Human Language Technologies – The Baltic Perspective
A. Tavast et al. (Eds.)

doi:10.3233/978-1-61499-133-5-96

Cross-linking Experience of Estonian WordNet

Neeme KAHUSK [a,1], Heili ORAV [b] and Kadri VARE [b]
[a] *Institute of Computer Science, University of Tartu*
[b] *Institute of Estonian and General Linguistics, University of Tartu*

Abstract. Our paper describes work we have done for Estonian WordNet according to META-NORD project tasks. We discuss the linking process of Estonian WordNet and Core WordNet from linguistic, lexicographical and technical point of view. Also, cross-language linking is briefly described.

Keywords. Estonian, WordNet, Core WordNet, cross-language linking, META-NORD, EuroWordNet

Introduction

WordNet is one of the most well-known lexico-semantic resources that is used in language technology. There is a growing number of national wordnets besides the original Princeton WordNet for English.

In spite of the fact that the number of publications about WordNet and followers is imposing [1], it would not do harm to repeat some of the most basic ideas over and over.

The main unit of a wordnet-type thesaurus is *concept*, presented as synonym set or *synset* for short in wordnet jargon. Synsets are organised according to part of speech, each synset belongs to exactly one part of speech. Words, or *lexical units*, to be precise (some of them are multi-word expressions) *in a certain sense* represent the synset. Lexical units that form a synset are interchangeable in certain context. A lexical unit, its sense number and part-of-speech tag make a unique combination to identify synset.

Synsets are connected with each other via *semantic relations*, also called *links*. The most prominent of them is hierarchy-building *hyponymy/hyperonymy* link. This kind of relations are between general and specific concepts. There are 17 different kinds of semantic relations in Princeton WordNet (version 1.5) [2]. Semantic relations are limited to certain parts of speech.

One of the first projects to create wordnets other than English was the EC financed EuroWordNet project. It was a European resources and development project supported by the Human Language Technology sector of the Telematics Applications Programme. In framework of this project seven national wordnets were created and linked to Princeton English WordNet. [3]

[1] Corresponding Author: Neeme Kahusk, Institute of Computer Science, University of Tartu, Liivi 2, 50409 Tartu, Estonia; E-mail: neeme.kahusk@ut.ee

Although the general principles of design are all the same for every wordnet, EuroWordNet layout differs from Princeton to some extent.

While Princeton WordNet has one definition (or *gloss* in wordnet jargon) per synset, then EuroWordNet format allows gloss and examples for every lexical unit in synset.

The sets of semantic relations in Princeton WordNet and EuroWordNet are a bit different. New relationships, including relationships across parts of speech, were introduced to EuroWordNet; hyponymy and hyperonymy across parts of speech are permitted. EuroWordNet uses hyponymy relation for verbs as well, in Princeton WordNet troponymy is used (see [4] for discussion about this topic). There are also role–patient relations and causality. Meronymy in EuroWordNet is more fine-grained than in Princeton WordNet, including `has_mero_madeof`, `has_mero_member`, `has_mero_part`, `has_mero_portion` (sub)relations. EuroWordNet has introduced a general unspecified relation, `fuzzynym` as well.

EuroWordNet introduced a whole set of specific semantic relations, that are used to interconnect wordnets in different languages. Inter-lingual-index (ILI) serves for that purpose. Basically, ILI is a special version of Princeton WordNet version 1.5. This enables use of EuroWordNet as multilingual lexicon. The most common ILI relation is `eq_synonym`, but there are also other relations possible, the more important ones being `eq_near_synonym`, `eq_has_hyperonym`, and `eq_has_hyponym`. The last ones can be used as ordinary hyperonymy-hyponymy relations, if ILI concept has broader or narrower meaning, respectively.

The technical implementations of Princeton WordNet and EuroWordNet differ too. Princeton WordNet data is stored in plain text files, there are files for each part of speech, plus index files. Detailed information about file structure is described in documentation that comes with wordnet distribution.

