

Merging specialist taxonomies and folk taxonomies in wordnets - a case study of plants, animals and foods in the Danish wordnet

Bolette S. Pedersen, Sanni Nimb, Anna Braasch

University of Copenhagen, Society for Danish Language and Literature,

E-mail: bspedersen@hum.ku.dk, sn@dsl.dk, braasch@hum.ku.dk

Abstract

In this paper we investigate the problem of merging specialist taxonomies with the more intuitive folk taxonomies in lexical-semantic resources like wordnets; and we focus in particular on plants, animals and foods. We show that a traditional dictionary like Den Danske Ordbog (DDO) survives well with several inconsistencies between different taxonomies of the vocabulary and that a restructuring is therefore necessary in order to compile a consistent wordnet resource on its basis. To this end, we apply Cruse's definitions for hyponymies, namely those of *natural kinds* (such as plants and animals) on the one hand and *functional kinds* (such as foods) on the other. We pursue this distinction in the development of the Danish wordnet, *DanNet*, which has recently been built on the basis of DDO and is made open source for all potential users at www.wordnet.dk. Not surprisingly, we conclude that cultural background influences the structure of folk taxonomies quite radically, and that wordnet builders must therefore consider these carefully in order to capture their central characteristics in a systematic way.

1. Introduction

In all lexical resources the problem of merging specialist taxonomies with the more intuitive folk taxonomies of the layman emerges to some degree or the other. Where traditional dictionaries survive surprisingly well with several inconsistencies between different taxonomies of the vocabulary, wordnets meant for computational aims can, however, not fulfill their role satisfactorily, unless some consistent methodology is pursued on this matter already in the development phase.

In this paper we investigate a specific case of this problem, namely the clash between the way we organize plants and animals in wordnets, typically highly influenced by the professional taxonomies of these domains, and the more intuitive and culturally-based way we organize foods. The way we organize foods obviously relate strongly to animals and plants but still seem to be quite another matter. A well-known example is the case of tomato which is *either* conceived as a fruit or as a vegetable depending on the view point.

In search for a methodology for a systematic treatment of such clashes of perspective, we apply Cruse's definitions for hyponymies (Cruse 1991, 2002), namely those of *natural kinds* on the one hand and *functional* and/or *nominal kinds* on the other hand. We pursue this distinction in the development of the Danish wordnet, *DanNet*. This wordnet has recently been built on the basis of a traditional dictionary, Den Danske Ordbog (DDO) and is made open source for all potential users at www.wordnet.dk¹).

2. Related work

Reusing traditional dictionaries in order to build wordnets is contrasted by the approach used in most recently

compiled wordnets of other languages, such as the Spanish wordnet (Fernández-Montraveta et al. 2008), the Arabic wordnet (Rodríguez et al. 2008), and the Hungarian wordnet (Miháltz et al. 2008). These wordnets apply the so-called expand approach where they basically develop new wordnets by translating Princeton WordNet (Fellbaum 1998) into the new source language. To our knowledge, only few other wordnet, such as the Polish wordnet (Derwojedowa et al. 2008) and the Norwegian wordnet (Fjeld & Nygaard 2009), apply a monolingually-based approach similar to ours.

This explains why the fundamental problem of how to treat the clash between specialist and folk taxonomies is not necessary a crucial one in the building of new wordnets, the taxonomies being mainly taken over from English and then reorganized for the particular target language. With respect to the food domain, as is further described in Section 3, it seems that Princeton WordNet has dealt more carefully with food senses than what is the case in many traditional dictionaries. To this end, some problems may be avoided using the expand approach, although a more loyal picture of linguistic conceptualisation in a specific language can be given by the monolingual approach.

The 'clash of perspectives' problem is also touched upon by the Chinese wordnet builders (Huang et al. 2008), where it is argued that the establishment of a so-called paronymic relation can help reorganize wordnets, so that sister nodes can now be described from different perspectives without causing an ontological clash (typically known as the ISA overload).

So, even if this discussion may not currently be central in the wordnet community, it is obviously a basic one in several classical studies of general linguistics and terminology. In our work, we refer to Apresjan for his account on regular polysemy (1972), to Wierzbicka (1996) and Cruse (1991, 2002) for their accounts of the varieties of hyponymies. Also recent terminology work by Madsen & Thomsen (2009) is relevant for our work

¹ DanNet is developed in collaboration between Center for Language Technology at the University of Copenhagen and the Danish Society for Language and Literature. The project is granted by The Danish Ministry of Research.

since they tackle the problem of ontological clashes from a terminological view point and suggest a solution similar to Huang's by establishing so-called 'subdividing dimensions' in the concept network.

