Linking and Validating Nordic and Baltic Wordnets

- A Multilingual Action in META-NORD

Bolette Sandford Pedersen
University of Copenhagen
Copenhagen, Denmark
bspedersen@hum.ku.dk

Lars Borin
University of Gothenburg
Gothenburg, Sweden
lars.borin@svenska.gu.se

Markus Forsberg
University of Gothenburg
Gothenburg, Sweden
markus.forsberg@gu.se

Krister Lindén
University of Helsinki
Helsinki, Finland
krister.linden@helsinki.fi

Heili Orav
University of Tartu
Tartu, Estonia
heili.orav@ut.ee

Eirikur Rögnvaldsson
University of Iceland
Reykjavik, Iceland
eirikur@hi.is

Abstract

This project report describes a multilingual wordnet initiative embarked in the META-NORD project and concerned with the validation and pilot linking between Nordic and Baltic wordnets. The builders of these wordnets have applied very different compilation strategies: The Danish, Icelandic and Swedish wordnets are being developed via monolingual dictionaries and corpora and subsequently linked to Princeton WordNet. In contrast, the Finnish and Norwegian wordnets are applying the expand method by translating from Princeton WordNet and the Danish wordnet, DanNet, respectively. The Estonian wordnet was built as part of the EuroWordNet project and by translating the base concepts from English as a first basis for monolingual extension. The aim of the multilingual action is to test the perspective of a *multilingual linking* of the Nordic and Baltic wordnets and via this (pilot) linking to perform a tentative comparison and validation of the wordnets along the measures of *taxonomical structure, coverage, granularity* and *completeness*. Currently, Danish, Finnish, Swedish and Estonian wordnets have been linked to Princeton Core WordNet, thereby providing a common, linked coverage of 5,000 core synsets.

1 What is META-NORD?

META-NORD is an EC project closely related to the META-NET initiative whose very general aim is to foster the technological foundations of a multilingual European information society. [1] More specifically, the META-NORD project aims to establish an open linguistic infrastructure in the Baltic and Nordic countries to serve the needs of the industry and research communities.

META-NORD runs from 2011 to the beginning of 2013 and focuses on 8 European languages - Danish, Estonian, Finnish, Icelandic, Latvian, Lithuanian, Norwegian and Swedish which each have less than 10 million speakers. The project aims at assembling, evaluating and linking across languages, and making widely available language resources of different types used by different categories of user communities in academia and industry.

Among these Nordic and Baltic resources are wordnets, which have been developed or are being developed for most of the involved languages. In this paper we investigate the different

[1] For more information on META-NORD and META-NET, see www.meta-net.eu.

nature of these wordnets, and we focus on the perspectives of a linking and evaluation of these in a multilingual context according to certain measures.

2 Multilingual Action on Wordnets

As briefly mentioned, during the last decades, wordnets have been developed for several languages in the Nordic countries including Finnish, Danish, Estonian, Icelandic and Swedish, and just recently a Norwegian wordnet is being initiated on the basis on the Danish wordnet. Of these wordnets, Estonian WordNet is the oldest one since it was built as part of the EuroWordNet project in the 1990s (see Vossen 1999). In contrast, the other three wordnets have been recently initiated; the oldest of them being the Danish wordnet which has been under development since 2005 (cf. Pedersen et al. 2009) and the latest the Norwegian wordnet which is initiated in 2011.

The builders of these wordnets have applied different compilation strategies: Where the Danish, Icelandic and Swedish wordnets are being developed via monolingual dictionaries and corpora and subsequently partially linked to Princeton WordNet; the Finnish and Estonian wordnets have applied the translation method by translating Princeton WordNet into their respective languages for later adjustment.

From the above mentioned different time perspectives and compilation, there was a need for upgrade of several of the wordnet resources to agreed standards, which was thus a preliminary task of this META-NORD action, in particular for Estonian WordNet.

A prerequisite for multilingual use of the resources is that the monolingually based resources are enhanced with regards to either synsets and/or more links to Princeton WordNet. From these links, which primarily constitute the so-called “core synsets” extracted at Princeton University,[2] pilot cross-lingual resources are derived and further adjusted and validated.

Currently, the linking of the monolingually based wordnets (Icelandic, Swedish and Danish wordnets) to the 5,000 “core synsets” has been completed to ensure common coverage between all wordnets. A tentative comparison of the resources is planned along the measures of *taxonomical structure*, *coverage*, *granularity* and *completeness* (see Section 4).

