CEUR-WS.org/Vol-1899/CfWNs_2017_proc5-paper_6.pdf

The Revision History of Estonian Wordnet

Neeme Kahusk and Kadri Vider

Institute of Computer Science
University of Tartu
J. Liivi 2, 50409 Tartu, Estonia

Abstract. Estonian Wordnet started in 1998 and since then at least once a year a new version has been released. The paper gives an overview of the main indicators for every year, as well as highlighting some problems that have been solved or as yet have to be solved.

1 Introduction

There is an old legend about the city of Tallinn, the capital of Estonia, that is situated on the northern coast of the country between the sea and a lake called Ülemiste. Once upon a time, a legend tells, there lived an old gray man, the gremlin of the lake. He used to ask the guards of the city whether the city was finished. Nobody was allowed to answer "yes, it is finished," as then the gremlin would have opened the gates of the waters and flushed the city down into the sea.

This old legend gives us a hint why creators of large and complicated language resources such as dictionaries, lexicons and thesauri, not to mention wordnets, tend to answer about the state of their product "well, under construction," instead of "yes, it is completed".

Whatever the reason, the users of such resources are not willing to wait forever for a completed resource. Half a loaf is better than no bread. Thus, there is a pressure from the users for a (at least intermediate) release of the resource. Should a grant be involved, the creators would be under financial pressure too.

Here we are going to present the twenty year long history of Estonian Wordnet (EstWN), and try to answer the question whether it is ready or not.

The history of EstWN started back in 1998 when the Estonian team joined the EuroWordNet-2 project.[1] Back in that time, the only available database following principles of WordNet structure was Princeton WordNet. The EuroWordNet project followed the same principles, but additionally had the possibility to link each synset in different languages to a central InterLingualIndex (ILI) that was then based on Princeton Wordnet version 1.5. The EuroWordNet database and software are described in the first place by Díez-Orzas, Forest, and Louw (1996).

[1] For overview of EuroWordNet project, see https://www.illc.uva.nl/EuroWordNet/

In construction of EstWN we used the expand method as a starting point. Base concepts from English were translated into Estonian as a first basis for a monolingual extension. The extensions have been compiled manually from Estonian monolingual dictionaries[2] (first of all Explanatory Dictionary of the Estonian Language) and other monolingual resources like terminological collections. In this sense, EstWN applies a hybrid method including both expand and monolingual techniques. EstWN includes nouns, verbs, adjectives and adverbs, and some multiword units among them. The EstWN data is available under a CC-BY-SA license and can be accessed via WordTies[3] and TEKsaurus query tool[4]. For details, see Pedersen et al. 2013.

2 Tools and Work Flow

For EuroWordNet project a new tool was developed. It was made by Novell (member of Groupwise software) and used the Flaim database format. It was a desktop tool: there were neither shared files nor a central server. For exporting and importing data, a plain text format based on GEDCOM was used. This file structure was originally intended for genealogical use (see e.g. Eastman 2014) and dates back to the 1980s, but the basic ideas have lived on in the formats like yaml or pug.

The file is structured hierarchically, consisting of level numbers, fields, and field values. A wordnet of a particular language consists of synsets. At every synset, there were descriptions of synset members, internal relations between synsets, equal relations (the links to ILI) between synsets in different languages, and properties or property values. The file syntax used to export a wordnet is described by Louw (1998).

As no client-server model was used, each collaborator worked individually and had to export her results via this export text file.

The Estonian team did not use any version control software back in these days, and in order to maintain conflict-free development, certain strict rules were applied and some helper scripts were used.

Each member was working on a specific topic, and added only synsets and literals strictly connected to the topic that was assigned to him. However, it is not possible to work in complete isolation and there were frequent discussions on some concept or word. It was not always possible to reach common ground, but most of the arguments were settled (Kerner 2008).

The individual files from each team member were sent to the project manager, who carefully imported each version individually and checked manually whether there were no overlaps in the vocabulary or synsets. There were always some lost and found synsets, relations or synset members after each manual upgrade. To decrease possibilities for such mistakes, some strict rules and some bash scripts

[2] `https://www.eki.ee/EN/dictionaries`

[3] `http://wordties.cst.dk/wordties-estwn/`

[4] `http://www.cl.ut.ee/ressursid/teksaurus/index.php?lang=en`

were implemented in fresh data collected from team members. It was a laborious effort, and as we see in section 3, some inconsistencies still have slipped in.

When a certain stage in development of EstWN was completed, or when it was time to report some results, the individual files were merged, checked for duplicates and inconsistencies, provided with a version number and released.

