The Danish Thesaurus: Problems and Perspectives

Sanni Nimb, Lars Trap-Jensen, Henrik Lorentzen
Society for Danish Language and Literature
sn@dsl.dk, ltj@dsl.dk, hl@dsl.dk

Abstract

In this paper, we present a new thesaurus for Danish and discuss some of the problems and decisions that the compilers have been faced with. The thesaurus is compared to other thesauri: Roget, Dornseiff and particularly to its smaller Danish predecessor Andersen. The different steps in the compilation process are outlined, with special attention being devoted to word ordering at the lowest level (alphabetical vs. semantic) and to the use of stylistic labels. In a comprehensive thesaurus, the conclusion is that semantic ordering is more useful for the user and that stylistic labels are necessary.
Keywords: thesaurus; onomasiological dictionary; semantic ordering; stylistic labels

During the last four years, a new Danish Thesaurus (*Den Danske Begrebsordbog*, DDB) has been compiled at the Society for Danish Language and Literature (DSL). Funded by the Carlsberg Foundation, the thesaurus is the first of its kind in 70 years and will at first be published as a printed book. The future plans are to publish it in an electronic version online at DSL's dictionary site *ordnet.dk* where it will be integrated with other dictionary resources at DSL. As it is completely based on, and systematically linked to, the approx. 100,000 lemmas and 135,000 word senses of the corpus-based The Danish Dictionary (DDO), we will have a thesaurus of modern Danish which contains all types of words, not least a wealth of compounds which is a common word type in a Germanic language like Danish. All types of fixed phrases and a number of frequent collocations from the DDO are given in the thesaurus. Well represented in the DDO, interjections and interjectional phrases are listed in the thesaurus together with semantically related verbs. Taxonomic vocabulary such as words for plants, animals, food, diseases, technical devices etc., is classified from a layman's point of view in common language, and all domains are thoroughly represented, including taboo areas. As an important side effect, the approach of common sense id numbers in the DDO and the DDB allows us to enrich a large number of the DDO lemmas and senses with supplementary information about synonyms and semantically related words in a future updated version of the DDO simply by making use of the thesaurus data. Experiments have likewise been carried out on the reuse of thesaurus data in the Danish WordNet, also based on the DDO and using the same sense id numbers (Nimb & Pedersen 2012; Nimb et al. 2013).
Unlike the comprehensive approach that we have adopted, some thesauri, for example the Norwegian Rosbach (2001) and our predecessor for Danish, *Dansk Begrebsordbog* (Andersen 1945, DB), focus instead on the core concepts of the language. We will discuss and compare these two approaches, focusing specifically on the difference between the DDB and the DB. Afterwards we will argue that a comprehensive approach has as consequence 1) that a semantic word ordering is necessary, and 2) that styli-

stic labels are necessary. But first a brief account of the different stages involved in the compilation of the DDB.

1 The Thesaurus-making Process

The basis of the compiling process of the DDB is an XML document representing all senses and their corresponding lemma forms in the DDO, supplied with all relevant semantic information from the dictionary: definition, domain, synonyms etc. On the basis of the sense unit document, the stock of words for the thesaurus is retrieved. As a case in point, consider the compilation of the section 'Bicycling'. First, the lexicographer retrieves all sense units where the string *cykel* is part of the definition and all units of which the corresponding lemma string begins or ends in *cykel*, picks out the relevant sense units and inserts them into the thesaurus document. Synonyms and near synonyms from the DDO as well as words discovered by introspection are added and linked to the DDO if not already part of its stock of lemmata. The selection of words are grouped into semantic categories, such as *persons* riding a bike (cyclist, rider etc.), concrete *objects* (bikes, saddles etc.), *events* (to cycle, to ride a bike etc.) or *properties* of the persons, objects etc., headed by the best representative of the group (annotated in the structure). They may also be put into groups where only a thematic relation holds between the words; this option is mainly used for cases that are difficult to categorize: single words with few near synonyms but nevertheless belonging to the thematic section. In this way, we make sure that not only the prototypical words and senses are covered but also the grey or peripheral areas of the vocabulary or the radial members of the categories established (cf. Dirven & Verspoor 2004: 33).

All the groups are tagged with formalized information about their type of category, and within each semantic group, the words are presented in a logical order according to semantic criteria (see below). In the digital document, semantics is the sole basis for classification. In a later phase, data is converted in order to present the material in four main divisions according to word class for the printed thesaurus (see below). But initially, verbs and their verbal nouns belong to one group tagged with the type 'act' and sometimes also followed by interjections related to the act. Words designating properties are grouped together, no matter if the property is expressed in the form of an adjective (*happy*), a noun (*happiness*), a verbal expression (*to tread/walk on air*) or an adverbial (*on cloud nine*), here exemplified by English words and expressions.