An additional aim of the multilingual task is to make the relevant wordnets accessible through a uniform web interface.

Wordnets provide semantically-based concept hierarchies for specific languages and are therefore ideal resources to use as a starting point for cross- and multilingual resources; actually they are conceptually better suited than bilingual dictionaries. With such linked resources, cross- and multilingual IR applying semantically-based query expansion becomes feasible. Another possible application for these resources is MT. The hierarchical structure of wordnets ensures that a translation can be found (going up or down in the hierarchy) even if a precise equivalent is not present between the specific languages (this feature is seen in the linking to Princeton Core where some links are constituted by means of the relation eq_has_hyperonym rather than the direct eq_has_synonym).

3 Background of Nordic and Baltic Wordnets

3.1 Estonian Wordnet, EstWN

Estonian wordnet, EstWN was initiated at the University of Tartu in 1990 as a part of the EuroWordNet project. The wordnet was developed by translating the basic concepts from English into Estonian and by building the rest of the wordnet on monolingual grounds. At present it contains more than 45,000 synsets, including nouns, verbs, adjectives and adverbs, as well as some multiword units.

EstWN has been compiled manually but there are some endeavors for automatic additions. For example, a number of words have been derived via suffixes. EstWn includes domain vocabulary from domains such as architecture, agriculture, transportation, and personality traits (Kerner et al 2010).

New updates here? Under the META-NORD project EstWN is being converted to XML-format in compliance with the recently completed KYOTO project. Furthermore, a state of the art editing tool, which can produce XML

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

markup, is being developed for further extensions.

3.2 Finnish WordNet, FinnWordNet

FinnWordNet is a recently built wordnet for Finnish developed at the University of Helsinki (cf. Lindén &Carlson 2010). It complies with the structure of Princeton WordNet. It was created by translating all the synsets in Princeton WordNet, and it is open source and contains over 117,000 synsets. After the translation, various things have been done in order to check the quality of the manual translations, e.g. spelling correction, word class consistency correction and some translation correction.

Currently, methods for improving and expanding the content of FinnWordNet are being developed. We have tested methods for finding a location for a new word in the FinnWordNet hierarchy. Since wordnets are structured ontologies, a location for a word can be pinpointed by its relations to other words. Finding a location for a new word means finding a hypernym, a hyponym or a synonym in FinnWordnet. The methods include searching for multiword terms, compound words and using lexico-syntactic patterns. It has also been explored which types of corpora are useful for this task and WikiPedia was found to be valuable in several ways due to its multilingual nature as well as its textual structure.

3.3 Swedish Wordnet, Swesaurus

Swesaurus, a free Swedish wordnet developed at Språkbanken, University of Gothenburg, is constructed by reusing information about lexical-semantic relations in a number of pre-existing freely available lexical resources: SALDO (Borin et al. 2008; Borin and Forsberg 2009), SDB (Järborg 2001), Synlex (Kann and Rosell 2006) and Swedish Wiktionary.

The SALDO resource constitutes the backbone of Swesaurus. SALDO is a large-scale lexical resource providing an inventory of 117k persistent sense identifiers, a morphology of 1.7 MW, and associative semantic relations connecting all senses (somewhat similar to 'evocation' in the WordNet context; Boyd-Graber et al. 2006).

A novel feature of Swesaurus is its fuzzy synsets derived from the graded synonymy relations of Synlex. The recognition of fuzzy synonymy raises many intricate methodological and theoretical questions, e.g., the effect on other lexical-semantic relations, such as hyponymy or meronymy.

As part of the META-NORD project, a linking between Swesaurus and Core WordNet has been completed. The linkage was bootstrapped by using the Lexin basic Swedish-English dictionary. Swedish lemmas in Lexin were automatically linked, in an overgenerating manner, to SALDO sense identifiers, giving us a set of senses for every lemma. The glosses of Core WordNet were subsequently, via Lexin, linked to these sense sets. Core WordNet has 5,000 entries, of which around 89% were covered by Lexin. Furthermore, 23% had a unique link to one SALDO sense, and the remaining an average ambiguity of 4.4 (a rather high ambiguity, but not unexpected for a core vocabulary).

Swesaurus is a part of a larger lexical project, SweFN++, and its development version is published through the lexical infrastructure of SweFN++ on a daily basis. Swesaurus and several other lexical resources are open source available for download and inspection at spraakbanken.gu.se/eng/sblex.

