Access to Multiple Lexical Resources at a Stroke: Integrating Dictionary, Corpus and Wordnet Data

Lars Trap-Jensen[1]
Society for Danish Language and Literature, Copenhagen, Denmark

Abstract

The paper presents a lexical resource, *ordnet.dk*, which brings together data from two dictionaries – both originally print dictionaries, one historical and one modern – with a contemporary reference corpus and a wordnet, all with Danish as the object language. Focus is on data exploitation across the components, dealing with onomasiological queries in the dictionary based on wordnet data, and on cross resource look-up possibilities from the three components.

Keywords: wordnet, dictionary, corpus, resource integration, Danish

1. Project background

It is quite likely that the technological prospects of e-media will gradually change the dictionary as a genre. No compelling reason exists why dictionaries should be confined to dealing with words and their descriptions. Many e-dictionaries already include spoken pronunciation and pictures, and why not continue with encyclopaedic articles, grammar paragraphs, and translation services until, eventually, the day arrives when we simply ask our computer a question and it provides us with the appropriate answer. However, until that day we must take one step at a time, and in this paper I will look at some of the very first steps we have taken towards increased resource integration based on our experience with a Danish online dictionary site providing access to two monolingual dictionaries and a corpus of contemporary Danish.

Ordnet.dk was developed during a six-year project period that ended in 2009. The interface gives separate dictionary and corpus access but at the same time their contents are combined in various ways. Furthermore, the contents have been supplemented with wordnet data for a new function in the online version.

A few words should be said about the original resources. Both dictionaries were conceived for print publication but at different times and under different circumstances. The *Ordbog over det danske Sprog* (The Dictionary of the Danish Language, henceforth ODS), is a historical dictionary in 28 volumes which was

[1] Society for Danish Language and Literature, Christians Brygge 1, 1219 Copenhagen K, DENMARK, ltj@dsl.dk

published between 1918 and 1956. Together with its five supplementary volumes, which appeared between 1992 and 2005, it covers approximately 250,000 words from the period 1700-1950. The digitization of the original manuscript was carried out as part of the current project (see *e.g.* Lorentzen and Trap-Jensen 2008), and a preliminary version has been publicly available since 2005. With regard to the current topic of integration, the ODS is, however, the least relevant of the three components.

Den Danske Ordbog (The Danish Dictionary, henceforth DDO) is where most of the integration is being explored. It is a dictionary of modern Danish covering the period following the ODS, *i.e.* from 1950 onwards. Conceived and published (2003-2005) as a print dictionary in 6 volumes, it was, however, prepared with the use of modern methods and technology. The SGML format of the data was converted to XML and considerable effort has been made to restructure the data to improve it for screen publication as part of the current project.

KorpusDK is a reference corpus of contemporary Danish. A subset of it was built as an integral part of the DDO project, this dictionary being the first corpus-based dictionary compiled for Danish. Spoken language and texts that were restricted for privacy or copyright reasons (such as private letters and diaries) were removed and new texts added when the corpus was made public under the name of *Korpus 2000* (see *e.g.* Andersen *et al.* 2000). It was mainly the name and design that were changed when it became *KorpusDK* as part of the current project but new texts have been collected on a regular basis since 2005.

DanNet was constructed on the basis of words and senses taken from the DDO in a joint work between the Society for Danish Language and Literature and the Centre for Language Technology at the University of Copenhagen (see Pedersen *et al.* 2009). Data from DanNet are used in the online version of the DDO.

It is worth noting that the data are more compatible than one might assume at first glance. The ODS and DDO were compiled at times when both theory and practice were different but even so, they were developed by the same institution, the Society for Danish Language and Literature, and within the same tradition of descriptive lexicography. There is an intimate relation between the empirical basis of the DDO and (parts of) KorpusDK, and the same is true of the DDO and the Danish wordnet, DanNet.