The standard tool for editing EuroWordnet was Polaris by Lernout and Hauspie. It used its own database format, and enabled to export and import data as plain text. As the inner database is proprietary and the description is not available, practical importance has the export-import text file. The Polaris tool and export file format are described in [5].

There has been done much work in providing various tools for accessing wordnet data, and to convert it into different databases and to provide libraries for several programming languages. The same stands for EuroWordNet. The need for a decent format and tools is even more urgent, as Polaris is very out-of-date. There are many good tools already, and plenty of them coming, we suppose. One of the best known and probably most used is DebVisDic and its predecessor VisDic (see [6] and [7], respectively).

EuroWordNet was started as a uniform multilingual lexical resource. Although there were used different strategies for building monolingual parts of the resource, the main idea and general layout was the same, the same stands for file format and tools used. META-NORD project sets up a more difficult task: to link across languages wordnets that have been developed under several projects, by using different strategies and distributed in different file formats. Hereby we will focus on our effort to manage with Estonian part of the task.

1. Estonian WordNet

Estonian WordNet contains at present (June 2012) more than 55 000 concepts (synsets) and the extending of the resource is still ongoing process. The lexical-semantic database contains nouns, verbs, adjectives and adverbs; there are multiword units as well. The main vocabulary of Estonian language was mostly covered by 2010 [8].

The design procedure of the Estonian WordNet during more than 10 years has followed different strategies. Firstly, the words (literals) to include were selected on frequency basis. Secondly, our chosen approach has been domain-specific, i.e we have added semantic domains like architecture, transportation, personality traits and so on. Thirdly, there are some endeavours to add derivatives automatically. Fourthly, we have used the results of sense disambiguation process [9].

Up to nowadays we have used the old Polaris tool for editing Estonian WordNet. This limits our possibilities to the semantic relations listed in Polaris files and ILI records to Princeton WordNet version 1.5. We have given a try to DebVisDic, but mostly because of the differences in formats, we have sticked to Polaris export format. DebVisDic uses XML, but the (default) schema resembles more Princeton wordnet, and if we would re-export the data into Polaris export file and would like to re-import data into Polaris, we would have some loss of information.

We have developed some tools to convert the Polaris export files to and from various formats. The tools are based on Python Eurowordnet module developed by us earlier [10]. There are included tools to convert Princeton WordNet into EuroWordNet (Polaris export), EuroWordNet into MySQL, EuroWordNet into various XML formats (DebVisDic and Kyoto [11] are included) in Python Eurown module. There exists a conversion of wordnet to MySQL [12], but this one is closely entwined with Princeton WordNet.

For browsing and searching Estonian WordNet, we have created a web-based application Teksaurus [13].

2. Wordnets in META-NORD Project

META-NORD is an EC project closely related to the META-NET network, whose aim is to take care of technological support to multilingual European information society. One of the key activities of META-NET is diminish high fragmentation and lack of unified access to language resources that hinder European innovation potential in language technology development and research.

The META-NORD project aims to establish an open linguistic infrastructure in the Baltic and Nordic countries. The project will focus on eight European languages — Danish, Estonian, Finnish, Icelandic, Latvian, Lithuanian, Norwegian and Swedish. See [14] for overview of the Project.

Besides general objectives META-NORD has several specific targets, and providing expertise in wordnets is one of them. The concept-based resource with ontology-like structure makes a good starting point for multilingual information retrieval or rule-based machine translation.

Currently there are six wordnets involved in META-NORD wordnet activities, Norwegian is the most fresh and Estonian the oldest one. See [16] for overview of wordnets in META-NORD.

3. Linking with Core Wordnet

There is a subset of Princeton WordNet called *core* wordnet. This consists of about 5000 most frequently used English word senses, compiled semi-automatically [15].

Core wordnet is part of WN ver. 3.0, and Estonian WordNet is by default linked to Princeton WordNet ver. 1.5 via many different semantic relations, so we had to map Estonian WordNet to Princeton WordNet ver. 3.0 at first, and later on adjust the results manually.

