A Danish thesaurus, *Den Danske Begrebsordbog* (Nimb et al. 2014) has recently been published by the Society for Danish Language and Literature (DSL). It presents 192,000 words and expression in 22 chapters and 888 thematic sections in which the words are presented in five word class subsections. The book is just one example of usage of the underlying, annotated XML manuscript where the vocabulary of the sections are grouped in subsections according to meaning, not word class, and furthermore formally annotated. In this talk I will describe the structure of the underlying XML manuscript and the types of semantic and thematic information which can be extracted from it. Furthermore, I will refer briefly to some experiments where this information has already proved its use, before I present some ideas on future exploitation of the thematic information in different projects at DSL.

The compilation of the thesaurus

The thesaurus was compiled on the basis of the 110,000 lemmas and expressions described in a corpus-based monolingual dictionary, *Den Danske Ordbog* (DDO, ordnet.dk/ddo). The hyponymy hierarchies established in *DanNet* of half of DDO's lemmas (Pedersen et al. 2009) were also exploited, making it easy to establish preliminary groups of words denoting e.g. 'furniture', 'clothes', 'animal' etc. in the relevant sections in the thesaurus. The three lexical resources are linked together at sense level, meaning that we are now able to combine the data in many different ways.
The Thesaurus Structure

The words and expressions in the 22 chapters and 888 sections of the XML manuscript are grouped in 8,300 subsections. For instance in the section “Nysgerrig” ('Curious') one subsection with words denoting the property 'to be curious', one with words for curious persons (kødrand ('circle of spectators'), pilfinger ('busybody')), and one with words for something making you curious (e.g. globryllup (wedding ceremony which you attend out of curiosity, as an uninvited guest), teaser ('teaser ad')). Each of the subsections contain formalized information in a header according to a restricted set of coarse-grained semantic types and semantic relations, established on the basis of our experiences with the compilation of the Danish WordNet (Pedersen et al. 2009). The formalized information makes it possible to identify groups of events, acts and properties as well as different hyponomy groups such as persons, artifacts, etc. throughout the XML manuscript. Fig. 1 describes the structure.

The structure also allows us to assign numbers representing thematic information at three levels to all words and expressions in the thesaurus, e.g. the noun lilleputland ('tiny country') belongs to one subsection of 8,300, one section of 888 (named “Stat, nation” ('State, nation')) and one chapter of 22 (named “Samfund” ('Society')). At the most fine-grained level the number is not yet connected to a topic in the form of a title, but we plan to identify one semi-automatically, based on the extraction of the first keyword of each subsection. We also plan to disambiguate the text strings in the 888 section titles, due to the fact that the strings in most cases occur as keywords with sense ids in the section, typically initiating one or two of the first subsections.
Fig.1: The XML structure of the thesaurus, exemplified with the subsection containing hyponyms of the noun *stat* ('state') initiated by the keyword *stat*_headword1 in the section “18.02 Stat, nation” in the chapter “Samfund” ('Society').

**USE OF THE XML STRUCTURE AND THE ANNOTATIONS**

The extractable semantic and thematic information has already been exploited in different projects. A Danish frame lexicon covering communication and cognition senses has been established on the basis of 'act' subsections (identified via their type number) of all sections covering these topics in the thesaurus, combined with the valency patterns of the identified vocabulary in DDO (Nimb & Pedersen, forthcoming). The method proved to be quite efficient, and funding has been provided to extend the lexicon in 2016 and 2017. A method of
supplying the Danish WordNet DanNet with more adjectives, as well as with information on the external argument of adjectives, based on information in the 'property' subsections in the thesaurus (also identified via their type number) is described in Nimb & Pedersen (2012). Experiments have also been carried out on the automatic transfer of the three thematic levels from the (at that time not yet completed) thesaurus data to DanNet. In the first test of the transferred data, information at the most coarse-grained level (the 22 chapters) was used to subdivide co-hyponyms in the WordNet thematically, and two percent of the assignments were manually judged in order to see to which extent the subdivisions of large hyponymi groups (eg. the many co-hyponyms of 'person') made sense. In spite of the coarse-grainedness, only 5 percent were not considered fully intuitive, and the experiment showed a considerable enrichment of the WordNet-data, see Nimb et al. (2013) for more details. Finally we should also mention that the printed thesaurus has been used to manually compare the lexicon of six medieval Danish romances of chivalry with two different provenances in order to obtain a clearer picture of the thematic differences between the texts (Akhøj Nielsen 2015).

**Future plans**

Apart from modern dictionaries, DSL also publishes dictionaries for elder Danish as well as classical Danish literature, in many cases on the basis of highly structured XML manuscripts. We consider the thesaurus a central lexical resource and plan to use it in several ways in the future. By creating (semi-automatic) links between the vocabulary of elder Danish dictionaries to the vocabulary of our modern Danish dictionary, DDO, the common ids between DDO and the thesaurus make it possible to transfer the thematic divisions to
these dictionaries, opening up for new ways of using and presenting the elder vocabulary. Along the lines of the manual comparison of the words in medieval texts mentioned above, the idea is also to use the thesaurus to identify topics automatically in the digitized editions of elder Danish literature published by DSL (ADL.dk), allowing new types of search facilities in the texts. It implies of course a manual extension of the thesaurus with the elder Danish words that cannot be linked to a modern equivalent. As a first step towards our goals, we are involved in a research application with the University of Copenhagen aiming at identifying topics automatically in the digitized editions. In the project, results based on standard topic modelling methods will be compared with results where the thematic information from the thesaurus is added to the automatic process.

LITERATURE


DanNet. See wordnet.dk, and andreord.dk (Anders Johannsen)

OOO. Hjorth & Kristensen (2003-2005), Det Danske Sprog- og Litteraturselskab, Copenhagen. Online 2015: ordnet.dk/ddo