2. Related words in DDO

As a new feature, the online DDO offers "related words" for a substantial number of word senses. "Related words" is a thesaurus-like function which is particularly useful for language production and for (advanced) language learning purposes. It assists users in "finding the right word" when writing a text and in developing their communicative skills to express themselves creatively, with nuance and accuracy. For language learners, it provides an overview of a semantic field that is important in vocabulary

training as words are not learnt in isolation but rather in comparison to words with similar meanings.

As data from the Danish wordnet is used for this feature, a few clarifying remarks about this resource are in order[2].

Wordnets are language technology resources that have primarily been used in informations systems, for example for information retrieval, word sense disambiguation and artificial intelligence applications. The basic unit is the *synset*: a set of one or more synonymous words that express the same concept. Each word sense belongs to a particular semantic class – called *ontological type* – established by a rough division of the conceptual world into approximately 200 semantic classes based on principles known from traditional componential analysis. Examples of ontological types are *Natural+Substance* (*ice, lava, sand*), *Plant+Object+Comestible* (*avocado, carrot, tomato*), *Human+Object+Occupation* (*accountant, nurse, taxi driver*) and *UnboundedEvent+Agentive+Mental* (*reflect, analyze, think*).

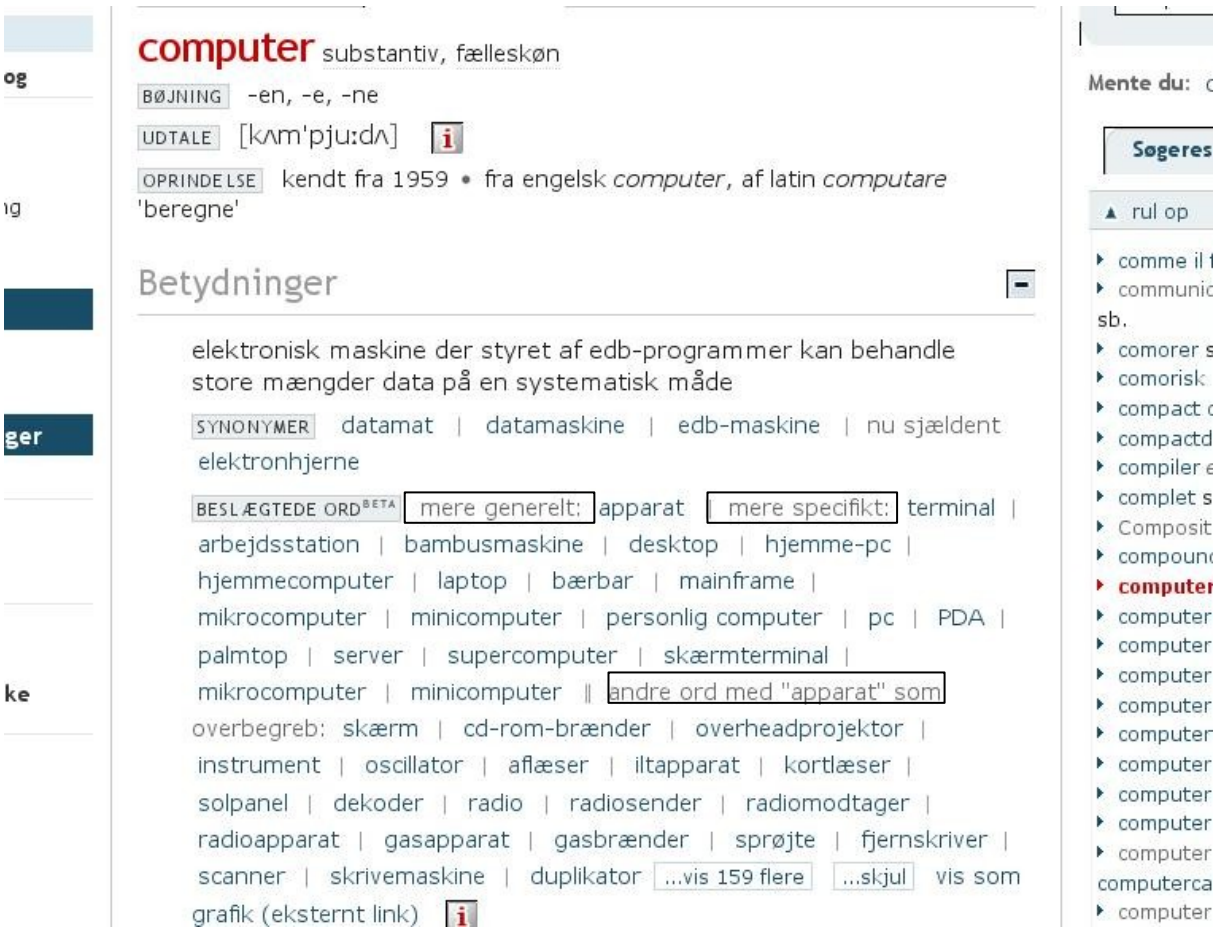

Figure 1. Information on related words in The Danish Dictionary.

[2] The account of DanNet data in DDO is based on the account given in Sørensen and Trap-Jensen (forthcoming).

A synset can be related to other synsets through various semantic relations, in DanNet a total of 18 are used. The most commonly used relations are: hyponymy, hyperonymy, part-whole, antonymy, near-synonymy, used for, concerns and involved agent. The relations have been encoded for each individual sense and the coded outcome is what is used to calculate candidates for "Related words".

An example is shown in Figure 1. For obvious reasons, this and the following examples are in Danish but hopefully they are internationally understood. According to the underlying hyponymy hierarchy, related words are selected from three levels: more general words (indicated by “mere generelt” in Figure 1) from the superordinate level are taken from the level immediately above, *i.e.* the word or words serving as *genus proximum*; more specific words (“mere specifikt” in Figure 1) are taken from the level below, *i.e.* among the hyponyms; and finally, the last group contains words at the same level as the sense looked up (indicated by the heading “andre ord med “apparat” som overbegreb” in Figure 1), *i.e.* words that are co-hyponyms or sister terms.

