Nordic and Baltic wordnets aligned and compared through "WordTies"

Bolette S. Pedersen[1], Lars Borin[2], Markus Forsberg[2], Neeme Kahusk [3]
Krister Lindén[4], Jyrki Niemi[4] Niklas Nisbeth[1], Lars Nygaard[5]
Heili Orav[3], Eirikur Rögnvaldsson[6], Mitchel Seaton[1,] Kadri Vider[3], Kaarlo Voionmaa[2]

(1) University of Copenhagen, Njalsgade 140 2300 Copenhagen S
(2) University of Gothenburg, Box 200, 405 30 Gothenburg, Sweden (3)
University of Tartu, Ülikooli 18, 50090 Tartu, ESTONIA (4) University of Helsinki, P.O. Box 33
(Yliopistonkatu 4) (5) Kaldera Language Technology, Oslo, Norway

bspedersen@hum.ku.dk,lars.borin@svenska.gu.se,markus.forsberg@gu.se,
neeme.kahusk@ut.ee,krister.linden@helsinki.fi, Jyrki.Niemi@helsinki.fi,
niklas@nisbeth.dk, larsnyga@gmail.com, heili.orav@ut.ee, eirikur@hi.is,
seaton@hum.ku.dk, kadri.vider@ut.ee

ABSTRACT

During the last few years, extensive wordnets have been built locally for the Nordic and Baltic languages applying very different compilation strategies. The aim of the present investigation is to consolidate and examine these wordnets through an alignment via Princeton Core WordNet and thereby compare them along the measures of taxonomical structure, synonym structure, and assigned relations to approximate to a best practice. A common web interface and visualizer "WordTies" is developed to facilitate this purpose. Four bilingual wordnets are automatically processed and evaluated exposing interesting differences between the wordnets. Even if the alignments are judged to be of a good quality, the precision of the translations vary due to considerable differences in hyponymy depth and interpretation of the synset. All seven monolingual and four bilingual wordnets as well as WordTies have been made available via META-SHARE through the META-NORD project.

KEYWORDS: wordnets, multilingual links, wordnet web interface, Nordic and Baltic languages, META-NORD.

1 Wordnets as a multilingual action in the Nordic and Baltic countries

Wordnets (cf. Fellbaum 1998, Vossen 1998) have emerged as one of the basic standard lexical resources in the language technology (LT) field. They encode fundamental semantic relations among words, relations that further in many cases have counterparts in relations among concepts in formal ontologies. According to the BLARK (Basic Language Resource Kit) scheme, wordnets along with treebanks, are central resources when building language enabled applications. The semantic proximity metrics among words and concepts defined by a wordnet are considered useful in applications such as information access systems and authoring tools because in addition to identical words, the occurrence of words with similar (more general or more specific) meanings contribute to measuring of the similarity of content or context or recognizing the meaning. Based on this impact, there is a crucial need for continuously comparing and improving such lexical semantic resources in order to approximate to a best practice. This is particularly relevant if we foresee an integration of them in cross-lingual LT.

For the Nordic and Baltic languages, an extensive development of wordnets has taken place during the last five years, excluding here the Estonian wordnet, which has existed for more than a decade as part of the EuroWordNet project. During the META-NORD project (2011–2013) these wordnets have been further consolidated via extensions, validations and documentation. Thus, the relevance of cross-lingual alignment and comparison of wordnets in this region has emerged only recently. Wordnets or wordnet-like resources of a considerable size exist now for the Finnish, Danish, Estonian, Swedish, Icelandic and the Norwegian languages. They have not, like several previous wordnet projects such as EuroWordNet, BalkaNet, Asian Wordnet, and IndoWordNet (cf. Section 2) been built as part of collaborative projects, but have rather emerged locally through national projects and initiatives. The META-NORD project thus poses a unique opportunity for actually coordinating a Nordic-Baltic action on wordnets and investigating the results of the very different compilation strategies that have been applied.

The aim of this investigation can be expressed as threefold:

- To facilitate browsing, alignment and comparison of the Nordic and Baltic wordnets through the development of an intuitive and easy-to-extend web interface.
- To estimate the perspective of alignment via Princeton Core WordNet by generating four bilingual wordnets and evaluating them.
- Via this alignment to perform a comparison of the involved wordnets along the measures of taxonomical structure, synset structure, and relational structure.

The web interface, WordTies, is documented in Section 4. Four pilot bilingual wordnets have been produced semi-automatically via established links to Princeton Core Wordnet: Danish-Swedish, Danish-Finnish, Estonian-Finnish, Finnish-Swedish. An evaluation of these linked resources is included in Section 5 and a further comparison of a selected set of the wordnets is given in Section 6. All seven monolingual and four bilingual wordnets as well as WordTies have been made available via META-SHARE: www.meta-share.org under a variety of open source licenses.