We haven't employed an English editor to establish the equivalents to Princeton WordNet, Estonian lexicographer who adds a new synset also adds an `equal_relation` to ILI. That has caused unfortunately often mistakes or imprecise links, especially if a specific vocabulary is concerned.

We have used mappings from Princeton WordNet 1.5 to 3.0 provided by the NLP group of the Universitat Politècnica de Catalunya (UPC, see [18], the method they used is described in [17]) to link Estonian WordNet synsets to Princeton WordNet 3.0. The mappings come as plain text files, for each part-of-speech one file. In files, there are rows for each synset in source-WN version. Each row contains at least three fields, separated by blanks. First field is for synset number of source-WN version, second and third fields are for target-WN version and probability of the mapping. If the probability is equal to 1, then there are no more fields; if the probability is less than 1, then there are one or more pairs of fields, each of them containing the target-WN version synset number and probability. There is not always the other fields present, in fact, there are no other fields if the probability is higher than 0.9.

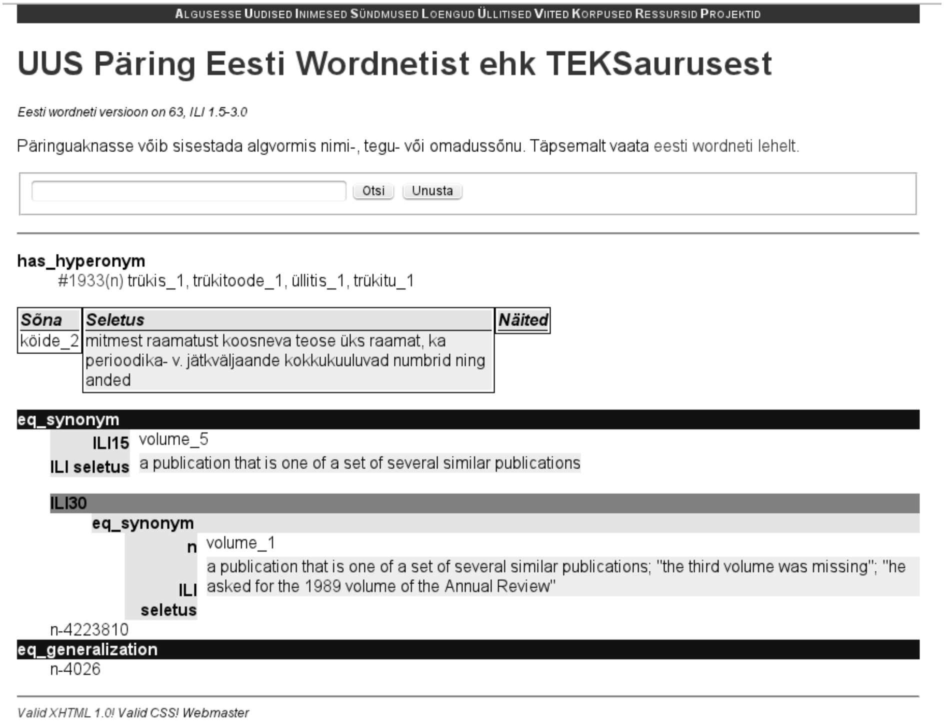
Since there are more types of ILI links in Estonian WordNet than `eq_synonym`, we have used a two-level mapping.

To map Estonian WordNet to core wordnet, we had to start from the core data. Core wordnet is laid out as plain text file that contains on each row part of speech, index, lexical unit and gloss. For our purpose it was enough to consider part of speech and index. According to these two fields we found ILI-compatible index that consists of wordnet-file offset and part-of-speech tag.

For mapping from core wordnet to Estonian WordNet we used UPC mapping files 3.0 to 1.5. There are mappings from older versions to newer ones and vice versa, and these mappings may be different because of the method that was used to generate the mappings (see [17]).

