Estonian Wordnet: Current State and Future Prospects

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

Kadri Vare
University of Tartu
kadri.vare@ut.ee

Sirli Zupping
University of Tartu
sirli.zupping@ut.ee

Abstract

This paper presents Estonian Wordnet (EstWN) with its latest developments. We are focusing on the time period of 2011–2017 because during this time EstWN project was supported by the National Programme for Estonian Language Technology (NPELT[1]). We describe which were the goals at the beginning of 2011 and what are the accomplishments today. This paper serves as a summarizing report about the progress of EstWN during this programme. While building EstWN we have been concentrating on the fact, that EstWN as a valuable Estonian resource would also be compatible in a common multilingual framework.

1 Estonian Wordnet: Project Progress

Estonian Wordnet is a lexical-semantic resource describing Estonian words and their lexical relationships. The history of EstWN starts already in 1998 when Estonian team joined the EuroWordNet (EWN) project (see also Vossen 1998). Back at 1998 the only available example was Princeton WordNet (PWN) (Fellbaum 1998), so the EWN project followed the same principles. The EWN added a completely new component – multilinguality – the possibility to link different languages via a central InterLingualIndex (ILI) that was based on PWN version 1.5 at that time.

At the beginning of 2011 the EstWN had reached around 40 000 concepts (including 10 000 synsets taken over automatically), by September 2017 there are around 85 000 concepts with 230 664 semantic relations and 135 497 senses in EstWN.

Over the years EstWN project has been mainly supported by the National Programme for Estonian Language Technology, the first programme lasted from 2006–2010 and the second one from 2011–2017. We greatly appreciate that the Estonian government has realized that it is crucial to support the creation of Estonian language resources so that the Estonian language is able to survive in the digital world among the larger languages.

There are two main directions in EstWN project – to add new and missing concepts and to improve the quality of existing data – for example performing the systematic revision of English equivalents and semantic relations or complementing EstWN with extra-information like sentiment, domain (see Bentivogli 2004) etc. Recently some wordnets have employed sentiment (opinion) information and also in EstWN 57 000 synsets have been automatically annotated with SentiWordNet's (see Baccianella et al. 2010) data. In addition to SentiWordNet, we have incorporated sense annotated vocabulary from the dictionary made for emotion detection (this vocabulary is manually tagged by linguists, see Pajupuu et al. 2016). Besides to the negative-positive-neutral scale, there is also contradictory-tag in this vocabulary, for example, *emotional, receptive* could be both positive or negative, depending on context. In the future, we plan to get sentiment tags for all synsets in the latest version of EstWN. In the long run, we expect that EstWN will be implemented more frequently as a language technology resource and for linguistic studies as well. Another important foresight is to belong into a unified global linguistic data infrastructure. While building EstWN we still follow general PWN principles and structure to enable linking, but at the same time, the EstWN should remain as language-specific as possible.

1.1 Where do new synsets come from?

Our team started to compile EstWN from translating base concepts and then we extended EstWN with the knowledge from different lexicons, corpora etc. Since EstWN has been mostly manual work of different people, then the semantic relations reflect largely human subjectivity. We have included vocabulary from dictionaries like Estonian Explanatory Dictionary, Orthological Dictionary, different terminology dictionaries, word frequency lists of corpora of written Estonian. Since general vocabulary of Estonian

[1] National Programme for Estonian Language, https://www.keeletehnoloogia.ee/en.

is covered, then we have moved on to special terminology. Although Martin Benjamin (2017) has written that "too many specialist terms would make PWN so unwieldy that the resource would become dysfunctional for users trying to sift through numerous esoteric senses" we continue to add vocabularies from different domains for the purpose of more broader usage of EstWN. Also, several students have contributed their work of the bachelor's thesis to improve EstWN – for example, the vocabulary of veganism, climate, transportation etc has deeply studied and semantic relations inside chosen vocabulary have been thoroughly examined. The computer game Alias which draws information from EstWN is also useful for feedback of the new and missing words and senses (we talked about it on last conference (Aller et al. 2016)).

1.2 Automatically generated synsets

At some point during the project, it seemed sensible to construct some part of the resource automatically. Only a few attempts have been made to increase the database (semi)-automatically before 2011. We have to admit, that these attempts haven't been overly successful and there are still problems to deal with.

Firstly, we included words that were missing from word sense disambiguation corpus but ended up with lots of proper names and words belonging already to some existing synset. Then synsets from the Dictionary of Synonyms were transferred automatically, but these synsets needed many corrections because the distinction between synonym and near-synonym was not clearly visible. Also, a lot of dialectal and archaic words were included, but not systematically or consistently.

