

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.se

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

Eiríkur 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 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*.

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

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

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 embarked 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 is a need for upgrade of several of the wordnet resources to agreed standards, which will thus constitute a preliminary task of this META-NORD action.

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 University², pilot cross-lingual resources will be 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” is being 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).

² <http://wordnetcode.princeton.edu/standoff-files/core-wordnet.txt>

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.

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 additives. 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).

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 formulate XML 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 recently been initiated. The linkage was boot-

strapped 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 5k 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 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

- Lexical Semantics, Lexicography, Terminology and Formal Ontologies*. Odense: NEALT, 7–12.
- Borin, L., M. Forsberg, L.Lönngrén 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 Sågvald 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 Ruth Vatvedt Fjeld Editors), *Proceedings of the Workshop "Wordnets and other Lexical Semantic Resources - 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.