The first group, the hyperonyms, is straightforward as there will always be a limited number of candidates in this group, in most cases just one. It is possible to have several word occurring as hyperonyms but only in the event that a) a concept is expressed by two or more synonymous words: the word *jeep* has as its hyperonym the synset consisting of *car, auto, automobile, machine, motorcar*, or b) if a word has more than one hyperonym: in DanNet the word for *roller skate* (Danish: *rulleskøjte*) has been encoded as a hyponym of both the synset *footwear, footgear* and of the synset *sporting requisite.* Accordingly, all the synonyms appear as more general words for *jeep* and *roller skate*, respectively.

More problematic are the co-hyponyms as there may be several thousands of them in the extreme cases. To help selecting the best suitable words, a score has been calculated to express the similarity between the entry word and the co-hyponyms. The entry word in the relevant sense is compared to each of the co-hyponyms, and first the ontological types are considered: the greater the similarity between the ontological types, the higher the score. Next, the relations describing the two are compared: having many relations in common yields a higher score but complete accordance is not obligatory. Finally, not only the number but also the kind of relations encoded is of importance: thus *petrol car* is more similar to *diesel car* than it is to *crane lorry* because although all three share the same relation HAS_PART, the relevant parts for the former two – *petrol engine* and *diesel engine*, respectively – belong to the same ontological type as opposed to the HAS_PART = ‘crane’ of the last.

Based on the similarity score, the co-hyponyms are presented in descending order. The list has been reduced to the 30 highest scoring words but with the possibility to see up to 200 as a clickable option.

The most problematic group is the list of hyponyms. As with the co-hyponyms, the list of hyponyms can be long but, unlike the co-hyponyms, we have found no meaningful

automatic way of selecting the best candidates. If we look up the word *car*, should *petrol car* be considered more relevant than *crane lorry*? So far, we do not know the answer and, as a provisional solution, we simply show a list of up to 200 words, randomly reduced. This is by no means expedient and for future updates we hope to develop a better method of presenting the information, for instance by grouping the hyponyms in relevant types.