2 Related work

Broadly speaking, wordnets can be compiled applying two different approaches: the merge vs. the expand method (cf. Rigau & Agirre 2002). By an expand method is meant that translations are made from Princeton WordNet and further customised to the target language, by a merge approach is meant that the wordnet is built monlingually and then (eventually) merged with Princeton WordNet.

The EuroWordNet project, which was concerned with the compilation of wordnets for a series of European languages (Vossen 1998), launched the idea of compiling and expanding wordnets via a so-called Inter Lingual Index (ILI) constituted by Princeton WordNet 1.5, cf. Peters et al. 1998. A successor of EuroWordNet was the BalkaNet project, where several wordnets were built in the Balkan area and aligned simultaneously. These projects all provide valuable reference points for a best practice within the expand approach, for example, BalkaNet uses a validation system based on word sense disambiguation for pinpointing wrong interlingual alignments, incomplete or missing synsets in one or another of the wordnets (Tufis, Ion & Ide 2004). Other works included mapping algorithms for aligning, tuning and validating wordnets as presented in Daudé, Padró & Rigau 1999, & Daudé, Padró & Rigau 2003 and several others. More recent collaborative wordnet projects include MultiWordNet (http://multiwordnet.itc.it) which relates Italian and Princeton wordnets, Asian WordNet which also applies the expand method for several Asian languages through a common management interface (Robkop et al. 2010), and IndoWordNet which include a series of Indian languages (Bhattacharyya 2010). Last but not least should be mentioned a recent initiative, Open Multilingual WordNet http://casta-net.jp/~kuribayashi/multi/ which aligns wordnets available through the Global WordNet Association's WordNet Grid (http://www.globalwordnet.org/gwa/gwa_grid.html).

In contrast, several recent European wordnets that have typically been compiled on a more local basis apply the merge technique (cf. Derwojedowa 2008, Borin & Forsberg 2010, Pedersen et al. 2009) applying monolingual language resources such as existing dictionaries and corpora as the initial source.

There are obvious risks related to both approaches. An expand approach based on Princeton WordNet runs the high risk of being biased towards the conceptual structure of the English language. However, with thorough customizations to the target language these risks can be reduced. A merge approach may reflect the target language better since it is based on more linguistic grounds (corpora and existing lexica) for that particular language. On the other hand, such wordnets typically differ so much from Princeton WordNet in structure that a merge becomes indeed very hard and extremely complex. These differences originate partly from different language cultures, partly from different levels of specialization depending on the source material used. For instance, a typical feature of wordnets based on monolingual lexica is that they adopt a perspective which is more geared towards the layman and therefore typically not so deep in taxonomical structure (cf. Pedersen et al. 2010).

3 Status of wordnets in the Nordic and Baltic countries

3.1 About META-NORD

During the last decade, linguistic resources have grown rapidly for all EU languages, including lesser-resourced languages such as the Nordic and Baltic ones. However they have typically been located in different places, have developed in different standards and in many cases were not well documented. The META-NORD project has aimed to establish an open linguistic infrastructure in the Baltic and Nordic countries to serve the needs of the industry and research communities. The project, which was completed in January 2013, has focused on 8 European languages – Danish, Estonian, Finnish, Icelandic, Latvian, Lithuanian, Norwegian and Swedish – each with less than 10 million speakers. The project has provided descriptions of the national landscapes in these countries via the META-NET White Paper Series "Europe's Languages in the Digital Age" and has assembled, linked across languages, and made widely available close to 500 language resources and tools of different types via the common network META-SHARE http://www.meta-share.org/. META-SHARE is a network of repositories of language data, tools and related web services documented with metadata, aggregated in central inventories allowing for uniform search and access to resources. The horizontal action on wordnets constitutes one of several cross-language initiatives in the project. In the following a brief status of each of the involved wordnets is given.

3.2 Estonian wordnet

The Estonian wordnet was built as part of the EuroWordNet project and thus used the expand method as a starting point. Base concepts from English were translated into Estonian as a first basis for a monolingual extension. The extensions have been compiled manually from Estonian monolingual dictionaries and other monolingual resources. In this sense, EstWN applies a hybrid method including both expand and monolingual techniques. EstWN includes nouns, verbs, adjectives and adverbs; as well as a set of multiword units, cf Kahusk et al. 2012. The database currently (Jan 2013) contains approx. 59 000 concepts are interlinked by175,000 relations and work is still in progress. The database is available under a CC-NY-NC license. The database can be accessed partly via WordTies, partly at http://www.cl.ut.ee/teksaurus and www.keeleveeb.ee.