If we again turn to the theme of bicycles, the section 'Bicycling' is one out of 29 sections in the chapter 'Sports and leisure', which in turn is one of 22 chapters that make up the full thesaurus. In total there are 888 sections in the thesaurus – some chapters have only 20 sections while others have more than 60. The chapters and sections were based on the division found in Dornseiff's *Der deutsche Wortschatz nach Sachgruppen* (2004), but have been thoroughly adjusted in all the cases necessary. New sections were added for mental diseases, for contraception and abortion, for medicine, labour market and unemployment, to mention but a few. Other sections were removed, such as the German sections *Bote*

('messenger', instead included in the section on communication), and *Autoindustrie* ('motor industry'), since they were either judged less important categories in modern Danish conceptualization or could easily be covered by other sections.

For various – not necessarily technical or logical – reasons, the grant was allocated for a printed dictionary. That is the reason why this paper is primarily concerned with the printed dictionary, but in all aspects of the compilation and editing process, data has been organized in a way that allows digital exploitation in a later phase, whether as an independent online dictionary or as an integrated semantic component of the DDO.

The manuscript for the printed thesaurus is produced on the basis of the thesaurus XML document. For each section, all words are extracted automatically and presented in four main groups according to word class (nouns, verbs, adjectives, remaining words) with clear marks (in the form of bullet points) of the shifts between semantic categories within each word class group, e.g. between the nouns for persons and the verbal nouns. The logical order within each semantic category, e.g. within the list of persons, respectively verbal nouns, is maintained in the text. The annotated initial words of each semantic group are automatically presented as keywords in the printed text. Some manual adjustment is still needed after the conversion, e.g. in the cases where a word end up being the only member of a group, or in case highlighted words conflict or have too large scope.

2 Comprehensive versus Restricted Thesauri

Where the aim of the DDB is to present all types of words in the DDO including many compounds, the editors of the predecessor DB chose to present only a restricted selection of Danish concepts. Andersen himself states that the vocabulary of the thesaurus was reduced to only a part of the vocabulary found in the contemporary monolingual Danish dictionary *Dictionary of the Danish Language* (ODS), and that he intentionally did not cover the vocabulary in detail, in contrast to the German thesauri which he also used as a model to improve the overall macro structure (Dornseiff's 2nd edition from 1940 and 3rd edition from 1943), and also in contrast to the approach of the contemporary Swedish thesaurus (Bring 1930, cf. Andersen 1945: X-XI). This should be viewed in light of the fact that a large amount of already categorized words were at his disposal in a manuscript of c. 1,000 pages established during a period of 25 years by a schoolteacher in Copenhagen. The manuscript was considered suitable for publishing provided that a thorough revision and modernisation was carried out, for which Andersen was responsible. He left out a substantial amount of compounds and dialect words, as well as many words for natural objects and other concrete objects, the advantage of this being, he writes, a dictionary which is "klarere og mere hændig" ('clearer and more handy'; Andersen 1945: X).

Another recent Scandinavian thesaurus, Rosbach (2001) has a limited stock of words for Norwegian. In our opinion, the restricted collection of words might be a reason why neither Andersen (1945) nor Rosbach (2001) have become widely used dictionaries in their respective countries, compared to the

position of *Roget's Thesaurus* (2002) within the English-speaking community (Hüllen 2009: 40 and 44). Roget and also Dornseiff (2004) for German constitute valuable linguistic resources, offering a large variety of words and expressions as it is the main purpose of thesauri, namely to support language variation and provide linguistic inspiration to users in text producing situations (cf. Hüllen 2009: 29 and 46).

3 The Comprehensive Approach: Lexicographic Challenges

The comprehensive approach involves some lexicographic challenges since some of the methods of the restricted thesauri are not applicable when the contents increase, both in terms of the number of lexical units and the types of words and expressions included. The first challenge concerns the order in which the words are presented; the second concerns the treatment of words which are not part of the unmarked standard vocabulary.

