There And Back Again – from Dictionary to Wordnet to Thesaurus and Vice Versa: How to Use and Reuse Dictionary Data in a Conceptual Dictionary

Henrik Lorentzen, Lars Trap-Jensen
Society for Danish Language and Literature
Christians Brygge 1, 1219 Copenhagen K, Denmark
E-mail: hl@dsl.dk, ltj@dsl.dk

Abstract

This is a story of elexicographical evolution and how lexical data are used and reused to develop new products and new presentations. At eLex 2009 we demonstrated how data from The Danish Dictionary were used to construct a wordnet for Danish, DanNet, and we showed how, in a betaversion, DanNet data could be used to improve the onomasiological component of the dictionary. Since then, we have used both DanNet and dictionary data in an ongoing project to create a conceptual dictionary – a thesaurus – for Danish. In this article we will show how the thesaurus data help us overcome the major problems connected with the direct use of wordnet data in a dictionary for human users. Focus is on the editing principles of the thesaurus and how data are presented in The Danish Dictionary online.

Keywords: thesaurus; wordnet; onomasiological search; e-dictionary

1. Background

A dictionary stock of 91,500 entries with 116,000 meaning descriptions and a wordnet consisting of 65,000 synsets connected through 75,000 internal semantic relations provide the starting point for a thesaurus project which is currently being developed at the Society for Danish Language and Literature in Copenhagen. Historically, data from The Danish Dictionary (Den Danske Ordbog, DDO) were used to construct DanNet, and since both serve as input for the thesaurus, all three resources are closely interconnected. The Danish thesaurus project set off in 2010 and will appear in 2013 as a publication of its own – and even as a printed dictionary.

In this connection, focus is on the thesaurus data and how they can be used to improve the onomasiological component of the online version of the DDO. In Louvain 2009, we gave a presentation of the dictionary site *ordnet.dk*, of which the DDO is but one element (the others being a comprehensive historical dictionary and a corpus component), and we showed how data from the Danish wordnet were exploited in a new element, *Related words*, that was introduced in the online version of the dictionary (Trap-Jensen, 2010). We demonstrated how candidates for *Related words* could be automatically extracted from DanNet but it was emphasized – as it still is in the version available to the public – that it is a beta version and not without its problems. In this article, we dwell on the nature of the problems involved and how they can be solved by using data from the thesaurus instead.

2. Shortcomings of wordnet data

In Trap-Jensen (2010), some problematic areas were mentioned and possible solutions suggested. Let us briefly recapitulate the central issues as well as other shortcomings that we have encountered since.

First, there is the problem of overgeneration. This pertains to both co-hyponyms and hyponyms and is due to the fact that the categories are often too broad and the hyponymy hierarchy too shallow. From time to time, Princeton WordNet has been criticised for having too deep and detailed a hierarchical semantic structure and it is therefore important to stress that DanNet is not a translation of Princeton WordNet but was built on original Danish data primarily extracted from the DDO, mainly to ensure that the conceptual world reflected by the Danish language is maintained in the resource.

The combination of a relatively shallow semantic hierarchy and a limited number of semantic classes – DanNet operates with approximately 200 ontological types as opposed to the 900-1,000 semantic groups found in thesauri like Roget (2002) for English or Dornseiff (2004) for German – implies a built-in risk of too broad semantic categories resulting in too many and not always the most evident related words when the candidates are automatically extracted from DanNet. Examples in DanNet are the large groups of vocabulary items for persons, plants and verbal nouns with hundreds or, in extreme cases, even thousands of co-hyponyms. To some extent this can be remedied by using combinations of other relations to narrow down the number of members in each category. Our solution to the problem, as described in Trap-Jensen (2010), has been to develop an algorithmic method that ranks words from large groups, primarily based on the number and nature of shared relations. Even if the algorithm has improved usability, the overall conclusion is nevertheless that a good deal of manual effort would still be needed for this element to work properly for the human user.

A second major problem with the DanNet data relates to the fact that synsets are seldom found in more than a single place within the semantic network. For example, the assignment of multiple hyperonyms to synsets,

although by no means impossible or forbidden, is still a relatively rarely used possibility: out of 65,000 synsets in DanNet, less than 500 have been assigned more than one hyperonym. This is primarily a consequence of the editorial process in DanNet: multiple hyperonyms are almost exclusively used when the dictionary definition contains more than one genus proximum. For example, in DDO *daughter* is defined as 'a girl or a woman of whom you are the father or mother' and *juniper* 'little tree or bush that bears juniper berries'. Because DanNet was constructed from the dictionary, the editor would be prompted that 'girl' and 'woman' were both possible hyperonyms of *daughter* and 'tree' and 'bush' possible hyperonyms for *juniper* but no independent routine was carried out to decide whether this was the case for other vocabulary items.

