Merging DanNet with Princeton Wordnet

Bolette S. Pedersen[1], Sanni Nimb[2], Ida R. Olsen[3], Sussi Olsen[4]
University of Copenhagen [1,3,4] & The Danish Society for Language and Literature[2]
Njalsgade 136, DK-2300 Copenhagen S[1,3,4,] Christians Brygge 1, DK-1219[2]
bspedersen@hum.ku.dk, sn@dsl.dk, jms862@hum.ku.dk, saolsen@hum.ku.dk

Abstract

In this paper we describe the merge of the Danish wordnet, DanNet, with Princeton Wordnet applying a two-step approach. We first link from the English Princeton core to Danish (5,000 base concepts) and then proceed to linking the rest of the Danish vocabulary to English, thus going from Danish to English. Since the Danish wordnet is built bottom-up from Danish lexica and corpora, all taxonomies are monolingually based and thus not necessarily directly compatible with the coverage and structure of the Princeton WordNet. This fact proves to pose some challenges to the linking procedure since a considerable number of the links cannot be realised via the preferred cross-language **synonym** link which implies a more or less precise correlation between the two concepts. Instead, a subpart of the links are realised through near synonym or hyponymy links to compensate for the fact that no precise translation can be found in the target resource. The tool WordnetLoom is currently used for manual linking but procedures for a more automatic procedure in future is discussed. We conclude that the two resources actually differ from each other quite more than expected, both vocabulary- and structure-wise.

1 DanNet - a monolingually compiled wordnet

In contrast to the majority of wordnets following the Princeton standard, DanNet (Pedersen et al. 2009) is constructed using the so-called merge approach where the wordnet is built on monolingual grounds and thereafter merged with Princeton WordNet (PWN, cf. Fellbaum 1998).

DanNet is open source and currently contains 65,000 synsets available from www.wordnet.dk in owl/rdf and csv formats (Pedersen et al. 2009). It can be browsed online from www.andreord.dk or from wordties.cst.ku.dk.

The wordnet has been compiled as a collaboration between the University of Copenhagen and the Society for Danish Language and Literature and is based on Den Danske Ordbog (DDO, Hjorth et al. 2003-2005). In other words, our starting point was the corpus-based, at that time newly completed dictionary of Danish, accessible in a machine-readable version and with genus proximum information explicitly specified for each sense definition (DDO). The motivation for a monolingual approach seemed obvious since by taking this approach we were enabled to compile the wordnet in a rather efficient and semi-automatic fashion using the genus proximum of the dictionary as the driving factor. The result was a resource truly based on the Danish language and vocabulary and not biased by English.

The SIMPLE lexicons (cf. Lenci et al. 2000) and particularly the Danish version of it (Pedersen & Keson 1999, Pedersen & Paggio 2004) have also influenced the construction of DanNet in the sense that it includes qualia information[1] such as the telic (PURPOSE) and the agentive role (ORIGIN), roles which corresponded well with the content of the word definitions in DDO. Qualia roles are encoded in DanNet in terms of relations such as *used_for*, *made_by* and *concerns* as well as by means of features such as SEX and CONNOTATION. Apart from these additional features, DanNet follows wordnet standards wrt. relation types and synset structure, and all synsets are tagged with EuroWordNet Top Ontology types (Vossen et al 1999).

[1] We apply Qualia Structure and Qualia information as proposed by Pustejovsky 1995.

2 Linking procedure – manual or semi-automatic?

Not surprisingly, a major disadvantage of applying the monolingual strategy is that subsequent linking to PWN becomes really complex and cumbersome, which is also why it was not prioritized in the first phase of the Danish wordnet project. Over time, however, it has become more and more evident that a full linking of the resource is indispensable if we want to operate in all sorts of multilingual contexts and if our vision of applying language transfer where it is meaningful and does not involve too strong a bias, should be realistic. To this end, we have been much inspired by the work around the Polish wordnet, plWordNet (Maziarz et a. 2014), a resource which is compiled monolingually in a fashion comparable to that of DanNet and subsequently merged with PWN. Thus, much of the linking experiences resembled in i.e. Rudnicka et al. (2012) such as differences in taxonomies/structures have counterparts in our work even if the difficulties are not exactly the same. [2]

Driven by the METANET/METANORD initiatives (cf. www.meta-net.eu) where we wanted to validate wordnets across the Nordic countries (cf. Pedersen et al. 2013), we initiated the merge with PWN by focusing on Princeton Core wordnet (http://wordnetcode.princeton.edu/standoff-files/core-wordnet.txt) which is a subset 5,000 central concepts of English. Going from English to Danish, these concepts where linked semi-automatically to DanNet and missing concepts where established in the Danish resource. A bilingual dictionary was used as a first automatic look-up and link suggestion for the core concepts and from here on the encoder could accept or modify the proposed links applying a wizard-like routine in the encoding tool.