2.1. Comparison with *Macmillan Online Dictionary*

A similar thesaurus function is offered by *Macmillan Online Dictionary* but in this case the contents have been manually edited. Figure 2 shows the thesaurus entry for 'car'.

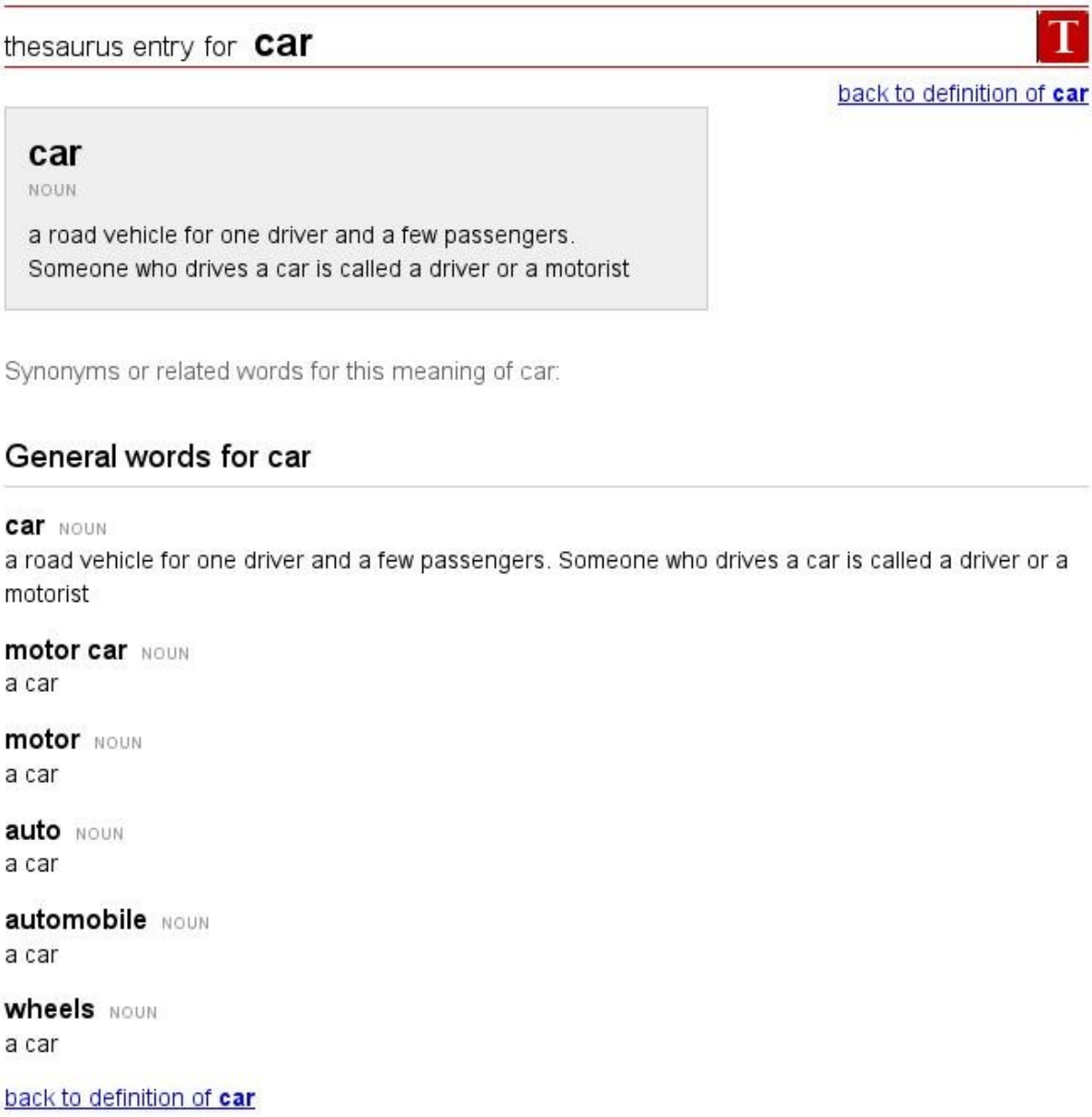

Figure 2. Macmillan's thesaurus entry for 'car'

