

# Growth and Revision of Estonian WordNet

**Kadri Kerner**

University of Tartu  
Institute of Estonian and  
General Linguistics  
Liivi 2–308, Tartu, Estonia  
kadri.kerner@ut.ee

**Heili Orav**

University of Tartu  
Institute of Estonian and  
General Linguistics  
Liivi 2–308, Tartu, Estonia  
heili.orav@ut.ee

**Sirli Parm**

University of Tartu  
Institute of Estonian and  
General Linguistics  
Liivi 2–308, Tartu, Estonia  
sirli.parm@ut.ee

## Abstract

Up to the present day, the main goal has been the increasing process of Estonian WordNet (EstWN) and currently it consists of more than 27 000 concepts (November 2009). Now we have started also the revision process of existing data in EstWN, since we have educated personnel and also the knowledge of the deficiency and needs of the thesaurus. Firstly an overview of current data in EstWN is given. Then we will address the different problems of automatically transferred synsets, also this paper describes the mini-test made to check the adverb's sense distinctions and definitions in EstWN. One of the problems is the revising the hierarchies and one study has been carried out – the checking of the taxonomy of human being in EstWN. Finally we address the representation of systematic polysemy in EstWN. The paper concludes with some of the future plans and connections with other projects.

## 1 Introduction

There are two thesauri available for Estonian. First thesaurus (Saareste, 1979) has an historic value (compiled by Andrus Saareste as war refugee in Uppsala in 1979) and second, the modern one is the *wordnet*-type thesaurus of Estonian. Estonian WordNet's basic idea is the creation of theoretically systematic and applicably proper network of meanings because it is useful for some natural language applications. WordNet as a valuable resource can be used for example in Semantic Web, ontology, word sense disambiguation systems, machine translation etc.

The Estonian team joined the wordnet community (EuroWordNet-2<sup>1</sup>) from the beginning of

January 1998. In the framework of the project of Estonian Language Technology the Estonian WordNet has been created during the years 1997–2000. After some discontinuation this project was awoken again. In 2006 started the project for increasing EstWN and is supported by Estonian National Programme on Human Language Technology. Thanks to governmental program our thesaurus has enlarged a lot. The EstWN at the present stage includes about 19600 noun synsets, 1700 adjective synsets, 4000 verb and 1700 adverb synsets.

Our chosen approach so far for enlarging has been manual and domain-specific, i.e we have added concepts from semantic fields like architecture, transportation, personality traits and so on. There are 45 different types of semantic relations present. The most frequent relation among nouns and verbs is hyperonymy/hyponymy. Near\_synonymy and near\_antonymy are more frequent among adverbs and adjectives. Since one person is dealing with one domain at the time, then it makes the relations between different concepts (in one domain) easier to determine. For example from the domain of architecture the concept *antiiktempel* ('antique tempel') has 1 hyperonym, 11 hyponyms, 1 has\_holo\_part and 8 has\_mero\_part relations.

Besides adding new concepts to EstWN we have started the checking of existing language material.

## 2 Revising automatically transferred synsets in EstWN

New synsets to EstWN have been and are mostly added manually, but during the increasing process of EstWN around 3000 noun synsets were automatically transferred from the Estonian Synonym Dictionary (Õim, 1991). The Synonym

---

<sup>1</sup> <http://www.illc.uva.nl/EuroWordNet/>

Dictionary represents synonyms both in written and spoken language, also there are present some old words and dialect words. This dictionary includes different word senses and in some cases diminutive forms (especially for adjectives). The synonym line in this dictionary does not indicate only to absolute synonyms, since that was not the goal. The goal according to the author was to provide guidance of how to express something with different words. This dictionary is meant for a sensible user, who can choose wisely between all the proposed synonyms.

The Synonym Dictionary does not include any semantic relations (except synonymy), so it was possible to import only synonym lines. The definitions, examples, semantic relations and ILI-links were planned to add manually. Also, the manual check-over of these synsets was required, because many of synsets needed to be revised and corrected in order to include them to EstWN. During the manual work different kinds of problems about the imported synsets aroused.