Multiple hyperonyms should not be confused with the situation known as systematic polysemy where a word, e.g. *school*, is ascribed a dual meaning: 'building' and 'institution' respectively. The difference is that the latter situation involves clearly distinct meanings belonging to different synsets. An inheritance mechanism in the editing tool helps the editor make sure that similar words and in particular hyponyms are coded in the same way.

In a traditional thesaurus, the picture is somewhat different. Here it is quite common for a word to appear in different thematic groups: for example, a *guinea pig* is at the same time a 'South American mammal', a 'rodent', a 'pet' and – at least in some parts of the world – an 'edible animal'. In other words, a word in a particular sense is not confined to occur under the nearest hyperonym alone. Sometimes it makes sense to use instead a subset of the hyperonym's hyperonym ('South American mammal' as opposed to 'rodent'), sometimes it is not the taxonomical position that matters at all but the role that the entity plays in a particular context ('pet' or 'edible'). Humans are overall more creative and flexible in the way they encode and decode meaningful categories than computers are.

People's ability to carve up the world in new ways and the assignment of multiple hyperonyms are both reflected in the index of a thesaurus: many words have several references to the systematic part. Of course, one has to remember that not all the references of a word concern the same meaning as the index only lists the form of each word and hence does not distinguish between homographs and different word senses. This is why the notion 'synset' (a set of word forms signifying a single concept) is so important in the wordnet universe whereas it has no counterpart in the common language. However, the point is here that even if we take this into account, it is not unusual for a particular word in a specific sense to be placed in several thematic groups of a thesaurus.

In the end we found that the difference between an NLP resource like DanNet and the human user's need for onomasiological assistance was beyond quick repair. Instead we have decided that the wordnet data should not be displayed directly in the dictionary but only indirectly – through the central role they play in building the thesaurus. How this is done is the subject of the next section.

3. Editing a thesaurus using DanNet and DDO

As a first step in the early phase of the project it was decided to copy the ontological structure of the German thesaurus *Der deutsche Wortschatz nach Sachgruppen* founded by Dornseiff, instead of building an ontology from scratch. This may seem a quick and dirty solution but considering the fact that Danish and German are closely related languages (and cultures) we concluded that the solution was justified. As a matter of fact, the predecessor of our Danish thesaurus (Andersen, 1945) was also to a large extent based on Dornseiff, which as a side effect makes it easier for us to supplement with material from the older thesaurus.

Dornseiff and consequently our thesaurus are divided into 22 main sections (chapters) ranging from *Natur und Umwelt* (Nature and Environment) over *Kunst und Kultur* (Art and Culture) to *Religion*. These chapters are further subdivided into 906 thematic groups. We are fully aware of the risk of taking over a meaning structure from German without it being fully adapted to Danish, and remind the reader that this merely serves as the starting point. Minor changes are to be expected as the editorial work progresses, for instance merging of groups or, on the other hand, splitting groups where Danish and German do not classify the world in exactly the same way.

When starting to edit a new thematic group the editor begins by extracting raw material from DanNet. Typically a central hyperonym, such as *orchestra* within the field of music, is selected as a starting point together with all its hyponyms. Relevant information about each concept such as definition, subject domain and part of speech is copied into the thesaurus database as well as ID numbers and other metadata necessary to ensure links between the two databases. The process can be reiterated any number of times in order to supply more central concepts and their hyponyms but usually this step does not exhaust the material entirely. Additional searches will then be made, notably in DDO, where relevant material can be extracted by combining search parameters like specific words in definitions and a particular subject field. Once an appropriate number of concepts has been obtained, the editor begins to sort the concepts into smaller groups. These groups are held together and labelled by a headword, often the hyperonym. The field of music can serve as a case in point: it contains groups like genre, volume, tempo,

rhythm, each one consisting of hyponyms such as *baroque music*, *chamber music*, *military music* (genre), *crescendo*, *forte*, *pianissimo* (volume), *adagio*, *moderato*, *allegro* (tempo), and *beat*, *triplet*, *syncopation* (rhythm). Generally, properties (adjectives) are grouped together: *instrumental*, *vocal*, *symphonic* as well as verbs: *play*, *practise*, *compose*. Other groups unite places and institutions like *academy of music*, *concert hall*, *discotheque* and parts of musical instruments such as *string*, *keyboard* and *pedal*. In the latter case it is the holonym that serves as headword.

