antonietta. (1998). *The EuroWordNet Base Concepts and Top Ontology*.