First steps in checking and comparing Princeton WordNet and Estonian Wordnet

Ahti Lohk
Tallinn University of Technology
Raja 15-117
Tallinn, ESTONIA
[ahti.lohk@ttu.ee]

Kadri Vare
University of Tartu
Liivi 2-308
Tartu, ESTONIA
[kadri.vare@ut.ee]

Leo Võhandu
Tallinn University of Technology
Raja 15-117
Tallinn, ESTONIA
[leov@staff.ttu.ee]

Abstract

Each expanding and developing system requires some feedback to evaluate the normal trends of the system and also the unsystematic steps. In this paper two lexical-semantic databases – Princeton WordNet (PrWN) and Estonian Wordnet (EstWN)- are being examined from the visualization point of view. The visualization method is described and the aim is to find and to point to possible problems of synsets and their semantic relations.

1 Introduction

Wordnets for different languages have been created for a quite a long time [1]; also these wordnets have been developed further and updated with new information. Typically there is a special software for editing wordnets, for example VisDic[2], WordnetLoom (Piasecki et al 2010), Polaris (Louw, 1998). These editing tools often present only one kind of view of the data which might not be enough for feedback or for detecting problematic synsets/semantic relations. The visualization method described here can be used separately from the editing tool; therefore it provides an additional view to data present in wordnet.

For initial data PrWN version 3.0[3] and EstWN version 63[4] have been taken. PRWN contains of 117 374 synsets and EstWn of 51 688 synsets. The creation of EstWN started in 1998 within the EuroWordNet project[5]. At present the main goal is to increase EstWN with new concepts and enrich EstWN with different kinds of semantic relations. But at the same time it is necessary to check and correct the concepts already present (Kerner, 2010).

The main idea and basic design of all wordnets in the project came from Princeton WordNet (more in Miller et al 1990). Each wordnet is structured along the same lines: synonyms (sharing the same meaning) are grouped into synonym sets (synsets). Synsets are connected to each other by semantic relations, like hyperonymy (is-a) and meronymy (is-part-of). As objects of analysis only noun synsets and hyperonymy-hyponymy relations are considered (of course, it is possible to extend the analysis over different word classes and different semantic relations). So, due to these constraints we have taken 82 115 synsets from PRWN (149 309 different words in synsets) and 41 938 synsets from EstWN (64 747 different words in synsets).

2 Method

We will explain our method's main idea with a small artificial example. Let us have a small separated subset presented as a matrix:

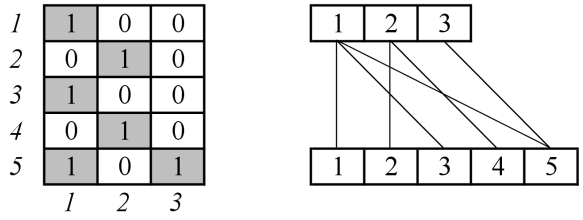


Figure 1. Relation-matrix and bipartite graph

In the rows of that table we have synsets and in columns hyperonyms. On the right side of

[1] http://www.globalwordnet.org/
[2] http://deb.fi.muni.cz/clients-debvisdic.php
[3] http://wordnet.princeton.edu/
[4] http://www.cl.ut.ee/ressursid/teksaurus/
[5] http://www.illc.uva.nl/EuroWordNet/

Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH, pages 25–29,
Avignon, France, April 23 - 24 2012.

that figure we have presented the same data as a bipartite graph where all column numbers are positioned on the upper line and all rows on the lower line. Every connecting line on the right side has been drawn between every "1"-s column and row number. As we see a lot of line crossings there exist even in our very small example. It is possible to reorder the rows and columns of that table into optimal positions so that the number of line crossings would be minimal possible. If there is full order then there will be no crossings of lines.

Generally this crossing number minimization is a NP-complete task. We are using the idea of Stephan Niermann's (2005) evolutionary algorithm to minimize the number of line crossings.

In our example the optimal result will be:

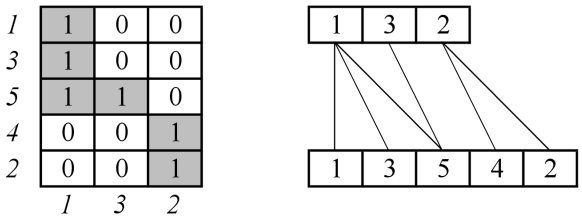


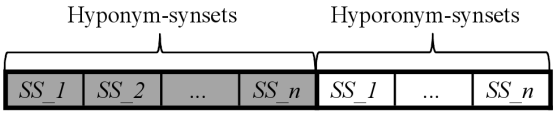
Figure 2. Reordered (arranged) relation-matrix and bipartite graph

As we can see there are no crossings and all connections are separated into two classes – let's call them closed sets. We have got a nice and natural ordering for rows and columns. With that kind of picture the relations between words (synsets) are easier to see and understand. We will present real cases from PrWN and EstWN later.