3.3 Finnish wordnet

FinnWordNet is compiled using the expand method and supplemented with monolingual localisations (see http://www.ling.helsinki.fi/cgi-bin/fiwn/search). FinnWordNet contains nouns, verbs, adjectives and adverbs grouped by meaning into synonym sets representing concepts. Version 1.0 of FinnWordNet was created by translating the word senses in the Princeton WordNet 3.0 (Lindén & Carlson 2010). To ensure quality, the word senses were translated by professional translators. This approach allowed a very rapid and cost-efficient creation of an extensive Finnish wordnet directly aligned with the Princeton WordNet providing a translation relation between English and Finnish.

It is often claimed that translating a wordnet from English is somehow problematic, so to dispel such doubts several rounds of evaluations were performed, only to discover very few translation or concept problems (cf. Lindén et al. 2012). During the evaluation some missing common Finnish words and concepts were added to FinnWordNet from a large corpus of Finnish as well as from Wiktionary and Wikipedia. The resulting FinnWordNet 2.0 has 120,449 concepts containing 208,645 word senses and linked to each other with 265,690 relations. It thus surpasses Princeton WordNet in the number of concepts and word senses. FinnWordNet is licensed under the Creative Commons Attribution (CC-BY) 3.0 licence. As a derivative of the Princeton WordNet, FinnWordNet is also subject to the Princeton WordNet licence.

3.4 Danish Wordnet

DanNet has been constructed using the merge approach where the wordnet is built on monolingual grounds and thereafter merged with Princeton WordNet. It currently contains 66,308 concepts which are interlinked by 326,564 relations (see also Pedersen et al. 2009). The wordnet has been compiled as a collaboration between the University of Copenhagen and the Danish Society for Language and Literature and is based on *Den Danske Ordbog* (Hjorth & Kristensen 2003). Furthermore, the Danish version of the SIMPLE lexicons (cf. Lenci et al. 2001) has influenced the construction of DanNet in the sense that it includes also qualia information such as the telic and the agentive role (purpose and origin). Qualia roles are encoded in DanNet in terms of relations such as used_for and made_by as well as by means of features such as SEX and CONNOTATION. DanNet is licensed under the Princeton WordNet licence.

3.5 Swedish wordnet (Swesaurus)

Swesaurus (Borin & Forsberg 2010, Borin & Forsberg 2011) is a Swedish wordnet developed at Språkbanken, University of Gothenburg. It is being built by reusing lexical-semantic relations collected from a number of pre-existing, freely available lexical resources: SALDO (Borin & Forsberg 2009), SDB (Järborg 2001), Synlex (Kann & Rosell 2006), and Swedish Wiktionary. A novel feature of Swesaurus is its fuzzy synsets derived from the graded synonymy relations of Synlex. Swesaurus and several other lexical resources are available for download and inspection at http://spraakbanken.gu.se/karp. Swesaurus is an integral part of a large and diverse lexical macroresource compiled in the Swedish FrameNet++ project (Borin et al. 2010). It includes 13,724 senses and is licensed under a CC-BY license. Due to its slightly different structure, Swesaurus is currently only partly visible through WordTies.

3.6 Norwegian Wordnet

A Norwegian Wordnet (NWN) has been developed as a part of The Norwegian Language Bank (*Språkbanken*). It consists of around 50,000 synsets, for both Norwegian Nynorsk and Norwegian Bokmål and covers more than 90 per cent of the senses of open word

classes in running newspaper text. Both wordnets are available via META-SHARE (http://www.nb.no/clarin/repository/search/?q=ordnett) under the Princeton WordNet License. The compilation is based on the Danish wordnet (DanNet), and thus NWN contains the same lexical relations and much of the same semantic analysis as DanNet. The data format and licence are also identical.

Semantically, Danish and Norwegian are very closely related, and word senses are mostly equivalent (though the frequency with which the senses are used often varies). Some synsets are dropped: some are only relevant for Danish society, and do not have natural equivalent in Norwegian. These synsets are almost exclusively infrequent and "peripheral" in the DanNet (i.e. they are leaf nodes in the synset graph). A partial semantic annotation of a Norwegian corpus has been developed to ensure that the most frequent senses for Norwegian text are covered. Using this method, it has been possible to create a very extensive wordnet for a fraction of the cost for development from scratch, and without the quality problems associated with translation from for example English.

3.7 Icelandic wordnet (MerkOr)

The semantic database MerkOr, which constitutes the Icelandic wordnet, has been developed using a monolingual approach with automatic methods for the extraction of semantic information from texts. Both pattern-based and statistical methods are used, as well as a hybrid methodology.