When embarking in 2018 the ELEXIS project (cf. elex.is, Krek et al. 2018), which is concerned with opening up linguistic and lexicographical data and language tools for European communities, we were finally prompted to start the full linking process of DanNet. This time the process is switched, going from Danish to PWN and thus taking point of departure in the Danish coverage and taxonomy. [3]

In this process, we also make use of a bilingual dictionary, but no semi-automatic linking to PWN is applied at the current stage. The reason for this is that it was not very evident which particular automatic procedure to pursue because of the many cases where no exact match can be found in PWN to a Danish synset, as also depicted in Figure 1.

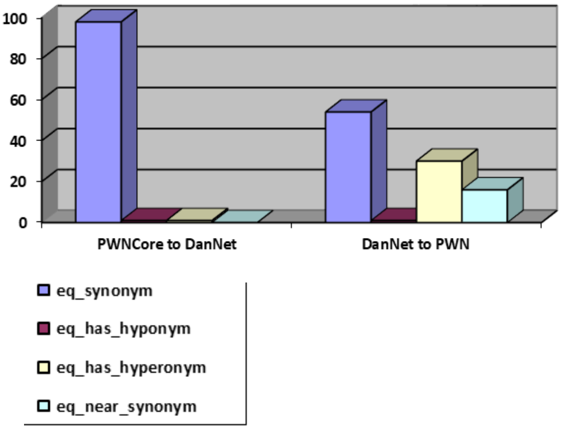


Figure 1. Percentage of different linking relations used when linking core concepts from English to Danish compared to linking general vocabulary from Danish to English.

Figure 1 illustrates how the use of linking relations differ quite radically when linking from PWNCore compared to when linking the other way around from DanNet to PWN. When going from PWNCore to DanNet, i.e. linking between core concepts in the two languages, almost all links are direct links in terms of eq synonym relations (for more details see Section 4). This means that the lexicographer has in almost all cases identified (through the semiautomatic procedure) what is considered to be an exact match between the English and the Danish resource.

The opposite proves to be the case when it comes to the linking of non-core concepts, now with the Danish resource as starting point for the linking process.[4] In the cases where no direct links are

[2] For instance, Rudnicka et al. (2012) show that since lexical units are the main building blocks in plWordNet (and not synsets as in PWN), linking to PWN is not straightforward.

[3] The linking is funded partly by ELEXIS, partly by The Carlsberg Foundation.

[4] Note however that DanNet contains less than one third of the number of senses in PWN. Nonetheless, the coverage differs quite substantially in particular when it comes to compounds, for more discussion see Section 4.

found, a rather complex cognitive procedure is initiated by i.e. looking up the Danish hypernym, finding the corresponding PWN synset, and looking for candidates among the related PWN hyponyms. Alternatively, by searching for a potential PWN hyponym to be linked to (for more details see Section 5).

To this end, we have at the current stage estimated that an automatic procedure for this process requires a rather precise cross-lingual hypernym or hyponym detection as a minimum. Nevertheless, some links can be established semi-automatically once a certain amount of relations have been established. Either vertically in cases where a Danish synset is synonym-linked to a PWN synset where it can be suggested that the hypernym of the PWN synset is also a hypernym of the Danish synset. Or horizontally, e.g. if two Danish synsets are near-synonymous, and only one is synonym-linked to PWN, then the second Danish concept can inherit that near-synonym link.

Another possibility is to apply an automatic prompt system as proposed by Kędzia et al. (2013) where the linguist/lexicographer is prompted in the process of manual mapping plWordNet on PWN. This system is based on the extended Relaxation Labelling algorithm, and suggests potential target synset candidates based on the synset positions in both wordnet structures, bilingual dictionaries and/or input from the linguist. Finally, the linguist verifies (or rejects) suggested links. It seems plausible to adjust this system to our mapping process and speed up the manual linking: it partially resembles the cognitive procedure described above, and also provides a possibility to determine the desired type of semantic relation.

At a later stage, when a more substantial part of the vocabulary has been linked, we will consider whether to follow for example Joshi et al. (2012) who generate lists of potential linking candidates with a heuristic based measure by pruning and ranking information from bilingual dictionaries. Better results are achieved with this measure when a number of links are already established. This approach could potentially be implemented when being able to utilize the high-quality established links to PWN already made by language experts. Arcan et al. (2016) use existing relations across wordnets and parallel corpora to identify contextual information for wordnet senses, and thereby expand the wordnets. Such an approach could also be adapted in our case and, again, build on the established links.

The approach of McCrae et al. (2017) for linking English-German knowledge graphs combines machine translation and cross-lingual ontology alignment. This approach, which makes use of the NAISC tool (McCrae et al. 2018), could be adapted for linking DanNet to PWN, and tested on the established links. It would require high-quality machine translation and sufficiently rich synset information, which additionally could be reinforced with contextual information as in Arcan et al. (2016).

Certainly, such automatic approaches would not achieve the precision of the manually created links, but they could be integrated as part of a semi-automatic procedure in order to speed up the process.