Firstly, there were cases of old and dialect words that are not present in nowadays general language. For the most part old and dialect words are not included to EstWN. So, some of the old word synsets and dialect word synsets were completely removed (for example the synset *kupulasja, kuppär, kupumoor*; in English it is a kind of a healer who uses special kind of glass-shaped instruments for healing cough). Some of the imported synsets had one or two old/dialect words and the inappropriate words were removed from the synset. For example the synset *seelik, undruk, kõrt* ('skirt'), where Estonian word *kõrt* is rarely used dialect word and which carries slightly different meaning than the other two words in this synsets. (Villem, 2009:65)

Secondly, there were synsets, which needed to be joined together, because there were no difference between the senses, even the words in the synset were mostly the same. Another reason for uniting similar word senses is to avoid the over-grained sense distinctions. For example, *põlgus1, põlgamine1, põlg1, põlastus1, halvaks-panu1, jälestus1* ('contempt', 'insolence') and *põlgamine2, põlgus2, põlg2, põlastus2, halvaks-panu2, halvustus1* ('derogative', 'contempt', 'insolence'). Also, if these synsets stood separate, then it could not be possible to connect both with *eq\_synonym* relation to ILI, instead each synset would have got *near\_synonym* link. (Villem, 2009:68)

One of the most frequent problems was that both hyperonyms and hyponyms were included

in one synset. Also, if hyponyms as more specific senses are in the same synset with their hyperonyms, there is a problem of choosing the appropriate ILI-link. For example *teller, klienditeenindaja* ('teller, customer service representative'), where *teller* is a customer service worker in a bank. In a wordnet it is necessary to keep specific meanings from the general ones. Another example is *vahuvein* ('sparkling wine'), *šampanja* ('Champagne, more a spoken language variant'), *šampus* ('Champagne'), where both sparkling wine and Champagne are hyponyms of wine. (Villem, 2009:78)

Many of the synsets transferred contained also plural forms. Some synsets contained only plural forms; some consisted of both plural and singular word forms. It was necessary to determine, which of these plural forms actually carry a different meaning than its singular form. In some cases nothing needed to be corrected, for example, *kedrid, säärised* ('spat, spats, gaiter'), because these are plural words consisting of two same objects. Also, some synsets containing both plural and singular word forms held the same meaning left unchanged. For example, *kard* (singular form), *inglijuksed* (plural form), ('tinsel'), because the singular form of the word *inglijuksed* ('tinsel') does not hold the same sense than the plural form. But there were quite many cases where the words were groundlessly in plural. Although in real texts these words are usually used in plural, there is no difference of meanings between plural and singular forms. For example *memoriaal, mälestusvõistlused* ('memorial, memorials'), where *mälestusvõistlused* ('memorials') should be in singular form. (Villem, 2009:90)

The next problem deals with singular and plural forms. In lexicography it is advised to mark the use of singular and plural more explicitly. So-called complex-words describe things, which are in a sense the same, they are tied with the same action, they are meant to perform together, for example, *bikini* and *parents*. Also, a group representing an individual is described with plural for example *boys*. (Langemets, 2004: 745)

Finally we faced the genus issues. Often one synset contained neutral, feminine and masculine gender. In some cases it was possible to organize the words in these synsets into hyperonyms and hyponyms, and then it was also possible to connect synsets with appropriate links to ILI. For example, *klassikaaslane, klassiõde, klassivend* ('classmate', "class-sister", "class-brother"),

where classmate could be the hyperonym for class-sister and class-brother. Another example, *lesk*, *leskmees*, *lesknaine* ('widow, widowman (widower), widow (widowwoman)'), where both genders are present in one synset, but it would be clearer if they were separated. (Villem, 2009:92)

### 3 Revising adverbs in EstWN

Another line of work has been with adverbs. We started to add adverbs into EstWN only few years ago and mostly we added adverbs of time, degree and also adverbs of manner and place. Estonian frequent adverbs of time have also multiple senses, for example adverb *veel* ('more, still, yet') has altogether 5 senses in EstWN. Previous work with nouns (Kerner, 2004) showed that human annotators cannot perceive too fine-grained sense distinctions in EstWN, so it was interesting to get some feedback from human judges about adverbs. So we carried out a questionnaire type of mini-test to check granularity of senses, definitions, and explanations of adverbs. Although we asked about three adverbs, it provided us with the information of how to improve the process of adding new adverbs in the future.