Ideally, we would want to have a broad coverage of vocabulary. That was the reason for our attempt to add automatically nominalizations, especially words with the suffixes *-ja* (equal to *-er* suffix in English) and *-mine* (equal to *-ing* suffix in English). In this way, almost 10 000 synsets were added. Unfortunately, very many of these derivations are not valid because both one internal and one external relation were generated automatically – internal with xpos_hypernym relation linked to a verb and external equal_hyperonym relation to a verb. This lead into a confusing situation, because both relations are not accurate and more importantly link only to another part of speech, which does not follow the principles of wordnet. For example, the verb synset 'say, state, tell' got automatically several xpos_hyponyms (all following synset are nouns):

lisamine, täiendamine 'adding'
andmine 'giving'
deklareerimine, kuulutamine 'declearing'
hõikamine, hõiskamine 'whooping'
protestimine 'protesting'
esitamine 'presenting'
kordamine 'repeating'
vastamine 'answering'.

Another problem occurred while transferring these derivations into EstWN – although the verb as a derivation base can have multiple senses, then the derived nouns with *-mine* and *-ja* suffix don't share the same senses – not syntactically and not semantically. For example, the word *andma* 'to give' has 14 senses in EstWN, but derivations *andmine* 'giving' and *andja* 'giver' are used only in some of these 14 senses. The revision of automatic derivations is quite challenging since they also miss definitions. We still deal with these derivations manually – either fix the set of relations and add definitions or delete the invalid concepts completely.

Because of rich Estonian morphology many derivations are possible, like adverbs which are easily derived from other word classes, for example, *ahne* 'greedy' – *ahnelt* 'greedily' (Kerner et al. 2010). However, the described experiments have made us cautious about fully automatic enlargements, since the manual correction is unreasonably time-consuming. Of course, we are open to implementing proven automatic extension methods, which measure up to the quality of manual work.

1.3 How to define synsets – general challenges

It is widely known that definitions are difficult to write and take a lot of time even in one's mother tongue, yet they provide clarity both for native speakers and foreigners (Benjamin 2017). Because a lot of synsets in EstWN are missing definitions, we have to provide them a proper one, if possible. The problem of definitions originates from our existing dictionaries of Estonian – we can find a lot of tautology – an unnecessary repetition of meaning. None of the dictionaries we have used contain information about hierarchical concepts. The explanatory dictionary features information about hypernym (also synonyms, near-synonyms or antonyms) for some headwords in definitions, but this information is, unfortunately, unsystematic and can be rather confusing.

In Estonian, it is possible (and common) to rewrite concepts with compound words, since patterns of compound word formation are productive in Estonian (Kerge 2016). Again, the problem of tautology arises if a synset contains a compound word, for example, *hüpertoonia+haige* 'hypertonia+sick person', *hüpertoonik*, *kõrgvererõhu+haige* – 'person, who suffers from hypertonia'. A good definition is meant to paraphrase the concepts, but tools (i.e words) seem to be missing. Lew (2015) has pointed out, that surprisingly people look up the explanation of meaning firstly through synonyms, so it might be more helpful in some cases to pay attention to synset members rather than to a (bad) definition. Similarly, from the Estonian Text Simplification application (Peedosk 2017) appeared that for the better understanding of a concept it is essential to be able to choose between foreign word and native word (encephalitis *vs. ajupõletik* 'inflammation of the brain' or *kõht* 'belly' *vs. abdoomen* 'abdomen'). Native words are often more informative to native speakers, whereas foreign word is understandable to foreigners (and through the foreign word they are able to learn and understand the native word).

2 EstWN odyssey from ILI1.5 to PWN3.0 and to CILI

Since we wanted EstWN to be linked to the Global WordNet Association repository with Collaborative Interlingual Index (CILI), the first step was to update the old ILI1.5 to the latest PWN3.0 version. As said before, different wordnets are generally similar but still need some effort to combine in a common interoperable multilingual framework (Bond, Piasecki 2017). As follows we describe our efforts and challenges of the CILI-linking process from the wordnet builders point of view.

EstWN was connected to ILI1.5 almost 20 years, and on 2017 we could finally update ILI1.5 to PWN3.0 thanks to our new wordnet editing tool – WordNet WorkBench[2]. The first ILI version (1.5) contained more than 90 000 concepts, yet it was often difficult to determine equal synonyms from Estonian to English. ILI1.5 missed suitable senses, especially regarding adjectives and adverbs. Another problem was that a lot of definitions were missing from ILI and it was complicated to decide the exact meaning of the ILI synsets. PWN 3.0 is of course much richer with different concepts to choose from, so we started to correct English equivalents systematically – changing other ILI-relations into more precise equal synonym relation.

In order to share the data with Open Multilingual Wordnet project, we still have to link EstWN's synsets to CILI, since the reference to CILI is the obligatory attribute of synset.