3 The Curse of Character Encoding

There are 27 letters in the Estonian alphabet. Besides the letters found in US-ASCII, there are a, o, u with umlauts (ä, ö, ü), o with a tilde (õ) and s and z with a caron (š, ž). See EVS 1999 for details about Estonian alphabet and recommended locale settings.[5]

Although best attempts were made to use only one single code page for EstWN, some encoding bugs have sneaked in. Since the Polaris tool was too old to support any Unicode encodings, so the work files had to be in some extended ASCII code page. As the number of lexicographers grew along with the number of computers used, it was practically impossible to keep track of the code pages actually in use. In fact, this problem only emerged around 2010 when we started to convert the export files to UTF-8.

The code pages that were in use are rather similar, there is no problem with the letters with umlauts. The first problems rose with Õ/õ (letter O with the diacritic mark tilde) that shares common code point with Ő/ő (letter O with double acute accent) in Windows 1252 and Windows 1250, respectively. These characters can be hard to separate, depending on the font design. Characters Š/š (letter S with a caron) and Ž/ž (letter Z with a caron) have different coding points in these code pages. If one lexicographer used Windows 1252 and another Windows 1250, after merging of the two files it could happen that synsets containing Š/š and/or Ž/ž were duplicated, one of them having nice literals and the other garbage. That led to synsets that differed in literals when really they should have been the same .

The "umlauts" and "carons" were not the only causes of trouble. In the process of creating the lexicon, lexicographers at times copied parts of text from other sources. The copied parts of text occasionally included typographic characters (quoting marks, em and/or en dash), mathematical symbols (° degree symbol), foreign characters from other languages (å, ø).

Finally we included only those characters that are in the extended Estonian alphabet. By doing this we lost some information, mostly foreign characters and typographic symbols, but the conversion was possible without any encoding errors.

[5] 27 is official number of characters in Estonian alphabet. This number does not include some letters found in US-ASCII: c, q, w, x, y. With all these included we will get the Estonian extended alphabet that contains 32 letters. If we will omit all "foreign" letters, we will get the number 23. This number includes the umlauts and õ; in addition to earlier mentioned letters, "carons" and f are excluded too.

4 The History of the EstWN

The numeration of EstWN releases has been very simple: the number of each subsequent version is incremented by one. The last number is 73, but this common ground does not mean, that we have exactly 73 releases of EstWN, and that all of them are numbered in this manner. There are 69 versions of EstWN at the Gitlab code repository[6] that is hosted at the Center of Estonian Language Resources (CELR)[7]. Most of them have version numbers, but two of them are identified by date. Those two we denoted as derivations of kb53 in Table 1.

Table 1. Some indicators for EstWN versions, one from each year

Version	Year	Synsets	Senses	Relations	ILI relations	Synsets without relations	Synsets without ILI relations	Synsets with many hyperonyms
kb01	1998	1,570	3,277	0	1,426	1,570	244	0
kb13	1998	3,726	6,626	5,584	4,000	1,068	259	115
kb29	1999	8,578	15,553	17,832	9,985	134	1	201
kb35	2000	9,347	16,849	19,634	10,819	154	1	245
kb39	2001	9,798	17,493	20,683	11,329	151	1	277
kb42	2002	10,439	18,192	22,017	11,558	162	1	279
kb47	2003	10,892	18,851	22,982	12,044	161	2	284
kb48	2004	10,893	18,855	22,984	12,044	161	3	284
kb51	2005	15,387	33,517	23,620	12,084	4,426	4,477	22
kb53	2006	15,386	33,514	24,802	16,522	4,025	4,073	58
kb53-1	2007	15,952	34,227	26,218	18,358	4,025	4,088	91
kb53-2	2008	17,544	36,404	30,836	20,786	3,967	4,089	243
kb55	2009	26,781	48,292	55,260	35,217	3,541	2,167	811
kb59	2010	42,341	72,579	99,014	58,787	1,730	5,360	1,267
kb62	2011	49,514	82,568	133,062	73,397	836	2,141	1,700
kb65	2012	56,929	93,914	163,785	84,146	518	1,804	1,717
kb69	2013	65,518	107,544	203,070	96,205	60	1,258	102
kb70	2014	67,676	110,881	210,424	98,710	47	1,231	14
kb72	2015	74,720	120,911	239,719	108,202	5	9	100
kb73	2016	77,878	125,646	248,996	112,283	0	0	70

For each new version of EstWN the summary of statistical measures has been made.

For the present overview we made an excerpt of the table, including the very first version (kb01 released on June 5, 1998), and thereafter the last version from each year. The date of the "last" version of the year varies a lot, from August to mid-December. The number of versions per year varies also. There were 13 releases in 1998, in 2006–2008 there was only one release per year.