3.4 Danish Wordnet, DanNet

In contrast to most other wordnets, DanNet has been constructed using the so-called merge approach where the wordnet is built on monolingual grounds and thereafter merged with PWN. 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 inspected in a browser from www.andreord.dk. The wordnet has been compiled as a collaboration between the University of Copenhagen and the Danish Society for Language and Literature.

Since the starting point of DanNet was a corpus-based, newly completed dictionary of Danish accessible in a machine-readable version with hypernymy information explicitly specified for each sense definition (Den Danske Ordbog), the motivation for the monolingual approach was obvious. Furthermore, the Danish version of the SIMPLE lexicons (cf. Lenci et al. 2001, and for Danish Pedersen & Paggio 2004) has influenced the construction of DanNet in the sense that it includes also qualia information such as the telic (PURPOSE) and the agentive role (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.

3.5 Icelandic Wordnet

Icelandic wordnet is in its early stage of development. It applies the monolingual approach and builds on previous work in the extraction of lexical semantic information from a monolingual dictionary of Icelandic (Nikulásdóttir and Whelpton, 2009; Nikulásdóttir, 2007 Nikulásdóttir & Whelpton. 2010) and seeks to use a mixture of pattern matching and statistical methods for relation extraction, given the promising results from this hybrid methodology in recent years (Cederberg and Widdows, 2003; Cimiano, 2006; Pantel and Pennacchiotti, 2008).

3.6 Norwegian Wordnet(s)

The compiling of a Norwegian wordnet for Norwegian bokmål and Nynorsk is being launched in 2011 by the language initiative Språkbanken and will be developed by the company Kaldera Language Technology. It has been decided to translate from the Danish wordnet, DanNet, and subsequently adjust to Norwegian. The goal is to complete the wordnet(s) in 2013.

4 Methodological Considerations on Linking and Validation

Where the establishment of multilingual resources has a very clear utility value in language technology applications, the purpose of this linking exercise is in fact twofold: Feasibility of cross-lingual linking via the “core synsets” as well as a comparison/validation of the monolingual wordnets. The linking is performed along the 5,000 Princeton Core synsets on the one hand and the selected languages on the other. The primary aim of the task is rather to provide a qualified feasibility study of the perspectives of such a linking at a larger scale and last but not least to give some valuable insights in the very diverse characteristics of the selected wordnets. The main questions to be examined in such a validation are the following:

- ***Taxonomical structure***: Do different approaches generally lead to different taxonomical structures of the lexical networks, and can we to some extend define best practice regarding depth of structure? (For instance, should wordnets generally cover the layman or the expert perspective?)

- ***Coverage.*** Are frequent concepts in the target language covered well enough when compiling a wordnet via English? And when deducing it from a traditional lexical resource? Can we define a coverage “pain threshold”? These and related issues will be evaluated using corpora and existing core vocabulary lists.

- ***Granularity of the described concepts.*** Does a specific approach result in many or few sense distinctions (i.e. synonym sets) for each lemma? Is it possible to identify a technology-oriented best practice for sense granularity (i.e. something that corresponds to main senses in traditional lexicography?)

- ***Completeness of synonym sets.*** Does a given approach bring about many or few semantic relations and/or semantic features per concept? And can a best practice set of semantic relations be established along the validated wordnets?

For illustration of difference in taxonomical characteristics, consider Figure 1 and 2 which show discrepant approaches regarding when to apply a zoological, highly taxonomical perspective and when to apply a simpler, layman approach.

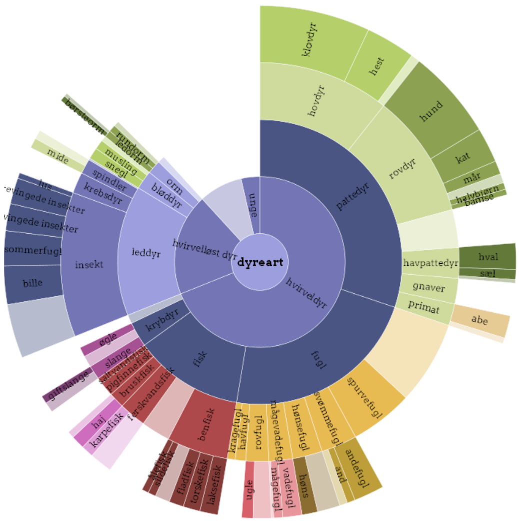


Fig. 1: Taxonomy of animals in DanNet, highly inspired by the zoological taxonomy.