3. Plants, animals and foods in our source dictionary: DDO

In DanNet, we apply the monolingual approach (also called the "merge" approach since it can subsequently be merged with Princeton WordNet) mainly because DDO, a modern corpus-based dictionary of Danish, was completed in 2005 right before the wordnet project started. It was accessible in a machine-readable version with hyponymy information explicitly specified for each of the approx. 100,000 sense definitions. First of all, this made it possible to build a Danish wordnet on monolingual grounds using semi-automatic methods (cf. Pedersen et al. 2009 for the full account of this process). But not less important, it guaranteed that the senses included in the wordnet were actually frequent in Danish general language texts since the selection of the lemmas and senses in the dictionary was strongly based on their representation in a balanced Danish corpus of 40 million tokens.

However, the monolingual approach also causes problems in the establishment of systematic and consistent wordnet hierarchies within certain domains such as for example foods. First of all, a corpus of 40 millions tokens, as well-balanced as it might be, may very well lack data on meals and recipes including other food ingredients than the most common ones in Danish cooking. This has had as a consequence that many types of food are not described in DDO. In contrast, we do find a very large number of e.g. fishes and vegetables in Princeton WordNet, represented with both an animal (or plant) sense as well as a food sense.

Secondly, we have noticed that many dictionary definitions in DDO tend to contain a genus proximum describing the physical properties of an object rather than containing a genus proximum describing its use or function. The function of the object is instead expressed elsewhere in the definition. In Princeton WordNet we also find such cases. Consider for example the definitions of 'ball' and 'doll' ('ball = round object that is hit or thrown or kicked in games'; 'doll = a small replica of a person, used as a toy'). In the structure of hyponymy relations, however, they have not been assigned the hypernyms 'object' or 'replica' but instead a hypernym expressing their function, in these cases 'game equipment' for ball and 'toy' for doll.

Since the semi-automatic reuse of DDO in the establishment of DanNet relies heavily on the already identified genus proxima, such cases have been treated manually. In DanNet we have aimed at establishing hierarchies for groups such as toys, game equipments, vegetables and foods, but it has not been possible to do it by semi-automatic methods since the genus expression is not useful for this aim. The domain of foods has been the most challenging one, though, due to the fact that the

dictionary descriptions in DDO within the domains that constitute the basis of food, namely animals and plants, in addition are heavily influenced by the fields of specialized domains such as zoology and botany. This has led to a quite heterogeneous description of edible plants and animals, the lexicographers using sometimes general language approaches, sometimes specialized language approaches as they are found in traditional encyclopedic work. The work was furthermore complicated by the well-known cases of regular polysemy between plants and vegetables (as well as between animals and meat) which are so recurrent that they are seen as fully conventionalized. However, regular polysemy has been treated heterogeneously in DDO without applying any general linguistic principles, depending mostly on the frequency of the lemma.

So, with respect to the reuse of a traditional and corpus-based dictionary when establishing food taxonomies in a wordnet, we are confronted with at least three problems:

- 1) A strongly corpus-based dictionary only describes the most common food types, leaving out food which is not represented in the corpus, often at the expense of a systematic treatment of regular polysemy.
- 2) Many dictionary definitions (of concrete objects) tend to use a genus proximum expressing a physical aspect rather than a functional one. A semi-automatic method depending on the genus expression does not work in these cases, and functional hierarchies in the wordnet must instead be established manually.
- 3) Within zoological and botanic domains, a traditional dictionary is strongly influenced by professional taxonomies. Lemmas with a food sense therefore often have as their starting point, or first sense, the biological animal or plant definition rather than a definition describing that they are used as food.

A corpus-based, traditional dictionary has to find a balance between the traditional description of such domains and the principles established about corpus frequency – as we shall see, this is not always easy, sometimes leading to a quite heterogeneous description in DDO, where very rare plant and animal senses have been described in the dictionary in spite of lack of corpus occurrences, whereas some common food senses are left undescribed.