Figure 1: Section from the thematic group *musical instruments* in the editing tool

It appears that DanNet and DDO are by far the most important resources and the main reason why it is possible at all to develop a dictionary of this kind within a limited period of time. Once a satisfactory semantic structure has been established, however, the editor also turns to look at other sources in order to supplement the thematic group in question. These sources include the older thesaurus (Andersen 1945), a Danish dictionary of synonyms (Schultz), a dictionary of slang and informal language (Politiken) as well as other relevant reference works. Schultz is available in electronic format which allows semi-automatic comparisons with the first draft version of the thematic group.

4. Presenting the data

At the moment 145 out of the 906 thematic groups have been completed in a draft version ready for futher editorial treatment, i.e. correction of errors, supplementing missing concepts, rearranging of groups and concepts. As mentioned in section 2, the current presentation of *Related words* will be replaced by thesaurus data once these are available. Let us consider how they can be used to offer fewer but more relevant candidates as *Related words* for a particular word meaning. We can use another example from the field of music, the aforementioned word *discotheque* in the sense 'a place to dance'. This concept is found in the following four groups: *Place*, *Dance, Popular music* and *Pleasure and leisure time*. In the dictionary entry for *discotheque* in the element *Related words*, the user will be presented with four snippets along these lines (with the relevant words in English translations):

1 PLACE discotheque, bar, night club; tivoli, amusement park, theme park
2 DANCE discotheque, dancing stage, ballroom, rehearsal room; show, striptease, ball
3 POPULAR MUSIC music venue, discotheque, jazz venue
4 PLEASURE restaurant, discotheque, night disco, dance restaurant

Each word is provided with a mouse-over function giving the definition from DDO. Furthermore, the headlines of the four groups are clickable, taking the user to a thesaurus presentation where each group of concepts are expanded to the level of thematic group; in the case of *discotheque* the first group is that of *Place* and the user will see a list of concepts denoting for instance places where humans perform an activity, ranging from *workshop* and *laboratory* over *stadium* and *sports centre* to *headquarters* and *executive's office*.

The abbreviated form of *discotheque*, *disco*, can be used in the same sense but also in the sense of 'type of music and dance'. It is important that the user is made aware of this and is allowed to navigate to this new thematic group. S/he can do so either via links to related groups or via a search field present at the top of every thesaurus page. In the case of *disco* in the second sense, the relevant thematic group will contain another type of related words: *dance, disco, breakdance, hiphop, headbanging, limbo, zumba*.

The reason why only snippets are given in the dictionary entry is of course to avoid imposing too much material on a user who perhaps is looking up for entirely different reasons. Only words from the very subgroup in which the entry word appears are displayed. If the user wants to see more s/he will have to go to the thesaurus page for a complete overview.

The thesaurus page shows the entire thematic group with the search word highlighted and with references to other groups that are semantically close. A thematic group contains an average of 270 words and is divided into subgroups on two hierarchical levels. Within each subgroup the first word functions as a heading for the following words.

The structure within the thematic group as well as the group-internal order of words are semantically based so that the reader intuitively recognizes a 'natural' or 'logical' organization and progression as s/he reads through the groups.

This organization also reflects the editor's way of organizing the thematic group. There has, however, been heated discussion among the editors whether this is also

the best way of presenting data. Some thesauri use alphabetical ordering at the lowest level of grouping – among them are Dornseiff and Andersen, whereas the order in Roget is 'logical' or semantical. And all thesauri that we have consulted use part of speech as a more general dividing criterion than semantics.

Following Wiegand (2004) the overall purpose of a monolingual conceptual dictionary depends on the dominant consultancy situation:

Bei der oben formulierten Charakterisierung des genuinen Zwecks von großen einsprachigen Sachgruppenwörterbüchern wurden bei den Mögligkeiten, die Wörterbücher dieses Typs dem Benutzer eröffnen, drei verschiedene Arten [unterschieden:

(i) Konsultationssituation wegen Ausdrucksfindungs-schwierigkeiten
(ii) Konsultationssituation wegen äquivalentbezogener Angemessenheitszweifel
(iii) Konsultationssituation wegen Wissensbedarf über einen bestimmten Aspekt des deutschen Wortschatzes]

Die erste Möglichkeit … wird wahrscheinlich am meisten genutzt. Das ist auch der Grund dafür, warum die Wörterbuchartikel im Formteil nach Ausdrucks-klassen und damit auch nach Wortarten gegliedert sind. Die kundigen Benutzer, die Ausdrucksfindungs-schwierigkeiten haben, werden durch diese artikelinterne Datenanordnung präferiert.