3 Linking complexities due to taxonomical differences

A major challenge when merging two wordnets concerns the often found discrepancies in taxonomical structure (Pedersen et al. 2013, Rudnicka 2012). Taxonomical discrepancies may have different origins, such as:

- different overall compilation approaches regarding how to organize the wordnet
- cultural differences in how to conceive a (group of) concept(s),
- idiosyncracies of the wordnet developers.

In our linking work, we encounter discrepancies of all three types. Where DanNet is compiled on the basis of a layman's dictionary of Danish, PWN is compiled without basis in any specific previous resource, but generally more true to expert knowledge in particular in relation to i.e. natural taxonomies. Consider the taxonomical complexity of the concept *plante* ('plant') in DanNet in Figure 2 compared to that of PWN in Figure 3. Even if the graphical interfaces differ, it proves quite evident that DanNet uses a layman's much simpler organization principles of plants than does PWN. Another overall discrepancy worth mentioning is different approaches taken wrt. the treatment of systematic polysemy. For instance, in DanNet all countries have a 'geographical' and a 'people' reading, a dichotomy which is not equally found in PWN and which makes a one-to-one linking procedure impossible.

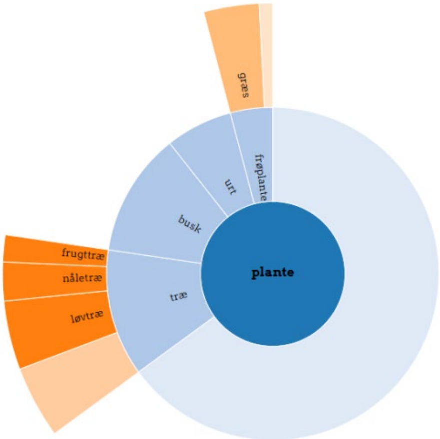
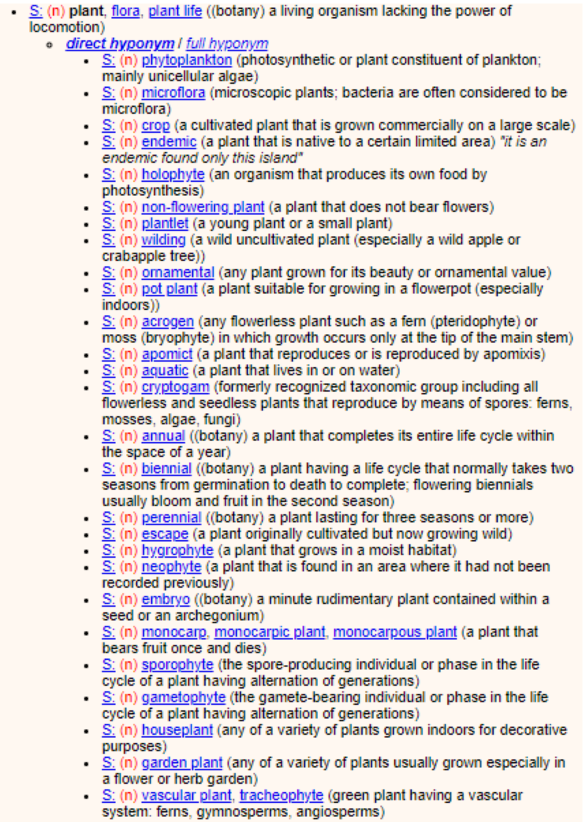


Figure 2: Taxonomical complexity of *plante* ('plant') in DanNet based on a layman's approach



- S: (n) **plant**, flora, plant life ((botany) a living organism lacking the power of locomotion)
  - *direct hyponym* / *full hyponym*
    - S: (n) phytoplankton (photosynthetic or plant constituent of plankton; mainly unicellular algae)
    - S: (n) microflora (microscopic plants; bacteria are often considered to be microflora)
    - S: (n) crop (a cultivated plant that is grown commercially on a large scale)
    - S: (n) endemic (a plant that is native to a certain limited area) *"it is an endemic found only this island"*
    - S: (n) holophyte (an organism that produces its own food by photosynthesis)
    - S: (n) non-flowering plant (a plant that does not bear flowers)
    - S: (n) plantlet (a young plant or a small plant)
    - S: (n) wilding (a wild uncultivated plant (especially a wild apple or crabapple tree))
    - S: (n) ornamental (any plant grown for its beauty or ornamental value)
    - S: (n) pot plant (a plant suitable for growing in a flowerpot (especially indoors))
    - S: (n) acrogen (any flowerless plant such as a fern (pteridophyte) or moss (bryophyte) in which growth occurs only at the tip of the main stem)
    - S: (n) apomict (a plant that reproduces or is reproduced by apomixis)
    - S: (n) aquatic (a plant that lives in or on water)
    - S: (n) cryptogam (formerly recognized taxonomic group including all flowerless and seedless plants that reproduce by means of spores: ferns, mosses, algae, fungi)
    - S: (n) annual ((botany) a plant that completes its entire life cycle within the space of a year)
    - S: (n) biennial ((botany) a plant having a life cycle that normally takes two seasons from germination to death to complete; flowering biennials usually bloom and fruit in the second season)
    - S: (n) perennial ((botany) a plant lasting for three seasons or more)
    - S: (n) escape (a plant originally cultivated but now growing wild)
    - S: (n) hygrophyte (a plant that grows in a moist habitat)
    - S: (n) neophyte (a plant that is found in an area where it had not been recorded previously)
    - S: (n) embryo ((botany) a minute rudimentary plant contained within a seed or an archegonium)
    - S: (n) monocarp, monocarpic plant, monocarpous plant (a plant that bears fruit once and dies)
    - S: (n) sporophyte (the spore-producing individual or phase in the life cycle of a plant having alternation of generations)
    - S: (n) gametophyte (the gamete-bearing individual or phase in the life cycle of a plant having alternation of generations)
    - S: (n) houseplant (any of a variety of plants grown indoors for decorative purposes)
    - S: (n) garden plant (any of a variety of plants usually grown especially in a flower or herb garden)
    - S: (n) vascular plant, tracheophyte (green plant having a vascular system: ferns, gymnosperms, angiosperms)