We selected three adverbs from the EstWN which are: *täna* ('today'), *veel* ('still/yet'; 'more'), *nimelt* ('namely'). The questionnaire was represented electronically on the Internet. We selected 10 sentences from the corpus of Estonian Written Language<sup>2</sup> for every adverb and asked people to choose an appropriate sense according to EstWN for each adverb. In addition, they were able to tag senses as 'not able to disambiguate' or 'correct sense missing'.

We received approximately 25 answers per each adverb, which gave us some information about these senses (usually manual disambiguation is performed by two people only).

The first adverb *täna* ('today') had two possible senses to choose from and we presented these sentences first, so the answerer could get the idea, and didn't feel too overwhelmed with too many senses.

**'täna1'** – today, at this moment, on this day (*I can't meet with you today.*);

**'täna2'** – today, at this time, at the present time, at present, nowadays, now, presently (*They now live in California*)

Based on human judges these two senses didn't combine with each other that often, so we could assume, that the sense distinctions in the EstWN are similar to the ones in real texts and also in human annotators mind.

The second adverb *veel* ('yet, still, more') had altogether 5 senses.

**'veel1'** – still (*It's still warm outside.*)

**'veel2'** – yet (*Mary is not yet at school; I have yet to see the results.*)

**'veel3'** – in an additional manner or place or at an additional time: more, in addition, else, further, other (*There's nothing more we can do.*)

**'veel4'** – more, to a greater extent (as in "more interesting")

**'veel5'** – a modal adverb with so called 'empty meaning'

For the adverb *veel* we assumed, that people do not distinguish between two time-related senses, sense number one and sense number two. The time senses 1 and 2 did not combine explicitly, but after a closer look, we discovered, that if people didn't know whether to choose one or the other time sense, they rather prefer the sense with empty meaning (the sense number 5). So we could assume, that these time senses one and two are still similar enough, so that human annotators cannot precisely disambiguate between them. Another conclusion is that the empty meaning of adverb *veel* is currently somewhat vaguely defined. The assumed result of clearly distinguishing between the time senses and in-addition sense also appeared.

The third adverb *nimelt* (eng namely, specifically etc) had three senses to choose from.

**'nimelt1'** – precisely, exactly, on the dot, just (*"Precisely, my lord," he said;*)

**'nimelt2'** – namely, viz., that is to say, videlicet (*That's how he is basically/namely;*)

**'nimelt3'** – spitefully, with spite, despitely (*He answered his accusers spitefully.*)

We assumed that some of the senses of adverb *nimelt* will combine with each other, and results show that indeed senses 1 and 2 tend to combine often and senses 1 and 3 also combine in some cases. Considering the problem of wordnets being too over-grained for some applications, we will consider probably joining these senses into

<sup>2</sup> <http://www.cl.ut.ee/korpused/baaskorpus/>

one. Also, it is useful to add antonymy relation to sense number three (for example: *accidentally* and *unwittingly, unintentionally* etc), in order to facilitate the distinction.

Considering this questionnaire it was draw some conclusions for future work: it is important to make the definitions of senses more clear and specific and to add more example sentences. The sense distinctions are not too fine-grained, they are quite in accordance with real texts.

#### 4 Revising the Taxonomies in EstWN

In this year one BA student defended her thesis which topic was taxonomy of human being in EstWN. In EstWN word *inimene* ('person') is the word with most hyponyms –more than 800 all together (Kirt, 2009). This study presented solutions of how to decrease the amount of the 'person's' hyponyms. For example the division of human beings was proposed accordingly to their professions (*professor, doctor*), place of living (*Estonian, Jew, American*), activity (*organizer, agitator*), relationships (*friend, mother, lover*), (emotional) state (*schlimazel, failure, beggar, waif*) and personality (*cheat, adventurer, hippie, hermit, yokel*). Every here described group has some sub-group also. All these were looked through and focus was concentrated on *hyponymy – hyperonymy* relation (Kirt, 2009). It became clear that it is necessary to review the current hierarchies, because it helps to correct the existing organization and creates a good foundation for the new accompanying synsets. Similar research has been done for Danish Wordnet (see Pedersen and Braasch, 2009) which shows that others compilers of wordnet-type thesauri have become to opinion how important is checking process parallel with composing.