Wiegand, 2004: 59 (i-iii summarized from 57-58)

If it is true that word-finding problems are the most common motive for thesaurus consultation, i.e. that the user is looking for paradigmatic alternatives to the search word, then part of speech seems well-placed as the superior criterion. If, on the other hand, knowledge needs are more common, then the semantic criterion is preferable. In this connection, Wiegand's second situation can be neglected as it refers to bilingual contexts only. In the printed dictionary, a decision must eventually be made, whereas – at least technically – it is an option to have two presentations in the online dictionary and leave it to the user to decide. This is one of the questions that have not yet been answered.

5. Perspectives

A remarkable feature of the resources in *ordnet.dk* is the fact that all the elements are closely interconnected, with mark-up of the building blocks at a fairly detailed level. A system of unique ID numbering maintains links between units, not only between lemmas in the dictionary but also at sub-lemma level between the senses of an entry in the dictionary and the synset members in DanNet and the concepts in the thesaurus. This way of organizing data is also what makes it possible for us to enrich the other resources and in that way facilitate further elexicographical evolution. When the thesaurus is complete the resulting data can be reused to improve the wordnet data. We cannot go into all the details here but just mention a few promising areas.

The thematic information contained in the thesaurus should be imported into DanNet. At present, there is little information about the degree of semantic closeness between the units in DanNet. Synonyms (members of a synset) and near-synonyms are specified if available but beyond that no finer distinctions are made. Even if the hyponymy hierarchy does say something about basic conceptual organization, it does not account sufficiently for the degree of similarity between concepts. Here, the semantic grouping on up to three different levels within a chapter represents an alternative way of categorization which could contribute significantly to the semantic richness of DanNet.

We mentioned earlier that multiple hyperonyms are under-represented in DanNet, primarily because the dictionary definitions usually do not contain more than one genus proximum. The thesaurus editors, however, often assign a second superordinate term, for example when they group objects according to their function instead of, say, physical shape. For example, the dictionary definition of *drumstick* reads 'a long wooden stick used for playing drums', and consequently *drumstick* is coded as 'a kind of stick' in DanNet but not as 'a piece of music equipment'. This situation could be remedied by using thesaurus data.

Finally, the thesaurus would be helpful to broaden the coverage of DanNet. An obvious area is the treatment of 2nd and 3rd order entities (as used by Lyons, 1977), for example properties that are not described in great detail in the current version of DanNet. In the thesaurus, however, adjectives are often grouped together on the basis of the nouns they modify: 'suntanned' is a property of humans, 'wire-haired' a property of dogs, 'carnivorous' a property of predators etc. Properties make up a surprisingly large proportion of the vocabulary – a rough estimate says that about 20 % of the vocabulary covered in the thesaurus so far pertain to properties. Another area where coverage could be broadened is the vocabulary from other parts of speech than the traditional content words as well as multiword units.

6. References

Andersen, H. (1945). *Dansk Begrebsordbog* (Danish Conceptual Dictionary). Copenhagen: Munksgaard.

DanNet. Accessed at: http://wordnet.dk.

Dansk Synonymordbog (1992). (Danish Dictionary of Synonyms), 8th edition. Copenhagen: J.H. Schultz Information Ltd.

Den Danske Ordbog (The Danish Dictionary). Accessed at: http://ordnet.dk/ddo.

Dornseiff, F. (2004). *Der deutsche Wortschatz nach*

Sachgruppen, 8th edition. Berlin/New York: Walter de Gruyter.

Lyons, J. (1977). *Semantics*. Press Syndicate of the University of Cambridge.

Politikens Slangordbog (1993). (Politiken's Dictionary of Slang), 4th edition. Copenhagen: Politikens Forlag.

Roget, P.M. (2002). *Roget's Thesaurus*, 150th anniversary edition edited by George Davidson. London: Penguin.

Trap-Jensen, L. (2010). Access to Multiple Lexical Resources at a Stroke: Integrating Dictionary, Corpus and Wordnet Data. In S. Granger, M. Paquot (eds.) *eLexicography in the 21st Century: New challenges, new applications. Proceedings of eLex 2009. Louvain-la-Neuve, 22-24 October 2009*. Louvain-la-Neuve: Presses Universitaires de Louvain, pp. 295-302.

Wiegand, H.E. (2004). Lexikographisch-historische Einführung. In F. Dornseiff (ed.) *Der deutsche Wortschatz nach Sachgruppen*, 8th edition. Berlin/New York: Walter de Gruyter, pp. 9-100.