Figure 3. *Plant* with hyponyms in PWN

Cultural differences regarding how for instance the educational or the juridical system is organized is also clearly reflected in the taxonomical structures. Finally, pure idiosyncracies are found all over the resources, maybe even to some extent also culturally based; for instance cheese has a taxonomical division of concepts in DanNet (Figure 4) based on whether the cheese is cut or spread on the bread (typically on open sandwiches of rye bread); a division which is not made in PWN.

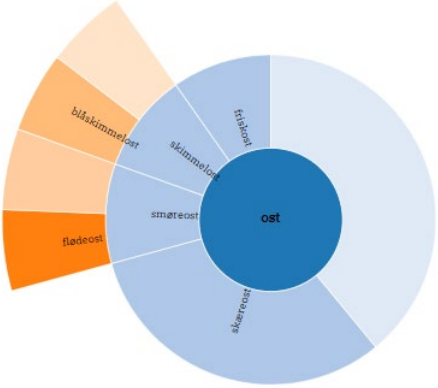


Figure 4. *ost* ('cheese') taxonomical complexity in DanNet.

4 Core concepts: Linking complexities and lexicographic characteristics

The core concepts of PWN have been selected based on two criteria: Importance of synsets measured by a) the number of relations with other synsets and b) a high position in the hierarchy.

Oflazer & Murat (2018) describes how the six Balkanet WordNets successfully used the latter criterion, a relatively high level of the English words in the PWN hierarchy, as a common starting point for the expand method, based on the assumption that language-specific information gets more important as one moves down the hierarchy. Also Green (2006) states that concepts at a basic level are more likely to be shared across classificatory systems than concepts at more general or more specific levels. In our case this is confirmed. As already described in Section 2, the linking process of the core concepts when going from PWN to DanNet results in many direct links, and equivalents were likely to be part of vocabulary covered by DanNet - only in a few cases new synsets had to be created.

The fact that DanNet is linked directly to a medium-sized corpus-based monolingual dictionary giving access to all types of lexical information now allows us to study the lexicographic characteristics of the core vocabulary in detail. We would in the case of Danish expect the core concepts to be simplex words rather than compounds and are now able to find out whether it is in fact the case. Simplex lemmas in DDO are opposite to compound lemmas characterized by often being part of the manually selected ~65,000 lemmas that constituted the vocabulary of the first printed version of the dictionary, and thereby to carry information on etymology, phonetics and compounding to a much higher degree than the ~35,000 lemmas added in the later years, after the first published edition. As seen in Table 1, the DanNet core-concept lemmas do in fact have a far higher number of all these types of information than the non-core lemmas.

Information on: DanNet Lemma	**Core**	**Non-core**
Etymology	65 %	33 %
Compounding	61%	8%
Phonetics	87%	45%
Part of DDO priority selection	99,98%	69%

Table 1. Comparison of information types across core and non-core vocabulary, percentage per lemma.

We would also expect the core concepts to be much more polysemous than the non-core concepts. The linking challenges we encountered when mapping the core synsets of PWN to DanNet are well-known to all WordNet developers (see for example Rudnicka et al. 2012, Cristea et al. 2004), typically being caused by the differences in sense distinctions and sense granularity. Often the case would be that one English synset corresponds to two or more Danish synsets, or vice versa, or even more challenging, the distinction between senses has been drawn in a slightly different way in the two resources. When looking at the number of senses of the Danish core vocabulary, it becomes obvious why the mapping was not trivial. Even though the core concept lemmas in DDO constitute only 4.6 % of the total number of lemmas in the dictionary, they cover 21.6 % of the senses in the dictionary. And while 69 % of the core lemmas are polysemous, this is only the case for 28 % of the non-core lemmas. The polysemous core lemmas have 2.65 times as many senses as the non-core polysemous lemmas. When it comes to fixed expression, the 4.6% core lemmas cover 56% of the total number in the dictionary, and they are much more likely to be part of one: 37% of them have at least one. This is only the case for 6.5% of the non-core lemmas. The core lemmas have an average of 2.76 times as many fixed expressions as the non-core lemmas, cf. Table 2. The high degree of polysemy and the high number of fixed expressions is of course a complicating factor when core concepts are linked between PWN and DanNet.