3.1 Semantic versus Alphabetical Word Ordering

The DB organizes words in word classes which are divided into different semantic groups (objects, events, persons etc.) separated by dashes. Persons are always preceded by double dashes as the final element of a noun group. Within each semantic group (between dashes) words are listed in alphabetical order, meaning that the first word is not necessarily a keyword for the whole group. For instance, words within the category 'milk' are presented in the order "Fløde, Kærnemælk, Mælk, skummet Mælk, Piskefløde, Sødmælk, Yoghurt" (cream, buttermilk, milk, skimmed milk, double cream, whole milk, yoghurt). In the case of words for 'water', we find "Brøndvand, Drikkevand, Gaasevin, Isvand, Kildevand, Kommunevand, Vand" (well water, drinking water, plain water, ice water, spring water, tap water, water). In both cases, the hypernyms (*milk* and *water*) are placed between different types of hyponyms and do not function as keywords for the established semantic category.

Dornseiff (1940 and 1943) used the same alphabetical presentation. In the 8th edition (2004), though, we find an important difference to this: a keyword, typically the hypernym, has been moved from its place in the alphabetical order to the initial position of the group. The only cue to the keyword function is the break in the alphabetical order, which may be challenging for the user. Consider the case of words for different types of dairy products in Dornseiff (2004): "Milch · Buttermilch · Joghurt · Kefir · Magermilch · Rahm · Sauermilch" (milk, buttermilk, yoghurt, kefir, skimmed milk, cream, curdled milk). One notes also that the choice of alphabetical order inevitably breaks down logical ordering, placing *skimmed milk* between different types of sour milk rather than next to *milk*, of which it is a direct hyponym. Another example from Dornseiff (2004) concerns words for coffee: "Kaffee · Blümchen · Cappuccino · Espresso · Lorke · Milchkaffee · Mokka · MuckeFuck" (coffee, weak coffee (informal), cappuccino, espresso, bad weak coffee (dialect), café au lait, mocha, coffee substitute/ersatz coffee (deroga-

tory)), where one could argue that the informal word *Blümchen* and the dialect word *Lorke* have too prominent positions and would according to logic be better placed after the different types of coffee. Where a restricted thesaurus, such as the DB, can get away with an alphabetically ordered list of a small amount of words within a semantic category, this is in our opinion not suitable for the comprehensive thesaurus. The higher the amount of words, the more disturbing the alphabetical order becomes. This is the case in Dornseiff (2004) where a given word is most likely to be semantically more closely related to the initial keyword of the whole semantic category than it is to its immediate neighbours. In that way, the semantics of one word cannot be used to understand the next word and thereby activate a forgotten word in the user's mind. Furthermore, the distance from a word to the keyword can easily get very long. By contrast, Roget (2002) presents words only in semantic order and has in fact done so from the very first edition (Hüllen 2009: 40), a principle also adopted by Bring (1930). Initial keywords at the highest level of the taxonomy are graphically highlighted in Roget (2002) by italic types.

> *soft drink*, teetotal d., non-alcoholic beverage; water, drinking w., filtered w., eau potable, spring water, fountain; soda water, soda, ..., coffee, café au lait, café noir, black coffee, white coffee, decaffeinated coffee, decaf, Irish coffee, Turkish c., espresso, cappuccino, latte ...

Example 1: Roget's Thesaurus (2002), soft drinks (excerpt).

In the DDB, we implement the same type of semantic ordering as Roget. It relies entirely on the lexicographer's judgment (cf. Hüllen 2009: 29), following two principles which go hand in hand, as reflected in linguistic theories on prototypes. The first implies that the prototypical, or central, members of a category are presented before less prototypical, radial members (see for example Dirven & Verspoor 2004: 17 for a description). An example is the section in the DDB on furniture, where chairs and beds are placed before lamps and carpets. The second principle is based on the idea of basic level categories in a language, from which a division of the vocabulary into three conceptual levels may be derived: a generic, a basic and a specific level (Dirven & Verspoor 2004: 37). Following this principle, basic and general level terms will be placed before specific level terms in the thesaurus. In many cases, the general level term is used as the title of the section in question and the basic level terms are highlighted as keywords. Illustrated by English words, general level terms such as *animal*, *plant* and *furniture* constitute section titles and are in their corresponding sections presented before basic level terms such as *dog*, *tree* and *bed*. Both the general and the basic level terms (i.e. *animal* as well as *dog*) are marked as keywords and listed in the thesaurus before the specific terms, in this case kinds of dogs, trees or beds, for example *poodle*, *oak tree* and *double bed*. In the case of coffee, the DDB presents the words as seen in example 2.