3.1 Plants and vegetables in DDO

We start our examination of DDO by considering a number of randomly chosen vegetable lemmas all of which also have a plant sense. In these data we encounter four different ways of description:

Description type 1:

Lemmas with only one sense, namely a vegetable sense, and with the hypernym *grøntsag* (vegetable) or *rodfrugt*

(vegetable root). This group contains in all 11 lemmas, with examples such as *agurk* (cucumber), *avokado* (avocado) and *aubergine* (aubergine). In these cases the plant sense is not described, either because the lemma does not cover the plant sense or because it is simply left out. These cases are unproblematic to reuse semi-automatically in the establishment of a vegetable hierarchy in DanNet, the genus proximum being functional.

Description type 2:

This group includes lemmas with two senses, both a plant sense and a vegetable sense which has the genus proximum vegetable (or fruit). Examples are lemmas like *tomat* (tomato) and *græskar* (pumpkin). We find five lemmas described in this way. The description of the lemmas are influenced by botanic specialized language since some of the plant senses are rare in general language (like *kål* (cabbage) and *løg* (onion)), but the established vegetable/fruit sense is easy to reuse in DanNet due to the genus proximum.

Description type 3:

In this more problematic group, 12 vegetables have been described as subsenses to a main plant sense and as parts of the plant, and what is important, not as vegetables. None of the lemmas have vegetable as genus proximum but instead words like root, tuber, tap root, beet, and stalk. Thus, it was necessary to look for expressions like 'edible', 'used as food', and 'used as vegetable' to identify the vegetable sense. The influence from specialized language is evident, furthermore because the main plant sense in many cases is infrequent in general language. The lemmas are such as *gulerod* (carrot) and *jordskok* (Jerusalem artichoke). The genus proximum of the plant sense is often taken from specialized language as seen in e.g. *skærmpilante* (umbelliferous plant), and *kurypilante* (composite).

Description type 4:

Also this group has to be treated manually in DanNet. Here we find five vegetables such as *artiskok* (artichoke) and *spinat* (spinach) described only by one sense definition: As plants used as a vegetable, edible, or used as food. The vegetable sense is in these cases not established as a separate sense, and furthermore, the plant sense is infrequent in general language while the vegetable sense is not. Seemingly, the DDO lexicographer has had problems describing the lemma, since in two cases she has chosen the word vegetable as genus proximum in the separate field for genus expressions, although the genus proximum in the definition itself is 'plant'.

Our investigations show that in cases 2, 3, and 4, DDO has been strongly influenced by the traditions from specialized language encyclopedic descriptions when it comes to the treatment of vegetables, and that this has in many cases overruled the overall strategy of DDO to base the sense distinctions on corpus frequency. In 17 of the 33 cases, the vegetable hierarchy in DanNet could only be established by manual means and not by the semi-automatic reuse of the genus expression.

3.2 Animals and meat in DDO

We have also investigated the description of 41 edible animals and 10 edible animal bodyparts in DDO in order to see how, and how often, the food sense was described and whether the genus proximum of the food sense was usable in the establishment of a food hierarchy in DanNet. The lemmas in DDO where the food sense is part of the sense description were found by searching in the definition for 'meat/edible/food etc.', indicating that a food sense was present. Many lemmas with a missing food sense were found manually during the DanNet encoding process from the encoders' own knowledge of the different types of meat we normally eat in Denmark. Other of the 41 edible animals were found by looking at a list of the 2035 lemmas in DDO having the domain information 'zoo' in a special field (hidden in the printed dictionary).

As in the case of vegetables, the ideal case in the perspective of being able to establish DanNet semi-automatically on the basis of DDO would be that the animal or bodypart lemma has two senses (either two main senses or a main sense and a subsense), namely an animal sense with the genus expression 'animal' and a food sense with the genus expression 'food' or 'meat'. But as we shall see, only half of the investigated lemmas have an established food sense of which the genus expression and thereby the food hierarchy can be semi-automatically established. The other half has to be treated manually and found by the encoders' own knowledge of typical Danish food.

The description in DDO of food from animals can be divided in three types.