At the moment in EstWN 22 345 synsets have the external reference with relation type 'eq_synonym' to PWN 3.0 and thereby are mapped to CILI. Number of CILI-links which are not linked with EstWN is 95 314. This number includes also 7556 proper names, connected with PWN via instance-relation. Thus other (approx 65 thousand) synsets require work in order to either find a relation with appropriate concept from CILI or in the future to define a new concept with a new definition and propose them to CILI.

It is also widely known, that some mistakes are inevitable and the solution is the manual correction of errors. Next, we describe the process of improving English part of EstWN through the English equivalents. Since it is complicated and unreasonable to check English equivalents from the first entry in EstWN, we composed different types of lists[3], which we considered to be problematic.

From these lists different types of mistakes occurred, for example, 940 English synsets were connected to 1881 Estonian synsets via the eq_synonym relation, which indicates that these synsets need to be either corrected or united. Some examples:

- Small variations in spelling – like between singular and plural (for example *helilaine(d)* – 'acoustic wave(s)') or spelling error between *diakoniss* and *diakoness* – 'deaconess').
- Indistinguishable senses which are dealt as mistakes and were united to one synset (for example *finaal* 'finale' ja *kooda* 'coda' as music terms; *brie* and *brii* (as Estonian adaption of the name of Brie cheese)).

[2] The tool is freely available, please contact EstWN team for further information. For detail see Jentson et al. forthcoming.

[3] For example, list of eq_has_hyperonym relation with frequency more than 4 times of usage, list of eq_near_synonym with frequency more than 2 times of usage etc.

After the linking process to CILI was completed, then other general types of errors were found from the composed lists, for example:

- Some cases where eq_near_synonym and eq_has_hyperonym have been in confusion, for example, English concept 'folk singer' has 12 near_synonym and 13 has_hyponym in Estonian and therewith *kerjuslaulik* 'beggar singer' being eq_near_synonym to 'folk singer, jongleur, minstrel, poet-singer, troubadour' and *rüütlilaulik* 'troubadour' being linked with eq_hyperonym relation to 'folk singer, jongleur, minstrel, poet-singer, troubadour'.
- 8411 cases, where the Estonian synset has an external link to English concept in the different part of speech, for example, adjective *nunnalik* 'nun-like' is connected via ILI with noun nun. The Estonian word *nunnalik* 'like a nun' is rich with nuances (different across cultures, looks, behavior, attitudes, mentalities) and it is complicated to link this particular Estonian adjective to English adjective. So the only way is to link it to a noun.
- One English synset may have too many hyponyms in EstWN, for example, 'denizen, dweller, habitant, indweller, inhabitant' has 42 hyponyms.
- We counted synsets which use the same eq_near_synonym more than 2 times and we got 347 such. For example, 'district, dominion, territorial dominion, territory' has eq_near_synonym relation 7 times in EstWN.
- Mistranslations: the meaning of the word often depends on context (see e.g Wittgenstein 2005) - English concepts don't fit into Estonian context and vice versa. Lexical caps can be roughly:
    - referential (missing concept, as snow for African people) and
    - lexical (missing word or expression, for example, onomatopoeic words in English and culture-specific words like *kama* (Estonian food made from grain).

As no lexicon can cover all words and senses there are lot's of concepts which are lexicalized in language but haven't found their way to a lexicon or wordnet yet. For example, the Estonian concept *piimasupp*, 'milk soup' in English, which is lexicalized also in English but is missing currently from PWN3.0. Same on the contrary, Estonian synset may have several near_synonym links to English synset, for example *härra, isand, saks* has link of near_synonym to 'landlord' and 'gentleman' and has hyperonym link to 'man of means, rich man, wealthy man' – in Estonian concept, different nuances are mixed from all three English concept. One possible solution is offered by Frankenberg-Garcia (2015) who emphasized that correct translation should be shown with 4-5 examples of usages (i.e to show broader context) or with clear definitions to understand nuances of differences.

The remarks above summarized and discussed only some challenges of our wordnet building, and not the whole project, which is still in progress.

3 Future plans

The EstWN project has most definitely achieved the initial goals of the project and at the end of this NPELT program, there is an appropriate time to set new goals and plan future activities. EstWN project has several quite challenging stages ahead: we continue to increase the size of EstWN with a special focus on the quality. Another direction is to find applications for EstWN – it has been proven for EstWN, that via these applications it is possible to perform different types of quality checks. We have to look more into the topic of the compound words because EstWN is missing some of the mostly used compounds. For compound extraction a corpus will be used, and compounds which occur more than 10 times in this corpus are considered as possible candidates as new concepts or senses.