kaffe, mokka (uformelt); espresso, café au lait, caffé latte, cappuccino, macchiato; filterkaffe, stempelkaffe, pulverkaffe, kolbekaffe, tyrkisk kaffe; sort kaffe; jordemoderkaffe (uformelt), mokka; en lille sort;

termokaffe; varmekaffe; mosevand (slang); en kop kaffe, kaffetår, refill; morgenkaffe, formiddagskaffe, eftermiddagskaffe, aftenkaffe

(**coffee**, mocha (informal); espresso, café au lait, caffé latte, cappuccino, macchiato; drip coffee, press pot coffee, instant coffee, coffee made in a coffee maker, Turkish coffee; black coffee; very strong coffee (informal), (a cup of) strong coffee; a cup of black coffee laced with spirits; thermo jug coffee; warmed-up coffee; bog water (slang); a cup of coffee, cup of coffee, refill; morning coffee, mid-morning coffee, afternoon coffee, evening coffee)

Example 2: Coffee words in the DDB.

An argument against the semantic ordering principle is that it becomes more difficult for users to find a specific, already known word as they cannot rely on the alphabet when browsing through a group. Instead, they must look via words that come closest in meaning in the text, and this is not always an easy task. To support overview and browsing as much as possible, we have chosen to highlight more keywords in the text than Roget does. In the case of soft drinks in example 1, also hypernyms at a lower level in the taxonymy will be highlighted, i.e. also *water*, *soda water* and *coffee*. Furthermore, keywords at the highest taxonomic level are presented in boldface in the printed DDB in order to obtain a clearer visual signal than the italics used in Roget. As a consequence, finally, the index is organized such that for each entry word the following information is given: entry, keyword, section indicated by its number, part-of-speech. This is similar to the solution in Roget and different from Dornseiff that refers only to section number and title.

3.2 Stylistic Labels

Comprehensive thesauri cover a much broader range of stylistic varieties, accentuating the need for stylistic labels. In this section we will discuss the types of information used in the DDB and compare it with DB, Roget and Dornseiff.

The DB does not have any information about register, for example we find several informal words for nose simply listed without any comment (*Mule* 'muzzle', *Næse* 'nose', *Næsebor* 'nostril', *Snabel* 'trunk', *Snude* 'snout', *Snydeskaft* 'hooter, conk', *Tryne* 'snout', *Tud* 'snout, schnozzle'). Nor does Dornseiff use stylistic labels, which in our opinion sometimes leads to confusing lists of words, for example those concerning coffee mentioned above where dialect words (*Lorke* (bad weak coffee), informal words (*Blümchen* (weak coffee), and *MuckeFuck* (substitute/ersatz coffee) are presented alongside standard German words for coffee (*Kaffee*, *Cappuccino*, *Espresso*, *Milchkaffee*, *Mokka*). Roget, on the other hand, uses labels to some extent but gives no precise description as to the use of these. Since the DDB contains a substantial number of slang and informal words from the DDO, we have in line with Roget chosen to assign labels of the stylistic and temporal status to the words which are not part of the standard vocabulary. The labels we use are extracted from the information in the DDO but presented in simplified

form. The detailed set of values in the DDO are here converted into five types: four stylistic (derogatory, informal, slang, jocular) and one temporal (archaic). Regional language is rare in the DDO and therefore not included in the set of labels. The method of direct transfer, however, is problematic, due to the fact that the labels were originally applied in relation to other senses of the same word and maybe a few synonyms given in the DDO entry, not to a large group of synonymous expressions as it is in the DDB. Manual adjustment is therefore needed in many cases. Only in the event where a few labelled words occur between numerous unmarked words in a semantic group, can they be kept as they are, without further editing. In other cases, we need to adjust the text in order to achieve homogeneous information about linguistic style. This is particularly relevant where the labels of adjacent words clash when seen in context. In these cases, we harmonize the information by choosing one label to cover both, if possible. For example, *lampe* ('lamp') and *pære* ('light bulb') are two expressions for 'brain' in Danish, the former labelled 'slang', the latter 'informal' in the DDO. In the DDB, both are labelled 'informal' as there is no clear-cut boundary between slang and informal language. In the case of large groups of synonymous expressions of the same stylistic value, the label of the initial keyword indicates that the following words have the same value. Information about temporal status is treated independently of other labels. Example 3 shows the case of derogatory words for women in Danish, before (1) and after (2) editing.