This is an alternative way of doing things and it is instructive to compare the results. The first thing to notice is that the sheer number of words is much more manageable due to the fact that the words belong to more or less the same level of abstraction. In many cases this gives just the relevant alternatives for the user trying to find other words in text production. Conversely, several of the examples in Figure 1, *e.g. overhead projector, oscillator* and *scanner*, are not likely ever to become real paradigmatic alternatives for *computer*. The problem of over-generation of candidates from DanNet is connected with the number of ontological types. Most thesauri from Roget onwards use 800-1,000 semantic groups whereas the norm of about 200 ontological types used in wordnets is bound to result in more members per group for a given vocabulary size – unless it is combined with other criteria, such as the ontological type of the target sense for a given relation. An example of an extreme case is the rich vocabulary connected with 'person'. Because of the lack of more subtle taxonomic subdivisions, a word like *catholic* has 3185 co-hyponyms, including *hippie, fascist, godfather, gourmet, ecologist* and *cat owner*, words that have little in common apart from the fact that they denote persons.

Another difference is that the heading of a superordinate term in DanNet as well as all the category members are always themselves words in the language. It is an essential feature of DanNet that it should reflect a "natural" categorization of the world, *i.e.* corresponding to the lexicalized labels for concepts of the Danish language. This is arguably the major difference between a wordnet and an ontology, and we deliberately wanted to avoid introducing conceptual categories that lack linguistic counterparts. We regard this as a strong point of DanNet, in particular with respect to its primary use in language technology. In thesauri for human users, however, the need for categories of a manageable size is more important than having categories with only single word headings. This remains a problem when it comes to the presentation of wordnet data in comparison with the manually compiled thesaurus.

On the other hand, the manual approach has its problems, too. However commendable the effort to arrive at numerically manageable categories, it all depends on the meaningfulness of the headings chosen. Take *chinchilla* as a case in point: if you are looking for alternatives in Macmillan you arrive at a category "Mammals found in North, Central and South America", with 27 members. My guess is that you are as likely to be interested in other rodents or in other pets as you are in *alpaca, coyote, grizzly* or *caribou* as alternative words for *chinchilla*. Likewise, if you are looking for other words for 'off-road vehicle' under the heading "Vehicles used away from roads and on snow" you will not find *Land Rover* because this has been assigned to "Makes of car"; and *four-wheel drive* is found under "Equipment and systems in cars and other road vehicles", whereas *jeep* has been placed under "Military and industrial vehicles".

And although Macmillan is generally economical and to the point, they sometimes also face the problem of having too many or ill-suited words. If you look up the word *confectionery*, for example, you arrive at a group labelled "Types of food or drink", a

sizeable group with almost 50 members but not many of them obvious alternatives for *confectionery*: *aphrodisiac, baby food, creole, macrobiotic, nutraceutical, slop* and *wholefood*, to name but a few. I hasten to emphasize that this is not a general impression: If, instead of *confectionery*, you look for *sweets*, the group "Sweets and other confectionery" contains an equivalent number of, pardon the pun, very palatable examples.

In our case, the solution to the over-generation of hyponyms seems to be either to introduce sub-categories manually, especially towards the more abstract end of the semantic cline, corresponding to Macmillan's groups of "general words for *person, vehicle, machines,* etc.", or to develop a quantitative method that would allow us to rank and select the most appropriate words as it is done for co-hyponyms. But at the time of writing we have not accomplished this, which is why a small "beta" sign has been attached to the function label for "Related words".

Finally, to meet the needs of language learners and others interested in systematic vocabulary training and semantic fields we have long wanted to bring an overall visual presentation of related words, along the lines of *The Visual Thesaurus* and other viewers that allow the user to browse wordnet data.