The structure of the database is not based on hierarchies, like the Princeton WordNet, but rather on clusters of strongly related words and semantic relations often describing common sense knowledge and associations. The database contains about 110,000 words, primarily nouns, but also a number of verbs and adjectives. About 2.93 million relations between these words are listed in the database which also contains 305 semantic clusters – lists of words that belong to the same semantic field. The database is distributed under the GNU Lesser General Public License and can be queried online at http://merkor.skerpa.com. This wordnet is not yet made available via WordTies.

4 WordTies: A common web interface for viewing aligned wordnets

WordTies (wordties.cst.dk) is a web interface developed to visualize monolingual wordnets as well as their alignments with the other wordnets, cf. Figure 1. In this browser the user can chose either of the (currently four) relevant wordnets as a source language and see how a concept is linked to its sister wordnets.

Figure 1: Introductory screen of WordTies

WordTies builds on a monolingual browser, AndreOrd, which was built to browse DanNet, cf. Johannsen & Pedersen (2011). In this browser, the semantic relations are made available in a more graphical fashion compared to what is found in most other wordnet browsers which tend to focus primarily on visualizing the *hyponymy* structure of the wordnet. The particular choice of graph very compactly encodes large numbers of relations – each represented by its own colour – and thus gives a good overview of the general structure of the wordnet. In order to make room for all relations in the graph – also the inherited ones –, only one representative sense is visualized per synset. However, all senses are presented below the graph. By clicking on a related synset in the graph the user can dynamically move around in the wordnet. For illustration, see Figure 2 where Danish has been chosen as the source and the Danish concept *håb* (‘hope’) has been looked up and aligned with Estonian, Swedish and Finnish wordnets.

A click on either of these links will bring the users into these particular wordnets and enable them to browse the wordnet and view the established relations as well as its taxonomical structure, as seen in Figure 3 where we see that the Finnish wordnet has a much deeper taxonomical structure (expert perspective) than the Danish (layman perspective) of the concept *tree*. In this way, the web interface eases comparison and evaluation of the wordnets.

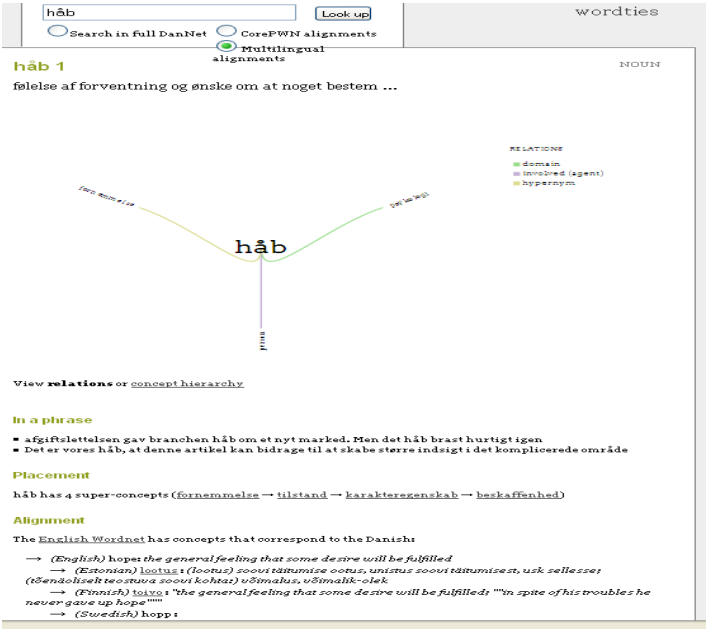

Figure 2: A Danish synset look-up (*håb* (hope)) with multilingual alignments to English, Finnish, Estonian and Swedish wordnets.

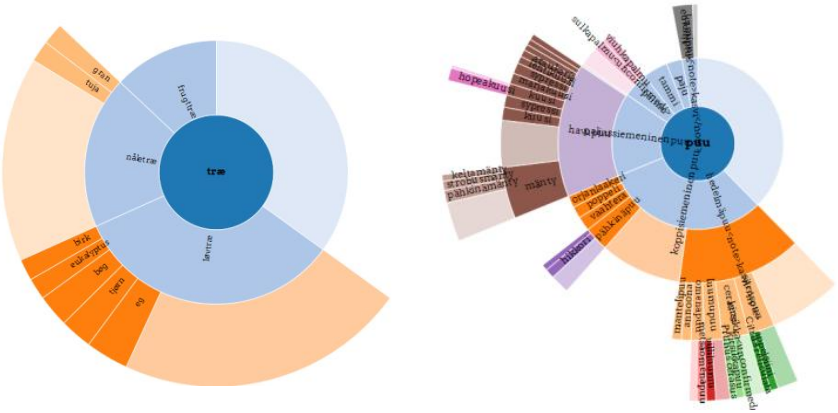

Figure 3: Graphical views of the taxonomical differences of the concept *tree* in the Danish and Finnish wordnets, respectively. DanNet includes three major (layman) subtypes of trees: deciduous trees, coniferous trees, and fruit trees, whereas the Finnish wordnet includes further subtypes based on a specialist, botanical structuring.