Description type 1:

(19 animals, 2 animal body parts). This type is the only directly reusable type when it comes to the semi-automatic construction of DanNet. The animal or bodypart lemma has two senses (two main senses or a main sense and a subsense), namely an animal sense and a food sense. They can be expressed in different ways: Sense 1: animal/bodypart sense 2: meat from this animal/bodypart (used as food) or bodypart/animal used as food. What is important in the case of DanNet, is that the senses have genus expressions corresponding to the two senses (namely animal or meat). This is also the case for those food definitions which are formulated as 'this animal/bird/fish used as food' and where 'animal/bird/fish' in some similar cases have been preferred as the genus expression (see type 2 below). The 'animal' and 'food' genus expressions make it easy to extract both an animal hierarchy and a food hierarchy directly from the genus proximum. Examples of this description type are: *kalv* (calf/veal) *lam* (lamb), *fisk* (fish), *musling*, and *snegl* (snail).

Description type 2:

(9 animals). Here, like in 1, we find two senses, an animal and a food sense, but the genus expression of the food sense is 'animal' due to the definition formulated as 'animal used as food'. This means that the food hierarchy cannot be established semi-automatically from DDO. Examples are such as *kylling* (chicken), *and* (duck), *gås*

frugt (fx bælg)
 plantedel der indeholder og beskytter plantens frø

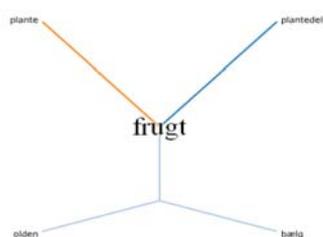


Fig. 2: *frugt* (fruit) in the botanical sense

We have not found any examples of such “false friends” among the zoological senses.

Specific to animal-derived food is also the distinction between the “grinding” sense of meat and the animal as a whole which is much more intuitive and clear than the fact that, i.e. fruit must be placed in two different hierarchies. This is set out explicitly by the fact that the meat senses of *okse* (beef), *lam* (lamb) and *svin* (pork) all have additional food terms in Danish: *oksekød*, *lammekød*, *svinekød* (lit. cattle meat, lamb meat and pig meat, respectively). In accordance with the general wordnet framework principles, these are organized pairwise in the same synsets in DanNet: {*okse*, *oksekød*}, {*lam*, *lammekød*}, and {*svin*, *flæsk*, *svinekød*}.

Particular for animals as food, is also the immense number of concepts that express which part of the animal the meat is cut from, the kind of animal it is cut from, as well as the cooking tradition in which a certain cut has emerged. For instance, Danish food culture is influenced heavily by several other cooking traditions, in particular the French, Italian and American traditions, and terms from these different cultures are all more or less integrated parts of the Danish food vocabulary. In DanNet, these are simply merged under one superconcept: *udskæring* (meat cut) independently of their mutual paronymic relations.

See Figure 3 for a small excerpt of the different meat cuts in the Danish taxonomy relating to these different cultures; concepts include such as *cordon bleu* and *cuvette* from French, *t-bone steak* with American inspiration, as well as *bov* (shoulder part of an animal) and *nakkefilet* (fillet from the back of an animal, typically a pig) which are more traditional Danish cuts.

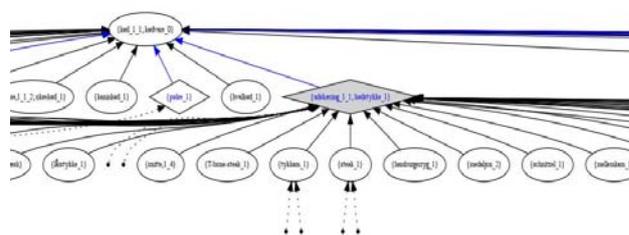


Fig. 3: Excerpt of the long list of meat cuts as subtypes to meat

To sum up on the DanNet encoding, instead of providing subdividing or paronymic dimensions as such, we strive towards keeping the natural and functional taxonomies apart; interrelating them, however, by means of either regular polysemy (i.e. animal vs meat) or multiple inheritance (i.e. tomato in the ‘part of plant’ sense with a ‘fruit’ hypernym AND a vegetable hypernym).

In contrast, terms which are only relevant for one of the taxonomies (like in the cases of *grøntsag* (vegetable) and *vildt* (game) are obviously placed in one hierarchy only. Likewise, the few previously mentioned cases of not recurrent polysemy (i.e. *nød* (nut)) are placed in each hierarchy and not interrelated³.