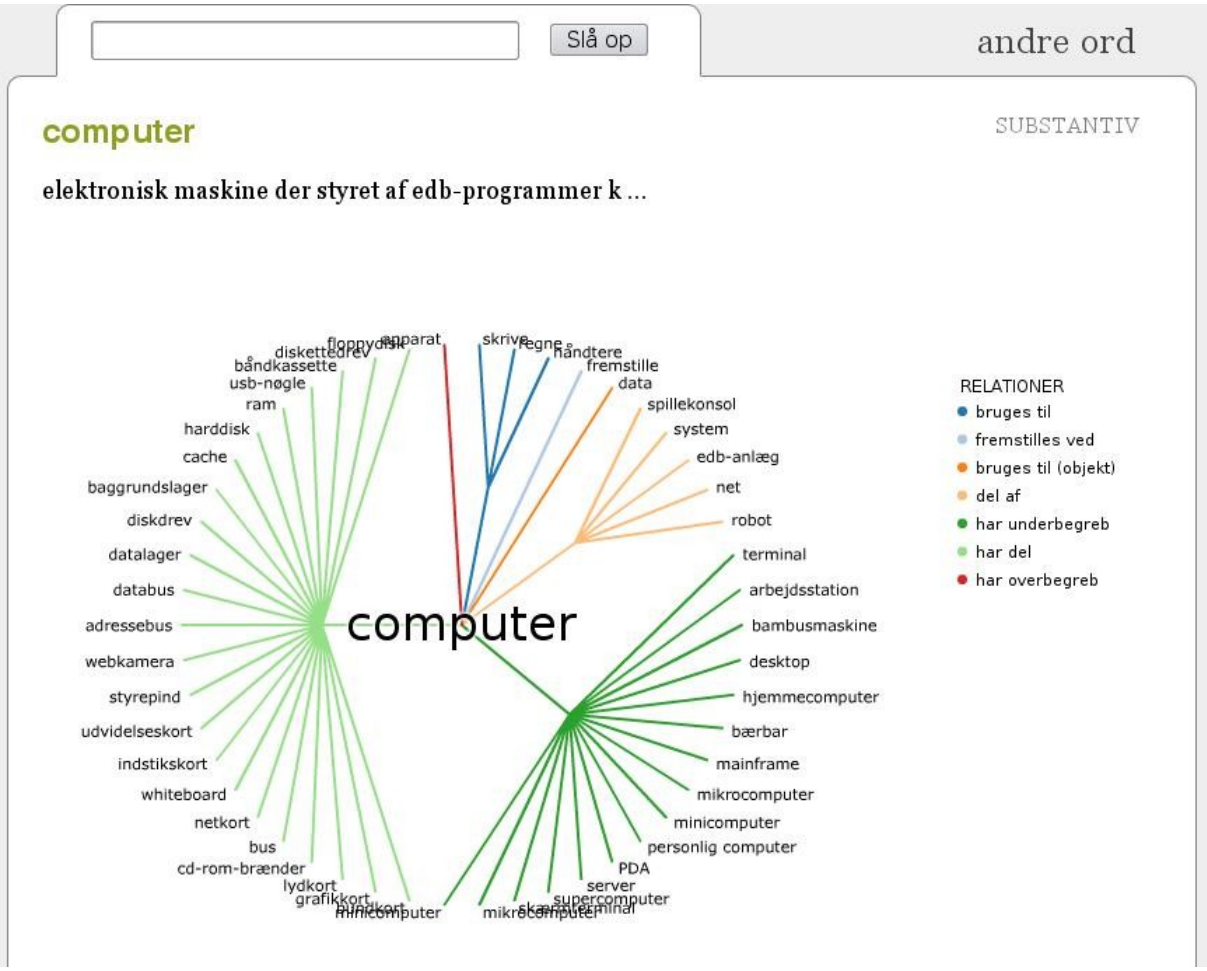

Figure 3. Visual representation of "computer" in andreord.dk

But as sometimes happens when you release data as open source, others do the job for you. This is what happened for us when we discovered andreord.dk ('other words'), a site that does precisely that. So, instead we have chosen to link to this external resource for the visual presentation. An example is shown in Figure 3.