Major changes to the AndreOrd source-code were model changes including the addition of instance, source model classes and modification to alignment and its relations. These three model classes handle the relational structure and data used to enable the

multilingual relations (connections), facilitating a link between application instances via a wordnet's imported Princeton Core WordNet relations.

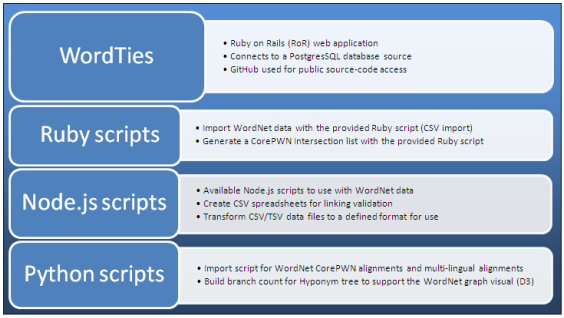

Figure 4: Overview and scripts used in WordTies.

WordTies can be dynamically extended to include more wordnets. There are two compulsory steps for import, firstly to calculate and update the hyponym count for each synset record, and secondly to import alignments to Princeton Core WordNet. Optionally, the import alignments script can be used to import multilingual alignments, alignments to other wordnet synsets via Princeton Core WordNet.

The application is able to have a customised locale, and language files. Currently, Danish and English are supported languages with the application. An index page is customisable based on the locale, and new language support can be easily added with valid translation of labels. Currently multi-locale support is not included with the application, a single locale is set for the application instance to operate in. Other customisations available, include filter values and path names (routes), and custom colour mappings for the relations graph.

5 Alignment and evaluation of bilingual wordnets

Four bilingual wordnets have been automatically processed on the basis of each wordnet's links to Princeton WordNet. In other words, English has functioned as an interlingua in a triangulation method, and a central aim has been to examine to which extent this strategy influenced the quality of the bilingual translations.

However, since the Nordic and Baltic wordnets were built locally using both the expand and merge techniques as we have seen, they differ in the extent to which they were bilingually linked before the META-NORD project was initiated. Therefore, a first task was to ensure a common linked coverage of all the involved wordnets. To this end, all wordnets were manually linked to Princeton Core WordNet containing 5,000 core synsets.[1] Princeton Core WordNet is a recent, semi-automatically compiled list of 5,000

[1] http://wordnetcode.princeton.edu/standoff-files/core-wordnet.txt

"core" word senses in WordNet corresponding approximately to the 5,000 most frequently used word senses, followed by some manual filtering and adjustment. This set of basic concepts is considered to be deduced on better statistical grounds than the previously applied "base concepts" used in the EuroWordNet and SIMPLE projects. Further, Princeton Core WordNet is characterized by being relatively coarse-grained compared to the full Princeton WordNet and thus much better suited for alignment tasks.

For the evaluation, a top 1000 set of this 5,000 synset intersection with a POS ratio of 6:2:2 for nouns, verbs, and adjectives, respectively was generated. The extract was also based on provided frequency data from Swedish and Finnish. Even if one-to-one synset alignments are by far the most frequent ones, one-to-many and many-to-one synset alignments occur as well. Valid relations to Princeton Core WordNet include eq_synonym, which are by the most frequent one, eq_has_hyponym, as well as eq_has_hyperonym, allowing thus in some cases for alignments to more or less specific synsets. All in all, four linked wordnets were processed and evaluated. Table 1 sums up the evaluation results.

	Slight mismatches	**Linking errors**
Danish-Swedish	2.3	2.4
Finnish-Swedish	2.6	1.1
Finnish-Danish	0.7	0
Estonian-Finnish	22.6	5.3

Table 1: Percentage of errors and mismatches in bilingual wordnets

As can be read, the semi-automatic alignments are judged to be of a relatively good quality even if translations are in several cases not 100% precise. An average of 2.2% errors and 7.0% slight mismatches is reported on. However, there are some clear divergences to be commented on in Table 1: the evaluation of Estonian-Finnish reports on 53 errors and 226 slight mismatches whereas no errors and only 7 mismatches are reported on for Finnish-Danish. Since the evaluations are made by different partners with the necessary bilingual language skills, part of this divergence is due to somewhat different interpretations of the concepts 'slight mismatches'. Furthermore, the different nature of the wordnets seems to have influenced the evaluations to a certain extent. Thus, some evaluators have focused on a good definition-to-definition match between two synsets, whereas others have applied a somewhat stricter criterion by evaluating the exact sense-to-sense correspondence.