5. Conclusions

Foods taxonomies are typically folk taxonomies that have emerged spontaneously in different cultures depending on the goods available and on the particular cooking traditions of a particular region. They are obviously highly inspired by and related to specialist views of botanical and zoological taxonomies in the sense that terms are taken over from them. However, as we have seen, such parallel terms are mostly organized differently in the two kinds of taxonomies. This fact stresses the necessity

(i) to base a wordnet for a particular language on monolingual grounds since the cultural background influences the taxonomical structure of folk taxonomies quite radically, and

(ii) to develop a framework that enables the wordnet encoder to distinguish clearly between the natural taxonomies and the functional taxonomies of the network.

In this paper we have shown that a traditional dictionary like DDO survives well with several inconsistencies between different taxonomies of the vocabulary and that a restructuring is therefore necessary in order to compile a

³ It should be noted that the revision of food terms in DanNet according to these principles is still in progress.

consistent wordnet resource on its basis. To this end, we have outlined a methodology of how to cope with these clashes of taxonomy based on Cruse's distinctions between different hyponymies, as well as on the acknowledgement of regular polysemy as a way of connecting related taxonomies.

6. References

- Apresjan, J. (1973). Regular polysemy. In: *Linguistics* 142, pp 5-32.
- Cruse, D.A. (1991). *Lexical Semantics*. Cambridge University Press.
- Cruse, D.A. Hyponymy and Its Varieties (2002). In: R. Green, C.A. Bean, & S. H. Myaeng (eds.) *The Semantics of Relationships: An Interdisciplinary Perspective, Information Science and Knowledge Management*. Springer Verlag.
- Derwojedowa, M. M. Piasecki, S. Szpakowicz, M. Zawislawska & B. Broda (2008). Words, Concepts and Relations in the Construction of the Polish WordNet. In: *Global WordNet Conference 2008* 162–177. Szeged, Hungary.
- DDO = Hjorth, E., Kristensen, K. et al. (eds.). (2003-2005). *Den Danske Ordbog 1-6 ('The Danish Dictionary 1-6')*. Gyldendal & Society for Danish Language and Literature. www.ordnet.dk/ddo.
- Fellbaum, C. (ed) (1998). *WordNet – An Electronic Lexical Database* pp. 23–47, The MIT Press, Cambridge, Massachusetts, London, England.
- Fernández-Montraveta, A, G. Vázquez & C. Fellbaum (2008). The Spanish Version of WordNet 3.0. In: *Text resources and Lexical Knowledge* pp. 175–182. *Text, Translation, Computational Processing*. Mouton de Gruyter, Berlin & New York.
- Fjeld, Ruth E Vatvedt; Nygaard, Lars. (2009). NorNet - a monolingual wordnet of modern Norwegian". In *Proceedings of Nodalida 2009, Odense, Denmark*. NEALT proceedings series, Vol. 4. Tartu University Library, Estonia.
- Huang, C., Hsiao, P., Su, I., Ke, X. (2008). Paronymy: Enriching Ontological Knowledge in WordNets. *Proceedings of the Fourth Global WordNet Conference* 221–228, Szeged, Hungary.
- Madsen, B.N. & H.E. Thomsen. (2009). CAOS— A tool for the Construction of Terminological Ontologies. *Nealt Proceedings Series VOL. 4. 17th Nordic Conference of Computational Linguistics* 279-283, Odense
- Miháltz, M., C. Hatvani, J. Kuti, G. Szarvas, J. Csirik, G. Prószéky, T. Váradi. 2008. Methods and results of the Hungarian WordNet Project. In: *Global WordNet Conference 2008*, 387–405. Szeged, Hungary.
- Pedersen, B.S, S. Nimb, J. Asmussen, N. Sørensen, L. Trap-Jensen, H. Lorentzen (2009). DanNet – the challenge of compiling a WordNet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation* vol. 43: 269-299.
- Rodríguez, H. D. Farwell, J. Farreres, M. Bertran, M. Alkhalifa, M. A. Martí, W. Black, S. Elkateb, J. Kirk, A. Pease, P. Vossen, & C. Fellbaum (2008). Arabic WordNet: Current State and Future Extension. In: *Global WordNet Conference 2008*, 387–405. Szeged, Hungary.
- Wierzbicka, A. (1996). *Semantics: Primes and Universals*. Oxford University Press.