DanNet vocabulary	**Core**	**Non-core**
Lemmas ≥ 2 senses	69%	28%
Sense per polysemous lemma (incl. fixed expressions)	6.55	2.47
Lemmas with fixed expression	37%	6,5%
Fixed expressions (of lemmas with fixed expression	4.41	1.6
% of definitions (total DDO = 98,944)	21,6% 21,407	78,4% 77,537

Table 2. DanNet - core and non-core vocabulary, polysemous lemmas and fixed expressions.

When it comes to the challenges caused by different sense granularities in the two lexical resources, the Danish lexicographers who mapped the core concepts often got the impression that the sense inventory of PWN was more fine-grained than the one of DanNet/DDO. This seems to be for a good reason. When studying 20 highly polysemous Danish nouns with their English equivalents (see Table 3), we calculated PWN to have an average of 10.3 % more senses. A similar comparison of highly polysemous verbs and adjectives would probably show an even bigger difference in the number of senses.

	Number of senses	
Lemma, Danish/ English	**DDO**	**PWN**
selskab / company	10	9
kontakt /contact	9	9
kort / card, map	10	11
Plads /room, space..	13	16
slag (stroke; blow; knock)	17	12 (*stroke*)
top /top	8	11
hul /hole	14	8
plade / plate; sheet	11	15 (*plate*)
lys / light	13	15
Model	8	9
skud / shot	12	17
kurs / course	3	9
hold / hold	12	9
ansigt / face	7	13
skade / damage; harm	4	5 (*damage*)
blik / look; gaze	5	4 (*look*)

stykke / piece; bit; part	18	13 (*piece*)
stand / condition; state	9	8 (*condition*)
støtte / support	5	11
vold_1 / violence; force	6	10 (*force*)
Total number	**194**	**214 = 10,3 % more**

Table 3. Number of senses for selected polysemous Danish nouns and their English equivalents.

5 Linking complexities of non-core concepts (going from DanNet to PWN)

When it comes to the vocabulary of the non-core concepts, the linking complexities are of a different nature. One might think that the task of mapping less polysemous words would be easier, but confirming the assumptions of Oflazer & Murat, (2018) mentioned in Section 4, it seems that language-specific peculiarities tend to evoke more translation difficulties as one moves down the hierarchy. In spite of their considerable size, the two lexical resources cover quite different vocabulary and it is often difficult to find exact equivalents. Although the two wordnets seem to have more or less the same level of specificity, it is not carried out in detail for exactly the same vocabulary. Sometimes PWN is more specific wrt. to hyponyms than DanNet, and sometimes DanNet covers the highest number of specific concept lemmas, typically in the form of compounds. As an example to this, the noun *forhandling* ('negotiation') has 13 hyponyms in DanNet, all compounds, e.g. *kontraktforhandling* ('contractual negotiations'), *skilsmisseforhandling* ('divorce proceedings'). The English equivalents are not included in OED, nor in PWN. And the English equivalent to the hypernym *forhandling* ('negotiation') has no hypernyms in PWN.

Even when it comes to mapping the hyponyms of concrete core concepts which are already mapped, and where we find roughly the same number of hyponyms in the two wordnets, we might still not find many equivalents among these hyponyms. Compare for example the types of carpets in DanNet, the hyponyms of *tæppe* (*axminstertæppe*, *bedetæppe*, *kludetæppe*, *kokostæppe*, *løber*, *perser*, *rya*, *måtte*, *filttæppe*, *forligger*, *sengeforligger*, *tæppebelægning*) in Figure 5 with the types of carpets in PWN, the hyponyms of *rug* in PWN in Figure 6. Among the 14 English hyponyms, only *prayer carpet*, *runner*, *scatter rug* and *shag rug* have Danish equivalents among the 12 hyponyms of *tæppe*.