Apart from the visual presentation, the site also provides a search box, an extract from the definition, synonyms, one or more example sentences, the path of hyperonyms, hyponyms and the ontological type.

3. Dictionary and corpus data

Dictionary data are linked with corpus data in various ways, as shown in Figure 4.

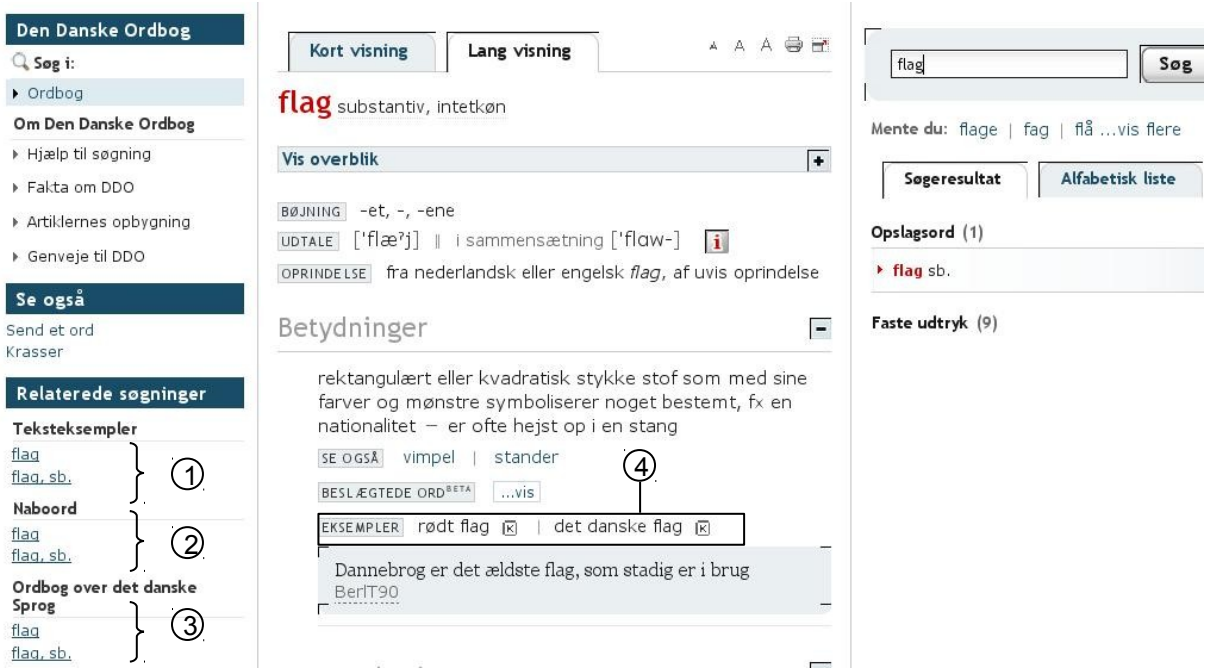

Figure 4. Look-up possibilities in KorpusDK and ODS offered by DDO

Figure 4 shows the entry for 'flag' (same word in English) in DDO. In the left column the user can:

- look up the relevant lemma in the corpus (1);
- calculate collocates of the lemma (2);
- look up the corresponding lemma in the ODS (3);

And, in the centre column, the user can

- look up corpus examples of the specific collocations shown (4).

For all three options in the left column, the user can choose whether the query should be for a) the string, or b) a particular part of speech. The latter is particularly useful in case of homonymy.

The "K" icon (K for *korpus*) in the centre column is clickable and when activated it submits a query for that collocation. The result for "rødt flag" (red flag) is shown in Figure 5 as a common concordance display with the searched words highlighted[3].