[6] https://gitlab.keeleressursid.ee/nemee/Estwn

[7] https://keeleressursid.ee/en/

The capacity of a wordnet can be measured by various indicators. The most prominent ones are the number of synsets, and the number of lexical items. A more precise indicator is number of synsets by part-of-speech. As a wordnet is also about semantic (or lexical, or lexico-semantic) relations, the number of relations is informative as well. In order to link to other lexicons, external references are important. In EuroWordNet project the equal relations or ILI relations served this purpose .

For EuroWordNet, the number of the Base Concepts (see Vossen et al. 1998) served as one of the indicators. It is not listed in this paper, but according to our raw data the Base Concepts were included in EstWN kb14 (first release of 1999) already, and their number has remained greater than 2,300 since then.

Following the running statistics has influenced the process of extending EstWN. This can give insight to features, in which the data is missing. One of the aims to complement EstWN has been adding semantic relations, the decreasing number of synsets without semantic relations is a good indicator of this process.

The last row in Table 1 entitled “Synsets with many hyperonyms” lists the number of synsets that have more than one hyperonym. Although more than one hyperonym per synset is a common practice in wordnets, their existence may cause anomalies in the wordnet hierarchy, as Lohk 2015 has shown.

Better overview about the dynamics of the number of synsets by part of speech is given in Figure 1. The chart shows the number of noun, verb, adjective and adverb synsets cumulatively. Instead of a linear growth there is a S-shaped curve, which has a plateu at the beginning of the Millenium and a sharp burst in 2010.

According to Figure 1 there is a increase in number of synsets from 2004 to 2005, but at the same time the number of synsets without relations, both internal and ILI, has changes too. This may be the co-effect of automatic addition of new synsets via derivations. For some derivations it was possible to make semantic relations in the same manner, but not for all of them. The method and problems are described by Kahusk, Kerner, and Vider (2010).

As for kb73, currently the the last version, the number of synsets without relations is zero. EstWN kb73 will probably remain the last version to be linked to Princeton WordNet ver. 1.5, the next version will be linked to to 3.0 at least, and perhaps we will use a new numbering scheme as well. Kb 73 is also browsable by the current version of TEKsaurus tool (Kahusk and Vider 2005). A slightly older version of EstWN has been included into EstNLTK toolkit (Orasmaa et al. 2016) along with the Eurown module for Python (Kahusk 2010).

As EstWN is compiled according to Estonian language, the meaning of words in Estonian is considered instead of translating synsets from English wordnet. This practice has lead to a rather small number of eq_synonym relations in EstWN (see Figure 2). There was a rather broad spectrum of equal relations used in EuroWordNet project, 15 equal relation types (besides eq_synonym) are used in EstWN. That explains the relatively high number of other ILI relations on Figure 2.