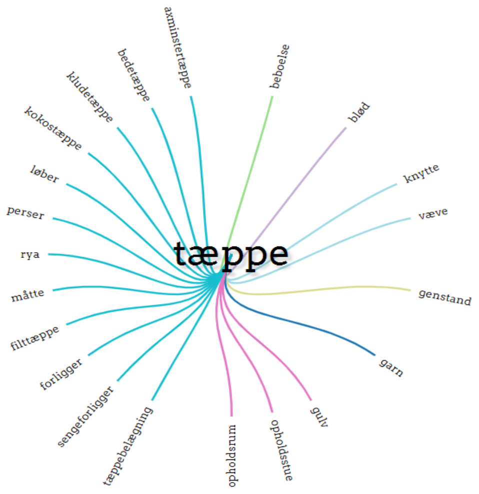
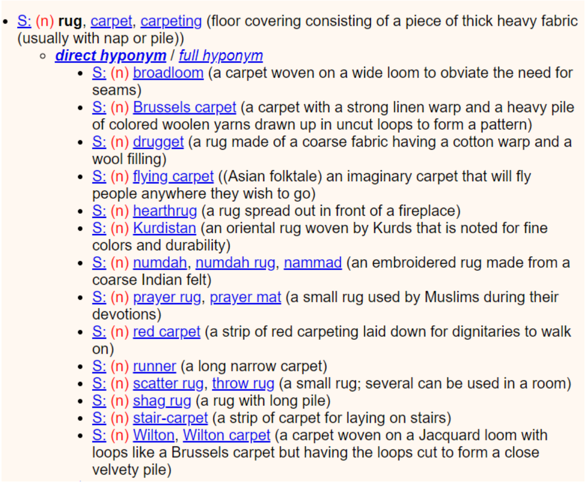


Figure 5. Hyponyms of *tæppe* ('rug') in DanNet



- S: (n) **rug**, carpet, carpeting (floor covering consisting of a piece of thick heavy fabric (usually with nap or pile))
  - ***direct hyponym*** / *full hyponym*
    - S: (n) broadloom (a carpet woven on a wide loom to obviate the need for seams)
    - S: (n) Brussels carpet (a carpet with a strong linen warp and a heavy pile of colored woolen yarns drawn up in uncut loops to form a pattern)
    - S: (n) drugget (a rug made of a coarse fabric having a cotton warp and a wool filling)
    - S: (n) flying carpet ((Asian folktale) an imaginary carpet that will fly people anywhere they wish to go)
    - S: (n) hearthrug (a rug spread out in front of a fireplace)
    - S: (n) Kurdistan (an oriental rug woven by Kurds that is noted for fine colors and durability)
    - S: (n) numdah, numdah rug, nammad (an embroidered rug made from a coarse Indian felt)
    - S: (n) prayer rug, prayer mat (a small rug used by Muslims during their devotions)
    - S: (n) red carpet (a strip of red carpeting laid down for dignitaries to walk on)
    - S: (n) runner (a long narrow carpet)
    - S: (n) scatter rug, throw rug (a small rug; several can be used in a room)
    - S: (n) shag rug (a rug with long pile)
    - S: (n) stair-carpet (a strip of carpet for laying on stairs)
    - S: (n) Wilton, Wilton carpet (a carpet woven on a Jacquard loom with loops like a Brussels carpet but having the loops cut to form a close velvety pile)

Figure 6. Hyponyms of *rug* in PWN

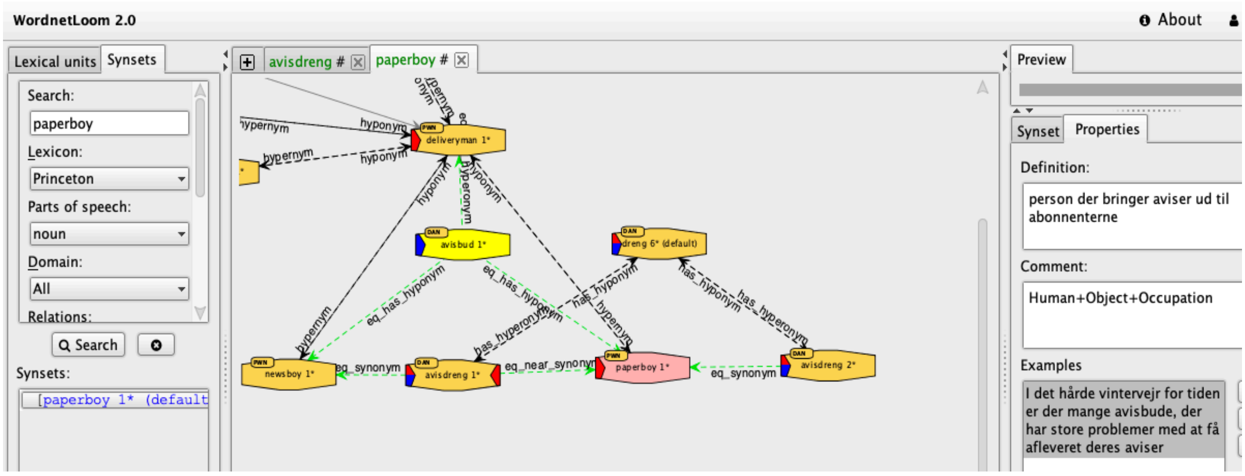
Also culture-specific differences as discussed in section 3 result in many lexical gaps. This is a problem that wordnet developers encounter even when applying the expand model in the first place. In BalkaNet for example, once a core wordnet was developed by translating from PWN, the 6 language partners had to add a number of language-specific synsets which were afterwards linked to PWN via hypernymy relations (Oflazer & Murat 2018, p. 328). In our case such synsets are already included in DanNet and have Danish hypernyms, and they are now supplied with an English hypernym as well, also in cases where an English translation equivalent does in fact exist but is not (yet) part of PWN. One example is the vocabulary of handball, a common sport in Denmark, however less important in the Anglo Saxon community and therefore not (yet) included in PWN.