Figure 5. Look-up possibilities in KorpusDK, DDO and ODS

The alternative query options are, as always, found in the left column (under the heading "Relaterede søgninger" in Figure 5). Here it is possible to:

- calculate collocates of any of the constituting words, based on the possible lemma forms (1);
- get a list of fixed phrases containing any of or all the constituting words (2);

[3] The default setting for multi-word queries allows up to three intervening words – hence the varying number of highlighted words – but the setting for multi-word queries can be customized at the user's will.

- look up the multi-word expression or any of the constituting words in the DDO, as a string or as a specific POS (3);
- look up any of the constituting words in the ODS, as a string or as a specific POS (4).

The list of fixed expressions in (2) is itself derived from DDO. It represents a subset of multi word expressions (bearing in mind that the DDO was itself based on evidence from a subset of KorpusDK), *viz.* those expressions that were selected for lemmatization by the editors during the manual compilation process. Concordances for words or expressions are easily generated with a click, both from the list of collocates and from the list of fixed expressions.

5. Perspectives

This is roughly the current state of affairs but, obviously, things do not stop here. Among the priorities for continued development are the following:

First, we need improved tools for tagging and parsing new corpus material. At present, the KorpusDK contains c. 56 million tokens and has not changed since it was released in 2002. Texts have been collected continuosly since 2005 but await mark-up. Provided we can develop new and expedient mark-up procedures, we hope to supply lemmas, variants and inflectional forms with corpus frequency information. With proper syntactic mark-up we would also be able to offer look-up possibilities for the valency patterns given in the grammatical section. And with a continuous influx of texts, corpus analysis can be automated to generate candidates for new lemmas in the dictionary.

Another perspective is the development of an integrated separate onomasiological presentation of the DanNet data where the user can query and navigate the semantic hierarchy, *e.g.* through a tree-structure view of the nearest superordinate and subordinate levels. Whether it should be via a separate search page or integrated in the entry (*i.e.* as the current presentation of "Related words") remains an open question at this point. Recently, The Society for Danish Language and Literature has received funding for a three year project to develop a traditional thesaurus based on the data from DanNet. This will allow us to address the problems and shortcomings that have been pointed out here.

Finally, we would like to incorporate more grammatical information in the dictionary. Our institution is involved in the edition of a new comprehensive grammar of Danish and an obvious perspective is to link directly from the grammar sections of a dictionary entry to the relevant paragraph in the grammar.

References

ANDERSEN, M.S., ASMUSSEN, H. and ASMUSSEN, J. (2000). The Project of Korpus 2000 Going Public. In Braasch, A. and Povlsen, C. (eds). *Proceedings of the 10th EURALEX International Congress*. Volume 1, Copenhagen: Euralex: 291-299.

LORENTZEN, H. and TRAP-JENSEN, L. (2008). The Dictionary of the Danish Language Online: From Book to Screen – and Beyond. In: *Lexicographie et Informatique – bilan et perspectives. Pré-actes*, ATILF / CNRS, Nancy-Université: 151-157.

MACMILLAN ONLINE DICTIONARY. http://www.macmillandictionary.com

ORDNET.DK. http://ordnet.dk

PEDERSEN, B.S., NIMB, S.; ASMUSSEN, J., SØRENSEN, N.H., TRAP-JENSEN, L. and LORENTZEN, H. (2009). DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. In: *Language Resources and Evaluation*, Volume 43, Number 3. Springer Netherlands: 269-299.

ROGET, P.M. (1998). *Roget's Thesaurus of English Words and Phrases*. 1998 edition by Betty Kirkpatrick. London: Penguin Books.

SØRENSEN, N.H. and TRAP-JENSEN, L. (forthcoming). Den Danske Ordbog som begrebsordbog. In: Lönnroth, H. and Nikula, K. (eds.). *Nordiske Studier i Leksikografi*. No. 10, NFL-skrift nr. 11, Tammerfors.