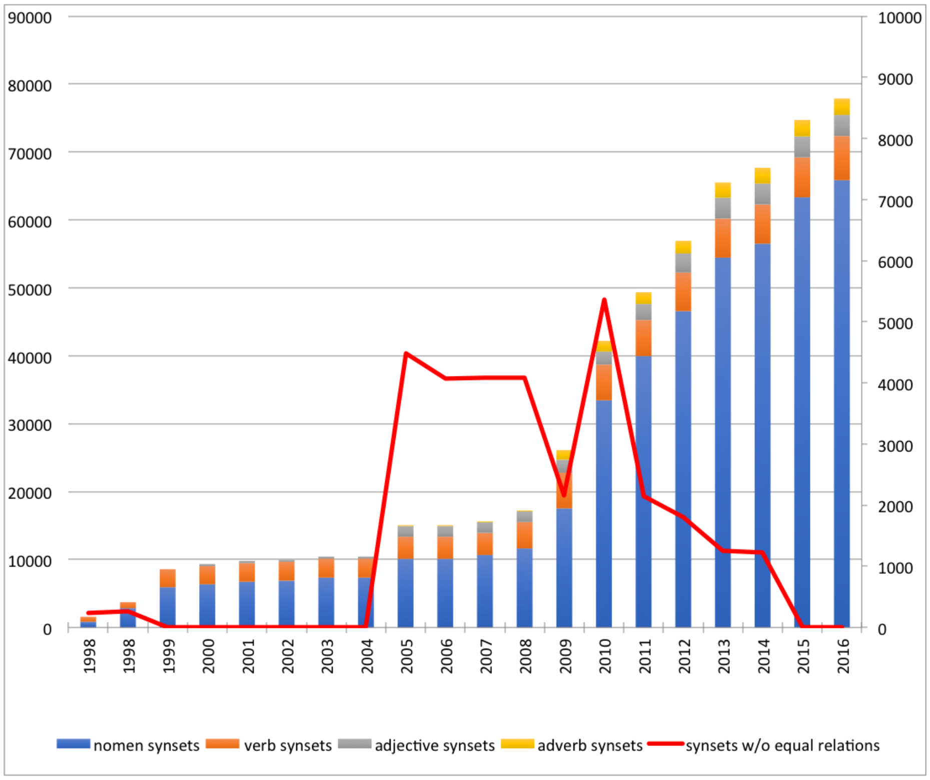

Fig. 1. The number of synsets by part of speech in Estonian wordnet versions and the number of synsets without ILI relations.

The most numerous of the other ILI relations are eq_has_hyperonym and eq_near_synonym. Both of them are relations that are used to cope different granularity and coverage of synsets in Estonian and many other languages.

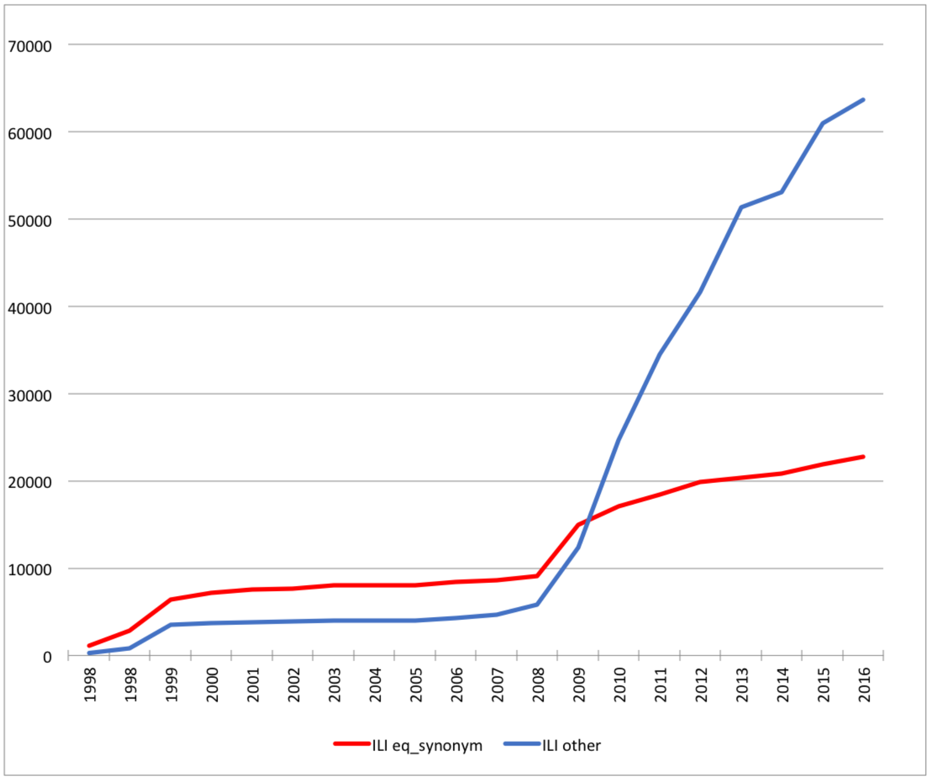

Fig. 2. The number of eq_synonym and other ILI relations in different versions of Estonian Wordnet.

5 Saving for Future

There are five versions of EstWN (see table 2) described in the CELR language resources metadata registry META-SHARE[8]. They have been viewed more than 600 times, more than 330 views imputed to the oldest version. As for downloads, the number is 113, the latest version has been downloaded most often.

[8] https://metashare.ut.ee

Table 2. EstWN versions in META-SHARE

Version	DOI	Views	Downloads
65	`https://doi.org/10.15155/1-00-0000-0000-0000-00084L`	333	37
69	`https://doi.org/10.15155/1-00-0000-0000-0000-00072L`	95	13
71	`https://doi.org/10.15155/1-00-0000-0000-0000-00087L`	73	10
72	`https://doi.org/10.15155/1-00-0000-0000-0000-000E9L`	61	7
73	`https://doi.org/10.15155/1-00-0000-0000-0000-0011DL`	89	46

The metadata from our repository is harvested by CLARIN[9] to the Virtual Language Observatory (VLO)[10], and META-SHARE[11]. As for the last registry, search for "wordnet" gives 74 answers[12], if Estonian is specified as the language, 7 of them will remain in results. In VLO 's case, the search for "wordnet" gives 106 answers, after limiting language to Estonian, 10 of the results remain. Although the numbers are not very large, it may take for a while to find out the last version. Probably version 73 will not remain last version of the EstWN, and every next version will add some more confusion.