Finally it should be mentioned that some linking complexities are caused by differences in word formation in Danish and English. Where noun-noun compounding is indeed very productive in Danish, English in many cases construct similar content by using an attributive and a noun. For example, compounds with *andels-* (co-op, cooperative) e.g. *andelssamfund* and *andelsbutik* translate into English by using an attributive and a noun as in 'cooperative society', 'cooperative store'. There seems to be a tendency that such terms are not lexicalized in English to the same degree and thus not present in PWN.

6 The linking tool

For the linking from DanNet to PWN (which is currently ongoing) we apply the wordnet editing system WordnetLoom 2.0 (Naskręt et al. 2017). WordnetLoom is a graph-based system where several users can access and edit the nodes (lexical units) edges (semantic relations), and synsets as well as view glosses and usage examples. The complex ontological types of the synsets (following The EuroWordNet top-ontology (Vossen 1999)) are also visible in the accustomed version suitable for browsing DanNet, developed by Tomasz Naskręt[5] and adapted by Mitchell J. Seaton[6].

An advantage of the system is that users can view and directly edit the relations in the interface, avoiding problems on manual editing of a wordnet representation file. As seen at the top of Figure 7, multiple bars of slices of the wordnet graph can be open at the same time, and are found by a given search query to the left. The results can, in the DanNet adjusted version, be filtered by part-of-speech, synsets, supersenses, lexical units, and lexicons. The presentation of results includes relations and nodes from both DanNet and PWN.



Search: paperboy

Lexicon: Princeton

Parts of speech: noun

Domain: All

Definition: person der bringer aviser ud til abonnenterne

Comment: Human+Object+Occupation

Examples

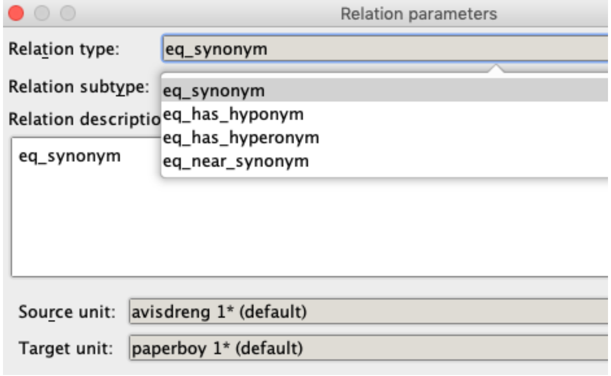
I det hårde vintervejr for tiden er der mange avisbude, der har store problemer med at få afleveret deres aviser

Figure 7: Linking synsets in WordnetLoom

Figure 7 shows an example where *avisbud 1* ('paper deliveryman') is placed between 'deliveryman 1' as a hypernym, and 'newsboy 1' as a hyponym. *avisdreng 1* is synonymous with 'newsboy 1', which is nearly the same as 'paperboy 1'. Every relation can be established, edited or deleted. The synonym, near-synonym, hypernym and hyponym relations (see the green lines) are prioritized (in that order) when linking. The relation is chosen from a drop-down menu as seen in Figure 8.

[5] G4.19 Research Group, Department of Computational Intelligence
Wrocław University of Science and Technology, Wrocław, Poland

[6] Centre for Language Technology, Department of Nordic Studies and Linguistics, Copenhagen University



Relation parameters

Relation type: eq_synonym

Relation subtype:

Relation descriptio

eq_synonym

Source unit: avisdreng 1* (default)

Target unit: paperboy 1* (default)

Figure 8: WordnetLoom drop-down menu of relation types.

7 Concluding remarks

The merging of DanNet with PWN is still ongoing and proves both cumbersome and complex as we have exemplified in the previous sections. To speed up the process, we hope to be able to introduce more semi-automatic procedures at a later stage when a substantial number of links have already been established, even if it has become evident that manual inspection and correction will always be a considerable part of the job. Within the ELEXIS project the NAISC tool (McCrae 2018) will soon be available and we hope to examine to which degree a semi-automatic linking with this tool involving interaction between lexicographers and developers can be useful.

It has generally been a surprise to us to acknowledge to which extent the two resources actually differ, both vocabulary- and structure-wise. A fact which has made us realize that a merge of the resources will really only be approximate. Nevertheless, it is our conviction that even such an approximate merge will be useful for several future NLP tasks where Danish is involved. Further, in line with the goals of the ELEXIS project, we hope that it will help interconnect existing resources in the lexicographical milieus in Europe. As such, the merge will provide the interlingual access to a substantial part of the lexical resources available for Danish.