(1) **kælling** (neds.), kone, gås (neds.), tante (neds.), tøs (neds.), hundyr, furie, rappenskralde (neds.), strigle (neds.), mokke, hystade, skude (neds.), sæk, (neds.), tudse (neds.), so (neds.), smatso (neds.), klidmoster (neds.), kran (slang), madamme (neds.), sladretaske (neds.), rendemaske (slang, gammeldags), sladderkælling, havgasse (neds.), harpe (neds.), ribs (neds.)

(**bitch** (derogatory), woman, silly goose (derogatory), aunt (derogatory), hussy (derogatory), female animal, fury, shrew (derogatory), termagant (derogatory), fishwife (derogatory), battleaxe (derogatory), virago (derogatory), cow, hag (derogatory), gadabout, .. etc.

(2) **kælling** (neds.), gås, tante, tøs, hundyr, madamme, klidmoster, furie, mokke, harpe, strigle, ribs, hystade, rappenskralde, havgasse, sladderkælling, sladretaske, rendemaske (gammeldags), skude, kran, sæk, tudse, so, smatso

bitch (derogatory), silly goose, aunt, hussy, female animal, fury, shrew (old-fashioned), termagant, fishwife, battleaxe, virago, cow, hag, gadabout, .. etc.

Example 3: Derogatory words for 'woman' in the DDB. The group is initiated by one label instead of keeping the automatically inserted labels from the DDO on each word.

A manual adjustment is required in approx. 1/8 of the 888 semantic groups in the DDB, concentrated in certain semantic areas such as men, women, drinking alcohol, body parts, sexuality and bodily functions, but also physical punishment, conflict, scolding and others. For the larger part of the DDB, however, stylistic labels occur only sparsely and are typically kept the way they appear in the DDO.

4 Conclusion

It is a challenge to compile a comprehensive thesaurus which truly reflects the vocabulary of a semasiological monolingual dictionary. Bringing such a project to a completion within a period of just a few years' time can only be successful by using computational methods and by means of a well-structured model which guides the lexicographer's categorization of the words and, maybe most importantly, offers ways of placing the many radial concepts of the language. The alphabetical ordering of the words adopted by previous works is impracticable and must be replaced by semantic guidelines to ensure a consistent logical order within the large vocabulary of each category. The close connection between the two dictionaries makes it possible to reuse data in various ways. In this paper, we have shown how stylistic labels from the dictionary can be transferred to the thesaurus, and in the future our plan is to extract information about the semantic relations between words in the opposite direction, from the thesaurus into the dictionary, in this way adding an onomasiological component to the way users may access dictionary data.

5 References

Andersen, Harry (1945). *Dansk Begrebsordbog*. København: Munksgaard.

Bring, S. C. (1930). *Svenskt ordförråd ordnat i begreppsklasser*. Stockholm: Hugo Gebers Förlag.

DDO = *Den Danske Ordbog*. Accessed at: http://ordnet.dk/ddo [11/04/2014].

Dirven, Rene & Marjolijn Verspoor (2004). *Cognitive Exploration of Language and Linguistics*. Philadelphia, PA, John Benjamins Publishing Company, USA.

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

Dornseiff, Franz (1943). Der deutsche Wortschatz nach Sachgruppen, 3. Auflage.

Dornseiff, Franz (1940). Der deutsche Wortschatz nach Sachgruppen, 2. Auflage.

Hüllen, Werner (2009). Dictionaries of synonyms and thesauri. In A. P. Cowie: *The Oxford History of English Lexicography, vol. II, Specialized Dictionaries*. Oxford: Oxford University Press, pp. 25-46.

Nimb, S. & B. S. Pedersen (2012). Towards a richer wordnet representation of properties – exploiting semantic and thematic information from thesauri. In *LREC 2012 Proceedings*. Istanbul, Turkey, pp. 3452-3456.

Nimb, S., B. S. Pedersen, A. Braasch, N. H. Sørensen & T. Troelsgård (2013). Enriching a wordnet from a thesaurus. In Workshop Proceedings on Lexical Semantic Resources for NLP from the 19th Nordic Conference on Computational Linguistics (NODALIDA). Linköping Electronic Conference Proceedings; Volume 85 (ISSN 1650-3740).

ODS = *Ordbog over det danske Sprog*. Vol. 1-28 (1918-1956). Det Danske Sprog- og Litteraturselskab og Gyldendal. Online version at: http://ordnet.dk/ods [11/04/2014].

Roget, Peter Mark (2002). *Roget's Thesaurus*, 150th anniversary edition edited by George Davidson. London: Penguin.
Rosbach, Johan Hammond (2001). *Ord og begreper. Norsk tesaurus*. Oslo: Pax Forlag A/S.