We may assume, that generally the user would be interested in finding the last version of the wordnet, but it may not always be so. A specific version of the wordnet or several (more than one) legacy versions may be needed in order to replicate a particular experiment or to develop a resource. If we poured all the more than 60 legacy versions of EstWN into our repository, then it would overflow not only our metadata registry, but CLARIN VLO and www.meta-share.org too.

An elegant solution for this problem would involve an elaborated CMDI profile with references to older and newer versions of the resource.[13] However, an upgrade of harvesting sites would be required as well. This would probably work for VLO, but not for www.meta-share.org or other harvesting services (our registry is harvested by Estonian E-Repository too).[14]

As a compromise, we are going to describe metadata of all the legacy versions as single collection of resources and upload it to our repository.

6 Acknowledgements

Development of EstWN is supported by the National Programme for Estonian Language Technology during 2006–2017.[15]

[9] `https://www.clarin.eu/`

[10] `https://vlo.clarin.eu/`

[11] `http://www.meta-share.org`

[12] The search site redirects to `http://metashare.ilsp.gr:8080/`

[13] `https://www.clarin.eu/cmdi1.2`

[14] `https://www.e-varamu.ee`

[15] `https://www.keeletehnoloogia.ee/en?set_language=en`

References

Díez-Orzas, Pedro, Philippe Forest, and Michael Louw (1996). *High-level Architecture of the EuroWordNet Database. A Novell ConceptNet-based semantic network.* Tech. rep. URL: `https://www.illc.uva.nl/EuroWordNet/docs.html`.

Eastman, Dick (2014). "GEDCOM Explained". In: *Eastman's Online Genealogy Newsletter.* URL: `https://blog.eogn.com/2014/05/24/gedcom-explained/`.

EVS (1999). *Requirements on information technology in estonian language and cultural environment. Unofficial final draft.* URL: `http://www.eki.ee/itstandard/2000/contents.html`.

Kahusk, Neeme (2010). "Eurown: an eurowordnet module for python". In: *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)*, pp. 360–364.

Kahusk, Neeme, Kadri Kerner, and Kadri Vider (2010). "Enriching Estonian WordNet with Derivations and Semantic Relations". In: *Frontiers in Artificial Intelligence and Applications* 219.Human Language Technologies—The Baltic Perspective. Ed. by Inguna Skadiņa and Andreijs Vasiljevs, pp. 195–200. ISSN: 0922-6389. DOI: `10.3233/978-1-60750-641-6-195`. URL: `http://doi.org/10.3233/978-1-60750-641-6-195`.

Kahusk, Neeme and Kadri Vider (2005). "TEKsaurus—the Estonian WordNet online". In: *The Second Baltic Conference on Human Language Technologies: proceedings.* Ed. by Margit Langemets and Priit Penjam. Tallinn: Institute of Cybernetics.

Kerner, Kadri (2008). "Proposing Some Methods of Improving Word Sense Disambiguation for Estonian Language". In: *Proceedings of the Fourth Global WordNet Conference: Proceedings of the Fourth Global WordNet Conference.* Hungary, Szeged: University of Szeged, Department of Informatics, pp. 229–239.

Lohk, Ahti (2015). "A System of Test Patterns to Check and Validate Wordnet-type Dictionaries". PhD thesis. Tallinn University of Technology. URL: `http://digi.lib.ttu.ee/i/?3161`.

Louw, Michael (1998). *Polaris User's Guide. The EuroWordNet Database Editor.* Tech. rep. URL: `https://www.illc.uva.nl/EuroWordNet/docs.html`.

Orasmaa, Siim et al. (2016). "EstNLTK—NLP Toolkit for Estonian". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).* Ed. by Nicoletta Calzolari et al. Paris, France: European Language Resources Association (ELRA). ISBN: 978-2-9517408-9-1.

Pedersen, S. Bolette et al. (2013). "Nordic and Baltic wordnets aligned and compared through "WordTies"". In: *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013).* Ed. by Stephan Oepen, Kristin Hagen, and Janne Bondi Johannesse. Linköping University Electronic Press, pp. 147–162. URL: `http://www.ep.liu.se/ecp/085/ecp13085.pdf`.

Vossen, Piek et al. (1998). *Set of Common Base Concepts in EuroWordNet-2*. Tech. rep. URL: `https://www.illc.uva.nl/EuroWordNet/docs.html`.