#### 5 Revising the Systematic Polysemy in EstWN

In this summer a PhD thesis about the systematic polysemy of nouns in Estonian was completed (Langemets, 2009) and EstWN can also profit from this thesis. This thesis provides the systematic patterns of Estonian noun polysemy. Systematic polysemy is defined as a situation where several senses of at least two words regularly imply a similar semantic relation (Langemets, 2009:28). Systematic polysemy is one type of metonymy (Peters and Peters, 2000), for example, the pattern BUILDING-INSTITUTION ('school, theatre'). In Estonian WordNet the sys-

tematic polysemy is not marked explicitly, also it has been marked quite arbitrarily, for example 'school' has both BUILDING-INSTITUTION senses, 'theatre' is only in a INSTITUTION sense, 'university' in only in a INSTITUTION sense etc. In this thesis around 80 different types of systematic polysemy patterns are presented, so it is possible to add the missing senses and to make the representation of systematic polysemy more persistent. Another issue is how to represent systematic polysemy, one way is to add it as a complementary sense; the representation should be clear to human users as well as useful to some applications (for example in word sense disambiguation so-called underspecified senses, where an option of further specification exists (Buitelaar, 2000)).

#### 6 Conclusion

The main line of work is the revising the present data and at the same time extending Estonian WordNet with new concepts. Since 2006 we have added around 12 000 new concepts, EstWN contains of more than 27 000 concepts at present. There are more than 55 000 semantic relations currently present from which approx. 25 000 are added during last 3 years.

From the 3000 automatically transferred synsets around 60% needed to be corrected and revised. Also we had to insert semantic relations manually and the correction-work took up more time than we originally expected.

More semantic relations for the adverbs are needed to determine the appropriate senses, for example the derivational relation. The list of semantic relations for adverbs in EstWN is actually currently in creation.

Also, we plan to include labels from WordNet Domains (Magnini and Cavaglià, 2000) to EstWN. From the project of manually disambiguated word sense corpora it is possible to get feedback to current sense divisions; also information about missing concepts and words.

#### References

- Bentivogli, Luisa; Forner, Pamela; Magnini, Bernardo; Pianta, Emanuela. 2004. *Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing*. In COLING 2004 Workshop on "Multilingual Linguistic Resources", Geneva, Switzerland, August 28, pp. 101–108.
- Buitelaar, Paul. 2000. *Reducing Lexical Semantic Complexity with Systematic Polysemous Classes and Underspecification*. In: Proceedings of the

ANLP2000 Workshop on Syntactic and Semantic Complexity in Natural Language Processing Systems, Seattle, USA.

Kerner, Kadri. 2004. *Sõnatähendused tekstides ja teauruses ühestajate erimeelsuste põhjal*. BA thesis. University of Tartu.

Kirt, Riin. 2009. *Inimesega seotud hierarhiapuu eesti wordnetis*. BA thesis. University of Tartu.

Langemets, Margit. 2004. *Mõnda nimisõnade semantikat*. - Keel ja Kirjandus nr.10.

Langemets, Margit. 2009. *Systematic Polysemy of Nouns in Estonian and its Lexicographic treatment in Estonian Language*. Phd Thesis. University of Tallinn.

Pedersen, Braasch. 2009. *What do we need to know about humans? A view into the DanNet database*. Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009. Editors: Kristiina Jokinen and Eckhard Bick, pp 158–166.

Peters, Wim; Peters, Ivonne. 2000. *Lexicalised Systematic Polysemy in WordNet*. In: Proceedings of the Second International Conference on Language Resources and Evaluation, Athens, Greece.

Saareste, Andrus. 1958—1968. *Eesti keele mõisteline sõnaraamat I-IV. Dictionnaire analogique de la Estonienne I—IV*. Kirjastus Vaba Eesti, Stockholm.

Villem, Olga-Annikki. 2009. *ILI-kirjete lisamine Eesti wordneti ja selle käigus ilmnenud automaatselt genereeritud sünohulkade probleemkohad*. BA thesis. University of Tartu.

Õim, Asta. 1991. *Sünonüümisõnastik*, Tallinn.