If there is `eq_synonym` or `eq_near_synonym` relation between ILI synset and Estonian WordNet synset, and mapping probability is 1, then the relation between WN 3.0 synset and corresponding Estonian WordNet synset would remain the same. If there is `eq_synonym` relation between ILI synset and Estonian WordNet synset, and mapping probability is less than 1, then the relation between WN 3.0 synset and corresponding Estonian WordNet synsets would be linked with `eq_near_synonym`. If other interlingual relations are between WN 1.5 and Estonian WordNet synsets, then the relations between Estonian WordNet and WN 3.0 would remain the same for every probability. Each relation between Estonian WordNet and WN 3.0 that has less probability than 1 would be marked for later revision.

The resulting mappings are also released as a (test) version of Teksaurus (see Fig 1).


ALGUSESSE UUDISED INIMESED SÜNDMUSED LOENGUD ÜLLITISED VIITED KORPUSED RESSURSID PROJEKTID
Otsi
Unusta

UUS Päring Eesti Wordnetist ehk TEKSaurusest

Eesti wordneti versioon on 63, ILI 1.5-3.0

Päringuaknasse võib sisestada algvormis nimi-, tegu- või omadussõnu. Täpsemalt vaata eesti wordneti lehelt.

has_hyperonym
#1933(n) trükis_1, trükitoode_1, üllitis_1, trükitu_1

Sõna	***Seletus***	***Näited***
köide_2	mitmest raamatust koosneva teose üks raamat, ka perioodika- v. jätkväljaande kokkukuuluvad numbrid ning anded	

eq_synonym
ILI15 volume_5
ILI seletus a publication that is one of a set of several similar publications

ILI30
eq_synonym
n volume_1
ILI seletus a publication that is one of a set of several similar publications; "the third volume was missing"; "he asked for the 1989 volume of the Annual Review"
n-4223810

eq_generalization
n-4026

Valid XHTML 1.0! Valid CSS! Webmaster

Figure 1. A test version of Teksaurus, running on MySQL backend, showing both links to WN 1.5 and WN 3.0

4. Results

Testing with core wordnet has suggested that about 30% of mappings to WN 3.0 would need some adjusting or editing.

Since Estonian WordNet has grown quite large in size it is difficult to detect mistakes that slip in due to the manual work of lexicographers. After linking Estonian WordNet with core wordnet it is possible to revise different types of problems related with cross-language semantic relations.

Multiple equal synonyms were linked to one and the same Princeton WordNet synset. This situation indicates that there is a possible mistake in Estonian WordNet — either the two synsets are too fine-grained and should be merged, or one of the synsets in Estonian WordNet should be linked to different Princeton WordNet synset altogether. For example, *kostma* (‘say in reply’) and *vastama, vastust andma* (‘answer’) both are linked with `eq_synonym` ‘answer’.

Synsets are linked with a wrong semantic relation. For the other example — *peatuma* (‘decide by choosing’) and *otsustama, otsusele jõudma* (‘decide’) — equal synonyms and equal near synonyms between Estonian WordNet and core wordnet were marked wrongly. Another example is *vein* `eq_synonym` ‘wine’ and *naturaalvein* (‘natural wine’) `eq_near_synonym` ‘wine’, but the latter relation should be `eq_has_hyperonym` instead. Of course in these cases the subjectivity of a lexicographer also appears; but still it is worth looking through synsets that are linked both `eq_synonym` and `eq_near_synonym`.

There are problems with part of speech also, for example, `eq_synonym` relation was set between adverb and noun, verb, or adjective.

And, of course, it is possible to add missing synsets (especially adjectives and adverbs) from the core wordnet list to Estonian WordNet in order to complete the core vocabulary.

Interesting would be also to compare linked synsets in core wordnet with the most frequent synsets and literals in the word sense disambiguated corpus of Estonian.

Acknowledgements

Estonian WordNet is deveoped mostly with the suppport from two grants: the National Programme of Estonian Language Technology and META-NORD.

Completing of Estonian WordNet has been financed grant EKT-2 from the National Programme of Estonian Language Technology.

The work within the project META-NORD has received funding from the ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, Grant agreement no 270899

References

[1] Andras Csomai. WordNet Bibliography, 2007. Accessible at http://lit.csci.unt.edu/~wordnet/. Last update: October 24, 2007.