WN synset	ET	FI	WN gloss	Comment
album%1:06:00::	Album_2	albumi; valokuva-albumi	a book of blank pages with pockets or envelopes; for organizing photographs or stamp collections etc	FinnWN synset member 'valokuva-albumi' is more specific (eq_has_hyponym)
apparatus%1:06:00::	aparaat_1, seadis_1	aparaatti; koneisto; laite; laitteisto; väline	equipment designed to serve a specific function	EstWN synset members are in singular, as one particular piece of equipment
gun%1:06:00::	tulirelv_1	kivääri; pistooli; pyssy; tykki	a weapon that discharges a missile at high velocity (especially from a metal tube or barrel)	FinnWN synset have more specific members

Table 2: Estonian-Finnish translations that are considered to be slight mismatches.

For instance, the Estonian-Finnish evaluator has registered differences in synonyms or differences in specificity as slight mismatches, influenced by the fact that not *all* senses in two aligned synset represent fully precise translations of each other. See Table 2 for comments on particular alignments between Estonian and Finnish.

An additional explanation to the divergences is that since the Finnish and Estonian languages are very close, the Estonian evaluator expected to see exact matches sense-to-sense in the translations between Finnish and Estonian, irrespective of the fact that the wordnets were built from more or less different starting points. The phenomenon with very close words (and mismatching translations) could presumably also have been observed between other close language pairs such as Swedish and Danish, but since these two wordnet are characterized by having less senses per synset, synonym mismatch is not observed to the same extent. The evaluator of the two "extremes" with respect to senses per synset, Danish and Finnish wordnets, reports on few mismatches, influenced by the fact that focus has here been more directed towards the definition-to-definition alignment.

Not surprisingly, wordnets that have been compiled via translations from Princeton WordNet have many senses per synsets (just as Princeton WordNet), whereas wordnets that are monolingually compiled and rather based on synonymy registrations in conventional dictionaries, have much less; (see also Section 5). As can be seen, it has proven difficult to 'neutralize' such differences initial to the evaluations.

With regards to alignment *errors*, there does not appear to be a systematic bias, some are due to false friends, others however, seem to be just random errors introduced during

the linking to Princeton Core WordNet, as in the following, where the English synset has been linked to a too specific sense of 'waste' in DanNet than what was actually indicated by the English gloss, cf. Table 3.

WN synset	DA	SV	WN gloss	Comment
waste%1:27:00::	{spildprodukt_0}	avfall..1	any materials unused and rejected as worthless or unwanted	refers to byproducts of uction; should be the more ral 'affald'.

Table 3: Example of Danish-Swedish link which is considered an error

6 Further comparison of selected wordnets

Via the evaluations presented in Section 5 and by browsing the wordnets in WordTies, further insights have been achieved wrt. the very diverse characteristics of the selected wordnets in terms of taxonomical differences, different understandings of the synset, and differences in compiling semantic relations.

First of all, we can observe some differences in average hyponym depth, number of senses per synset and average number of relations connected to a synset as shown in Table 4.

	DanNet	FinnWordNet	EstWN
Hyponym depth/SynSet	4.38	7.49	5.93
Word Senses/SynSet	1.09	1.74	1.65
Relations/SynSet	4.97	2.21	2.91

Table 4: Hypnonym depth, word sense per synset and relations per synset for Danish, Finnish and Estonian wordnets

FinnWordNet has the highest average of hyponymy depth, relating well to our intuition of this wordnet being more expert oriented at least in the fields of botany and zoology (see also Figure 3). In contrast, EstWN and DanNet which rely more on monolingual dictionaries and the genus proximums given in the definitions of these, have less depth. This fact can also be illustrated by extracting the path to the top from a botanical concept like *tree* in Danish and Finnish, respectively:

træ (tree) has 4 super-concepts (plante → organisme →fysisk genstand→entitet) (plant → organism → physical entity→entity)

puu (tree) has 9 super-concepts (puumainen kasvi → putkilokasvi → kasvi → eliö → elävä olio → kokonaisuus → esine → fyysinen entiteetti → entiteetti) (woody plant → vascular plant → plant → organism → animate thing →whole → object → physical entity → entity)