In the case of animals, DanNet adopts a specialist view, where the Icelandic has taken a one-dimensional, layman perspective.[3]

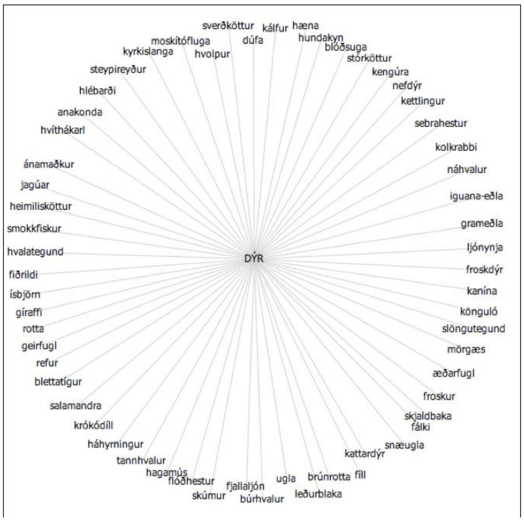


Figure 2: Flat taxonomy of animals in Icelandic Wordnet following a layman approach.

Also if we consider the sets of semantic relations established in the relevant wordnets, we find substantial differences which require further examination. Although the Danish and Swedish wordnets both adopt monolingual approaches, DanNet relates in a stricter way to classical wordnet relations than SALDO/Swesaurus, as is shown in Figure 3 and 4.

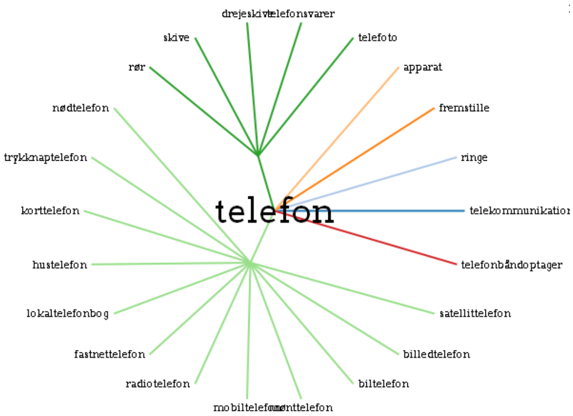
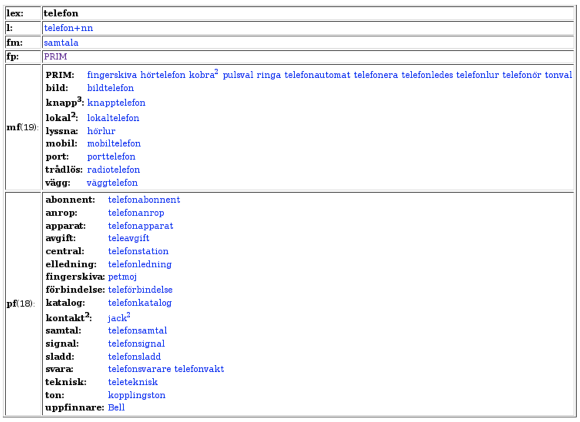


Figure 3: Semantic relations to telephone in DanNet following basically the lines of the Princeton WordNet relations (light green illustrates has_hyponym, yellow has_hyperonym, dark green has_mero_part, light blue purpose_of etc).

In the Swedish wordnet we find a slightly more associative approach to semantic relations where telephone is furthermore associated to concepts like *samtala* 'hold a conversation', *telefonledes* 'by phone', *mobiltelefon* 'mobile phone'.



lex:	**telefon**	
l:	telefon+nn	
fm:	samtala	
fp:	PRIM	
mf(19):	**PRIM:**	fingerskiva hörtelefon kobra[2] pulsval ringa telefonautomat telefonera telefonledes telefonlur telefonör tonval
	bild:	bildtelefon
	knapp[3]:	knapptelefon
	lokal[2]:	lokaltelefon
	lyssna:	hörlur
	mobil:	mobiltelefon
	port:	porttelefon
	trådlös:	radiotelefon
	vägg:	väggtelefon
pf(18):	**abonnent:**	telefonabonnent
	anrop:	telefonanrop
	apparat:	telefonapparat
	avgift:	teleavgift
	central:	telefonstation
	elledning:	telefonledning
	fingerskiva:	petmoj
	förbindelse:	teleförbindelse
	katalog:	telefonkatalog
	kontakt[2]:	jack[2]
	samtal:	telefonsamtal
	signal:	telefonsignal
	sladd:	telefonsladd
	svara:	telefonsvarare telefonvakt
	teknisk:	teleteknisk
	ton:	kopplingston
	uppfinnare:	Bell

Fig. 5: Semantic (associative) relations for *telefon* in SALDO.