[2] Sandra M. Harabagiu and Dan I. Moldovan. Knowledge Processing on an Extended WordNet. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 379–405. MIT Press, 1998.

[3] EuroWordNet Website. http://www.illc.uva.nl/EuroWordNet/.

[4] Christiane Fellbaum. A semantic network of English verbs. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 69–104. MIT Press, 1998.

[5] Michael Louw. *Polaris User's Guide: The EuroWordNet Database Editor*. Deliverable D024, WP6.5, EuroWordNet, LE2-4003, 1998. Downloadable at http://www.illc.uva.nl/EuroWordNet/docs.html.

[6] Aleš Horák, Karel Pala, Adam Rambousek, and Martin Povolný. DEBVisDic — First Version of New Client-Server Wordnet Browsing and Editing Tool. In Petr Sojka, Key-Sun Choi, Christiane Fellbaum, and Piek Vossen, editors, *GWC 2006, Proceedings*, Brno, 2005. Masaryk University.

[7] Aleš Horák and Pavel Smrš. VisDic — Wordnet Browsing and Editing Tool. In Petr Sojka, Karel Pala, Pavel Smrš, Christiane Fellbaum, and Piek Vossen, editors, *GWC 2004, Proceedings*, Brno, 2003. Masaryk University.

[8] Haldur Õim, Heili Orav, Kadri Kerner, and Neeme Kahusk. Main Trends in Semantic-Research of Estonian Language Technology. In Inguna Skadiņa and Andrejs Vasiljevs, editors, *Human Language Technologies: The Baltic Perspective. Proceedings of the Fourth International Conference Baltic HLT 2010*, pages 201–207. IOS Press, 2010.

[9] Kadri Vider and Neeme Kahusk. Estonian WordNet Benefits from Word Sense Disambiguation. In *Proceedings of the 1st International Global WordNet Conference: 1st International Global WordNet Conference, Mysore, India*, pages 26–31, Mysore, India, 2002. Central Institute of Indian Languages.

[10] Neeme Kahusk. Eurown: an Eurowordnet module for Python. In Pushpak Bhattacharyya, Christiane Fellbaum, and Piek Vossen, editors, *Principles, Construction and Application of Multilingual Wordnets. Proceeding of the 5th Global Wordnet Conference: The 5th International Conference of the Global WordNet Association (GWC–2010)*, pages 360–364, Mumbai, India, 2010. Narosa Publishing House.

[11] KYOTO Project Website. http://www.kyoto-project.eu/.

[12] WordNet SQL Website. http://wnsql.sourceforge.net/.

[13] Neeme Kahusk and Kadri Vider. TEKsaurus — the Estonian WordNet online. In *The Second Baltic Conference on Human Language Technologies*, pages 273–278, 2005.

[14] META-NORD Project Website. http://www.meta-nord.eu/.

[15] Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson, and Robert Schapire. Adding dense, weighted, connections to WordNet. In Petr Sojka, Key-Sun Choi, Christiane Fellbaum, and Piek Vossen, editors, *Proceedings of the Third Global WordNet Conference, GWC 2006*, pages 29–35, 2006.

[16] Bolette Sandford Pedersen, Lars Borin, Markus Forsberg, Krister Linden, Heili Orav, and Eirikur Rögnvaldsson. Linking and Validating Nordic and Baltic Wordnets — A Multilingual Action in META-NORD. In Christiane Fellbaum and Piek Vossen, editors, *Proceedings of 6th International Global Wordnet Conference: 6th International Global Wordnet Conference, Matsue, Japan*, pages 254–259, 2012.

[17] J. Daudé, L. Padró, and G. Rigau. Mapping wordnets using structural information. In *Proceedings 38th Annual Meeting of The Association for Computational Linguistics (ACL00). Hong Kong*, 2000.

[18] WordNet Mappings Website. http://nlp.lsi.upc.edu/web/index.php?option=com_content&task=view&id=21&Itemid=57.