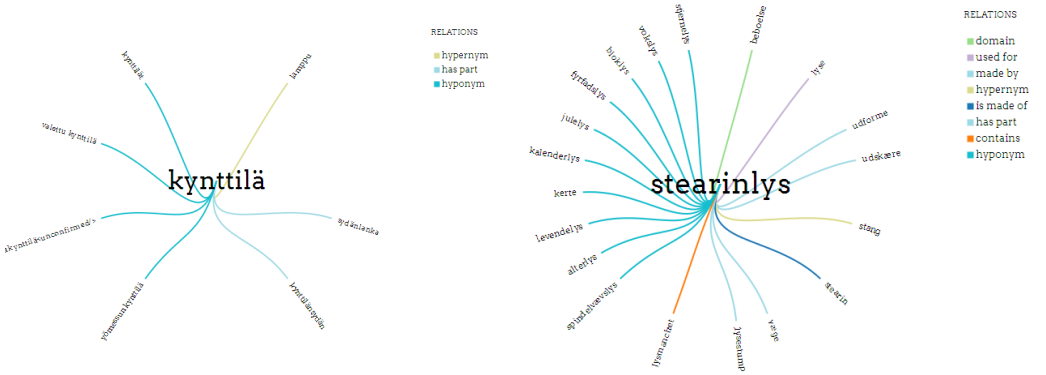

Figure 5: Differences in number of relations in Finnish and Danish, respectively, attached to the concept candle (*kynttilä, stearinlys*). For example, Danish includes relations such as used_for=light and is_made_of=stearin whereas the Finnish wordnet includes only hyponyms, parts and hyperonym.

The number and selection of relations in the wordnets also differ; some have included only Princeton relations, others include EuroWordNet relations (i.e. Estonian, Danish, Norwegian) and others again have adapted qualia-inspired (Pustejovsky 1995) relations also from the SIMPLE project (Lenci et al. 2001), such as the used_for and made_by relations in the Danish and Norwegian wordnets. This extension of the relation set to include also purpose and origin is again influenced by sense definitions in conventional dictionaries where it is typically expressed for which purpose a given artifact is made and eventually how it is made (i.e. baked, grown, cooked, produced), see Figure 5 for such differences in number of relations between Danish and Finnish wordnets.

7 Conclusion and further steps

Apart from consolidating, extending and providing richer documentation for the Nordic and Baltic wordnets, the META-NORD multilingual wordnet initiative has ensured an alignment and comparison of the most mature of these wordnets and have made them all easily accessible through META-SHARE. Central aims have been to understand better the different nature of the lexical-semantic resources in order to approximate to a best practice, to test the perspective of linking them and to make them visible in an intuitive way in a common web interface. Four core bilingual wordnets have been compiled,

made visible and evaluated with diverging, but still promising, results. The evaluations and comparisons have exposed a considerable variety of the wordnets wrt. taxonomical structure, structure of the synset (many or few senses per synset) and number of relations attached to each synset, a variety which proves to originate from the different compilation strategies used for the different Nordic and Baltic languages. As we have shown, the two compilation strategies (expand versus merge) have considerable impact on how the lexical-semantic information is represented and on the depth of the lexical hierarchies. In spite of these differences, an alignment through Princeton Core WordNet has proven feasible.

Three wordnets were not fully mature when the META-NORD project started and have therefore not yet been aligned or made visible in WordTies, namely the Icelandic and Norwegian wordnets; the plan is to include them during 2013.

References

Rigau, G. and Agirre, E. (2002). Semi-automatic Methods for WordNet Construction. Tutorial at 2002 International WordNet Conference, Mysore, India.

Bhattacharyya, P. (2010) IndoWordNet. Proceedings of LREC 2010. Valletta: ELRA.

Borin, L., Danélls, D., Forsberg, M., Kokkinakis, D. and Gronostaj, M.T. (2010). The past meets the present in Swedish FrameNet++. In Proceedings of the 14th EURALEX International Congress, pp. 269–281. Leeuwarden: EURALEX.

Borin, L. and Forsberg, M. (2009). All in the family: A comparison of SALDO and WordNet. In Proceedings of the Nodalida 2009 Workshop on WordNets and other Lexical Semantic Resources – between Lexical Semantics, Lexicography, Terminology and Formal Ontologies, pp. 7–12. Odense: NEALT.

Borin, L. and Forsberg, M. (2010). Beyond the synset: Swesaurus – a fuzzy Swedish wordnet. In Workshop on Re-thinking synonymy: Semantic sameness and similarity in languages and their description. Helsinki.

Borin, L. and Forsberg, M. (2011). Swesaurus – ett svenskt ordnät med fria tyglar. LexicoNordica vol. 18, pp. 17–39.

Borin, L., Forsberg, M. and Lönngren, L. (2008). The hunting of the BLARK – SALDO, a freely available lexical database for Swedish language technology. Joakim Nivre, Mats Dahllöf and Beáta Megyesi (eds.), Resourceful language technology. Festschrift in honor of Anna Sågvall Hein, pp. 21–32. Acta Universitatis Upsaliensis: Studia Linguistica Upsaliensia 7. Uppsala: Uppsala University.