At the current stage of the project, Danish, Finnish, Swedish and Estonian wordnets have been linked to Princeton Core WordNet, thereby providing a common, linked coverage of the previously mentioned 5,000 core synsets. Next step is to provide a viewer which enables evaluators to see the cross-lingual links in a flexible manner so that validation along the lines described above can be performed in a direct fashion. A possible approach could be the one adapted in the Danish viewer 'andreord.dk' were cross-lingual links between DanNet and Princeton Core are shown as direct or indirect alignments.

5 Conclusions

According to the BLARK (Basic Language Resource Kit) scheme, wordnets along with treebanks and other resources, are crucial when building language enabled applications. BLARK lists Computer Assisted Language Learning (CALL), speech input, speech output, dialogue systems, document production, information access and translation applications as dependent of wordnets. The semantic proximity metrics among words and concepts defined by a wordnet are very useful in such applications because in addition to identical concepts, the occurrence of words with similar (more general or more specif-

[3] Actually, the Danish wordnet differs internally with regards to layman or expert perspective. Based on The Danish Dictionary, the general approach is that of the layman, but in certain corners of the resource, an expert view has proven dominant.

ic) meanings contribute to measuring the similarity of content.

As has been presented in this paper, most Nordic and Baltic countries are in the fortunate situation where wordnets are already built or are being built right now. However, it is crucial that we continuously adapt them to currents standards and let them undergo cross-lingual comparison and validation in order to ensure that they become of the highest possible quality and usefulness for future, hopefully also multilingual applications. META-NORD provides a unique opportunity for such a validation across languages.

References

Borin, L., M. Forsberg 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. Odense: NEALT, 7–12*.

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

Boyd-Graber, J., C.Fellbaum, D.Osherson, R. Shapire 2006. Adding dense, weighted connections to WordNet. In: *Proceedings of the Global Wordnet Conference 2006*. Brno: Masaryk University, 29-35.

Cederberg, S. and D. Widdows. 2003. Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction. In *Proceedings of the International Conference on Natural Language Learning (CoNLL), pages 111–118.*

Cederberg, S. and D. Widdows. 2003. Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction. In *Proceedings of the International Conference on Natural Language Learning (CoNLL), pages 111–118.*

Cimiano, P. 2006. *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer.

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

Kann, Viggo, Magnus Rosell 2006. Free construction of a free Swedish dictionary of synonyms In: *Proceedings of the 15th NODALIDA conference*. Joensuu: University of Eastern Finland, 105–110.

Kerner, K.; Orav, H. Parm, S. (2010). Growth and Revision of Estonian WordNet. In: Principles, Construction and Application of Multilingual Wordnets. *Proceeding of the 5th Global Wordnet Conference: 5th Global Wordnet Conference; Mumbai, India*; (Ed.) Bhattacharyya, P.; Fellbaum, Ch.; Vossen, P.. Mumbai, India: Narosa Publishing House, 2010, 198 - 202.

Lindén, K. and L.Carlson. 2010. FinnWordNet – WordNet på finska via översättning [FinnWordNet - WordNet in Finnish via Translation]. LexicoNordica – Nordic Journal of Lexicography, 17:119–140.

Nikulásdóttir, A. B. 2007. *Automatische Extrahierung von semantischen Relationen aus einemeinsprachigen Isländischen Wörterbuch*. MA-Thesis, University of Heidelberg.

Nikulásdóttir, A. B. and Matthew Whelpton. 2010. Lexicon Acquisition through Noun Clustering. *LexicoNordica* 17:141-161.

Nikulásdóttir, A. B. and M. Whelpton. 2009. Automatic extraction of semantic relations for less resourced languages. In Bolette Sandford Pedersen, Anna Braasch, Sanni Nimb, and RuthVatvedt Fjeld Editors), *Proceedings of the Workshop "Wordnets and other Lexical SemanticResources - between Lexical Semantics, Lexicography, Terminology and Formal Ontologies",NODALIDA 2009.* Odense, Denmark: NEALT Proceedings Series Volume 7, pages 1-6. Northern European Association for Language Technology (NEALT), Tartu University Library.

Pantel, P. and M. Pennacchiotti. 2008. Automatically Harvesting and Ontologizing Semantic Relations. In Paul Buitelaar and Philipp Cimiano, editors, *Ontology Learning and Population: Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text*. IOS Press.

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, Computational Linguistics Series.* Volume 43, Issue 3:269-299.

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