References

Arcan, M., McCrae, J.P., & Buitelaar, P. (2016). Expanding wordnets to new languages with multilingual sense disambiguation. In *Proceedings of {COLING} 2016, the 26th International Conference on Computational Linguistics: Technical Papers, p. 97,* Osaka.

Cristea, D.; Mihaila, C., Forascu, C., Trandabat, D., Husarciuc, M., Haja, M., Postostolache, O. (2004): Mapping Princeton WordNet Synsets onto Romanian Wordnet Synsets. In *Romanian Journal of Information Science and Technology, Vol. 7*, Numbers 1–2, p. 125–145.

Fellbaum, C. (ed.) (1998). *WordNet: An Electronic Lexical Database*. MIT press.

Green, R. J. (2006). *Vocabulary alignment via basic level concepts*. OCLC/ ALISE research grant report published electronically by OCLC Research. http://www.oclc.org/research/grants/reports/green/rg2005.pdf.

Hjorth, E. & Kristensen, K. red. (2003-2005). *Den Danske Ordbog, bind 1-6*, DSL / Gyldendal, Online: ordnet.dk/ddo

Joshi, S., Chatterjee, A., Karra, A. K., and Bhattacharyya, P. U. (2012a). Eating your own cooking: automatically linking wordnet synsets of two languages. In *Proceedings of COLING 2012: Demonstration Papers*, p.239—246, Mumbai.

Kędzia P., Piasecki M., Rudnicka E., Przybycień K. (2013). Automatic Prompt System in the Process of Mapping plWordNet on Princeton WordNet. *Cognitive Studies* 13: 123–141.

Krek, S., Kosem, I., McCrae, J., Navigli, R., Pedersen, B. S., Tiberius, C., Wissik, T. (2018). European Lexicographic Infrastructure (ELEXIS). In *Proceedings of the XVIII EURALEX International Congress, Lexicography in Global Contexts*, pp 881-892.

Lenci, A., Bel, N., Busa, F., Calzolari, N., Gola, E., Monachini, M., et al. (2000). SIMPLE—A general framework for the development of multilingual lexicons. *International Journal of Lexicography,* 13(4), 249–263.

Maziarz, M., Piasecki, M., Rudnicka, E., Szpakowicz, S. (2014). plWordNet as the Cornerstone of a Toolkit of Lexico-semantic Resources. In *Proceedings of the Seventh Global Wordnet Conference*, 2014.

McCrae, J. P. & Arčan, M. & Buitelaar, P. (2017). Linking Knowledge Graphs across Languages with Semantic Similarity and Machine Translation. *The First Workshop on Multi-Language Processing in a Globalising World* (MLP 2017).

McCrae, J. P. & Buitelaar, P. (2018). Linking Datasets Using Semantic Textual Similarity. *Cybernetics and Information Technologies. 18*, p.109-123. 10.2478/cait-2018-0010.

Naskręt, T., Dziob, A., Piasecki, M., Saedi, C., & Branco, A. (2018). WordnetLoom – a Multilingual Wordnet Editing System Focused on Graph-based Presentation. In *Proceedings of the 9th Global WordNet Conference (GWC2018)*, Singapore.

Oflazer, K., Saraçlar, M. (eds.) (2018). *Turkish Natural Language Processing.* Springer International Publishing AG, Switzerland.

Pedersen; B.S. & Keson, B. (1999). 'SIMPLE - Semantic Information for Multifunctional Plurilingual Lexica: Some Danish Examples on Concrete Nouns'. In: *SIGLEX99: Standardizing Lexical Resources.* Association of Computational Linguistics, ACL99 Workshop, Maryland.

Pedersen, B. S., Paggio, P. (2004). The Danish SIMPLE Lexicon and its Application in Content-based Querying, in *Nordic Journal of Linguistics Vol 27*:1 p.97-127.

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. *Language Resources & Evaluation* 43:269–299.

Pedersen, B. S., Lindén, K., Vider, K., Forsberg, M., Kahusk, N., Niemi, J., Nygaard, L., Seaton, M., Orav, H., Borin, L., Voionmaa, K., Nisbeth, N. and Rögnvaldsson, E. (2013). Nordic and Baltic wordnets aligned and compared through “WordTies”. In *Proceedings from the 19th Nordic Conference on Computational Linguistics (NODALIDA).* Linköping Electronic Conference.

Pustejovsky, J. (1995). *The Generative Lexicon*. MIT Press, Cambridge, MA

Rudnicka, E., Maziarz, M., Piasecki, M., Szpakowicz, S. (2012). A strategy of Mapping Polish WordNet onto Princeton WordNet. In *Proceedings of COLING 2012*, Posters, pp. 1039–1048, Mumbai.

Vossen, P (ed). (1999). *EuroWordNet, A Multilingual Database with Lexical Semantic Networks.* Kluwer Academic Publishers, The Netherlands.

.