Derwojedowa, M., Piasecki, M., Szpakowicz, S., Zawislawska, M. and Broda, B. (2008). Words, concepts and relations in the construction of the polish WordNet. In Global WordNet Conference 2008, pp. 162–177. Szeged, Hungary.

Daudé J., Padró L. and Rigau G. (2003). Validation and Tuning of Wordnet Mapping Techniques. Proceedings of the International Conference on Recent Advances on Natural Language Processing (RANLP'03). Borovets, Bulgaria.

Daudé J., Padró L. and Rigau G. (1999). Mapping Multilingual Hierarchies Using Relaxation Labeling. Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC'99). Maryland, US.

Fellbaum, C. (ed) (1998). WordNet – An Electronic Lexical Database. Cambridge, Massachusetts: The MIT Press.

Hjorth, E. and Kristensen, K. (2003). Den Danske Ordbog. Gyldendal, Denmark.

Järborg, J. (2001). Roller i Semantisk databas. Research Reports from the Department of Swedish, No. GU-ISS-01-3. University of Gothenburg: Dept. of Swedish.

Johannsen, A. and Pedersen, B.S. (2011). "Andre ord" – a wordnet browser for the Danish wordnet, DanNet. In Proceedings from 18th Nordic Conference of Computational Linguistics, NODALIDA 2011, Riga, Latvia. Nothern Association for Language Technology, Vol. 11 pp. 295–298, University of Tartu.

Kann, V. and Rosell, M. (2006). Free construction of a free Swedish dictionary of synonyms In Proceedings of the 15th NODALIDA conference, pp. 105–110. Joensuu: University of Eastern Finland.

Martola, N. (2011). FinnWordNet och det finska samhället. In: Symposium om onomasiologiske ordbøker i Norden. Schæffergården, Copenhagen.

Kahusk, N., Orav, H. and Vare, K. (2012). Cross-linking Experience of Estonian WordNet. In: Human Language Technologies – The Baltic Perspective: The Fifth International Conference on Human Language Technologies – The Baltic perspective. Tartu, Estonia, October 4-5, 2012. (Ed. Arvi, Tavast; Kadri Muischnek; Mare, Koit). IOS Press, pp. 96–102. Online access: doi:10.3233/978-1-61499-133-5-96

Lenci, A., Bel, N., Busa, F., Calzolari, N., Gola, E., Monachini, M., Ogonowski, A., Peters, I., Peters, W., Ruimy, N., Villegas, M. and Zampolli, A. (2000). SIMPLE: A general framework for the development of multilingual lexicons. International Journal of Lexicography, vol. 13, pp. 249–263

Lindén, K. and Carlson, L. (2010). FinnWordNet – WordNet på finska via översättning. LexicoNordica – Nordic Journal of Lexicography, vol. 17, pp. 119–140

Lindén, K., Niemi, J. and Hyvärinen, M. (2012) Extending and Updating the Finnish Wordnet. In Diana Santos, Krister Lindén and Wanjiku Ng'ang'a (eds.), Shall We Play the Festschrift Game? Essays on the Occasion of Lauri Carlson's 60th Birthday, pp. 67–98. Springer: Berlin, Heidelberg. ISBN 978-3-642-30773-7.

Pedersen, B.S, Nimb, S., Asmussen, J., Sørensen, N., Trap-Jensen, L. and Lorentzen, H. (2009). DanNet – the challenge of compiling a WordNet for Danish by reusing a monolingual dictionary. Language Resources and Evaluation, Computational Linguistics Series, pp. 269–299.

Pedersen, B.S., Nimb, S. and Braasch, A. (2010). Merging specialist taxonomies and folk taxonomies in wordnets. - a case study of plants, animals and foods in the Danish wordnet In: Proceedings from the Seventh International Conference on Language Resources and Evaluation, pp. 3181–3186. Malta.

Peters, W., Vossen, P., Díes-Orzas, P. and Adriaens, G. (1998). Cross-lingual Alignment of Wordnets with an Inter-Lingual-Index. In: EuroWordNet – A Multilingual Database with Lexical Semantic Networks, pp. 149–179. Kluwer Academic Publishers.

Pustejovsky, J. (1995). The Generative Lexicon. Cambridge, Massachusetts: MIT Press.

Robkop, K., Thoongsup, S., Charoenpron, T., Sornlertlamvanich, V. and Isahara, H.. (2010). WNMS: Connecting Distributed Wordnet in the Case of Asian WordNet. In: Proceedings of the 5th International Conference of the Global WordNet Association (GWC 2010), Mumbai, India.

Tufiş, D., Ion, R. and Ide, N. (2004). Word Sense Disambiguation as a Wordnets Validation Method in BalkaNet. In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), pp. 1071–1074. Lisbon: ELRA

Vossen, P. (ed.) (1998). EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Dordrecht: Kluwer Academic Publishers.