The new editing tool WordNet WorkBench enables us to create, change or delete semantic relations, so we can create (and rename) new semantic relations valid for Estonian and adopt relations from other resources, for example, domain relation from PWN. Also, we plan to integrate domain labels from WordNetDomains automatically; of course we have to validate if the domains initially created for English apply also in the context of Estonian.

Summing up, we can say that EstWN has reached a level where it can be used in several language technology applications and in research as a valuable language resource.

References

Aller, Sven; Orav, Heili; Vare, Kadri; Zupping, Sirli. (2016). Playing Alias – efficiency for wordnets(s). – *Proceedings of the 8th Global WordNet Confer-*

ence [GWC 2016]: Bucharest, Romania, January 27–30, 2016. Ed. by V. Barbu Mititelu, C. Forascu, C. Fellbaum, P. Vossen. Bucharest: Alexandru Ioan Cuza University of Iași, pp. 16–21; http://jiangbian.me/papers/2016/gwc2016.pdf (15.09.2017).

Baccianella, Stefano; Esuli, Andrea; Fabrizio Sebastiani. 2010. SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. – *LREC 2010 Proceedings*: LREC 2010, Seventh International Conference on Language Resources and Evaluation. http://nmis.isti.cnr.it/sebastiani/Publications/LREC10.pdf (15.09.2017).

Benjamin, Martin. 2017. Inside Baseball: Coverage, Quality, and Culture in the Global WordNet. – *Proceedings of the Workshop on Challenges for Wordnets,* http://ceur-ws.org/Vol-1899/CfWNs_2017_proc9-paper_5.pdf (15.09.2017).

Bentivogli, Luisa; Forner, Pamela; Magnini, Bernardo; Pianta, Emanuele. 2004. Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing. In *COLING 2004 Workshop on "Multilingual Linguistic Resources"*, Geneva, Switzerland, August 28, pp. 101-108.

Bond, Francis: Piasecki, Maciej. 2017. Introduction: Comtemporary Challenges for Development and Application of Wordntes. – *Proceedings of the Workshop on Challenges for Wordnets,* http://ceur-ws.org/Vol-1899/wordnet_preface.pdf (15.09.2017).

Fellbaum, Christiane (ed.) 1998. WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Frankenberg-Garcia, Ana. 2015. Dictionaries and Encoding Examples to Support Language Production. *Oxford University Press International Journal of Lexicography*. Repository URL: http://epubs.surrey.ac.uk/808172/ (15.09.2017).

Jentson, Indrek; Orav, Heili; Vare, Kadri; Kahusk, Neeme. Forthcoming. LiLT Paper on Estonian Wordnet. *Special Issue on Linking, Integrating and Extending Wordnets. Linguistic Issues in Language Technology – LiLT.* Volume 10, Issue 4 Sep 2017.

Kerge, Krista. 2016. Word-formation in the individual European languages: Estonian. – Word-Formation. *An International Handbook of the Languages of Europe.* Ed. by P. O. Müller, I. Ohnheiser, S. Olsen, F. Rainer. Berlin, New York: De Gruyter. (Handbooks of Linguistics and Communication Science ; 40.5), pp. 3228−3259.

Kerner, Kadri; Orav, Heili; Parm, Sirli. 2010. Semantic Relations of Adjectives and Adverbs in Estonian WordNet. − *LREC 2010 Proceedings:* LREC 2010, Malta, Valletta, May 17-23, 2010. ELRA, pp. 33−37.

Lew, Robert. 2015. Dictionaries and Their Users. *International Handbook of Modern Lexis and Lexicography.* Springer-Verlag Berlin Heidelberg.

Lohk, Ahti, Orav, Heili; Vare, Kadri; Võhandu, Leo. 2016. Experiences of lexicographers and computer scientists in validating Estonian Wordnet with test patterns. – *Proceedings of the 8th Global WordNet Conference* [GWC 2016]: Bucharest, Romania, January 27-30, 2016. Ed. by V. Barbu Mititelu, C. Forascu, C. Fellbaum, P. Vossen. Bucharest: Alexandru Ioan Cuza University of Iaşi, 184−191.

Pajupuu, Hille; Altrov, Rene; Pajupuu, Jaan. 2016. Identifying polarity in different text types. − *Folklore. Electronic Journal of Folklore*, 64, 25−42.

Peedosk, Martin. 2017. Applying Estonian Digital Resources and Technologies in a Text. Simplification Program. University of Tartu, BA thesis, https://comserv.cs.ut.ee/ati_thesis/datasheet.php?id=58269&year=2017 (15.09.2017).

Vossen, Piek (ed.) 1998. EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Kluwer, Dordrecht, The Netherlands.

Wittgenstein, Ludwig (2005). Filosoofilised uurimused. Tartu: Ilmamaa.