3 Practical application of the method

Next we will describe the steps that should be taken in order to obtain visual pictures for lexicographers.

- First the word class and a semantic relation of interest is chosen from wordnet. For nouns and verbs hyperonymy and hyponymy are probably the most informative relations, for adjectives and adverbs near_synonymy (but of course this method allows us to choose different semantic relations in combination with different word classes).
- In order to find closed sets we use the connected component separating algorithm for graphs given in D. Knuth (1968). For example using hyponym-hyperonym relation and word classes of nouns then there will be 7 907 closed sets for EstWN and 15 452 closed sets for PrWN. Every closed set is presented in a table as a row with different lengths. An arbitrary closed set is similar to the following picture in Figure 3.



SS1 - synset 1, SS2 - synset 2, ...

Figure 3. Example of a closed set

- As a next step we use all connections for those two sets in a wordnet to get the relation matrix as it is shown in Figure 1 left part.
- Then the minimal crossing algorithm is used (result is seen on the right side of Figure 2).
- As the last step a lexicographer analyzes the figures.

It is still important to mention that our approach is not quite useful for analyzing the large closed sets. The reason is that in Nierman's evolutionary algorithm if the size of the matrix grows than the time increases with the speed $O(n^2)$. For example, to solve the 30x30 matrix, it takes 3 minutes and to solve 60x60 matrix, it takes 60 minutes. That is the reason why in this paper only closed sets that do not exceed the 30 hyponym sets are considered. The pictures from closed sets (Figure 4, 5, 6) were solved as follows: Figure 4 (3 x 5 matrix) 0,28sec, Figure 5 (4 x 11 matrix) 1,5sec, Figure 6 (4 x 12 matrix) 1,7sec.

For larger closed sets it is better to use the modified Power Iteration Clustering method by Lin and Cohen (2010) instead of Niermann's algorithm.

As a matter of fact, the largest closed set in EstWN has 4103 hyponyms-synsets x 405 hyperonym-synsets and the largest closed set in PrWN has 2371 hyponyms-synsets x 167 hyperonym-synsets (Figure 3). As for large closed sets, it could be sensible to use only the relation matrix (Figure 2, left side) to detect where possible problematic places occur.

4 Intermediate results

In this paper we focus on the synsets having two or more hyperonyms, which is the reason of closed sets, since it is more likely to find problematic places in these synsets.

For example in EstWN only one hyperonym for a synset should ideally exist (Vider, 2001). In EstWN there are currently 1 674 concepts with two hyperonyms, 145 concepts with three or more hyperonyms and the concept which has the most hyperonyms - 9 - is ’alkydcolour’.

In PrWN there are 1 442 concepts with two hyperonyms, 34 concepts with three or more hyperonyms and the concept with the most hyperonyms – 5 – is 'atropine'.

Of course in wordnets a synset can have multiple hyperonyms in many cases, in EstWN many of the onomatopoetic words, for example (typically they have hyperonyms which denote movement and sound). But also there are cases where one of the hyperonyms is in some ways more suitable than another. Even if a synset has multiple hyperonyms a cluster still often presents a homogeneous semantic field.

One of the purposes of the visual pictures is to help in detecting so called human errors, for example:

- in a situation where in the lexicographic (manual) work a new and more precise hyperonym is added during editing process but the old one is not deleted;
- lexicographer could not decide which hyperonym fits better;
- lexicographer has connected completely wrong senses (or words) with hyperonymy relation;
- lexicographer has not properly completed the domain-specific synsets etc.

The first three points can indicate the reason of why one synset has multiple hyperonym-synsets.

For example, in Figure 4 all the members of the cluster seem to form a typical set of allergic and hypersensitivity conditions and illnesses. In EstWN currently allergies and diseases caused by allergies do not form such a cluster, because they do not share hyperonyms. But also different clusters exist where some problems can appear.

For example, in Figure 5 where all the other characters (suicide bomber, terrorist, spy etc) except ‘programmer’ are bad or criminal by their nature. This leads to a thought that maybe ‘programmer’ as a hyperonym to ‘hacker’ and ‘cracker’ is not the best; it might be that ‘programmer’ is connected with some other semantic relation.

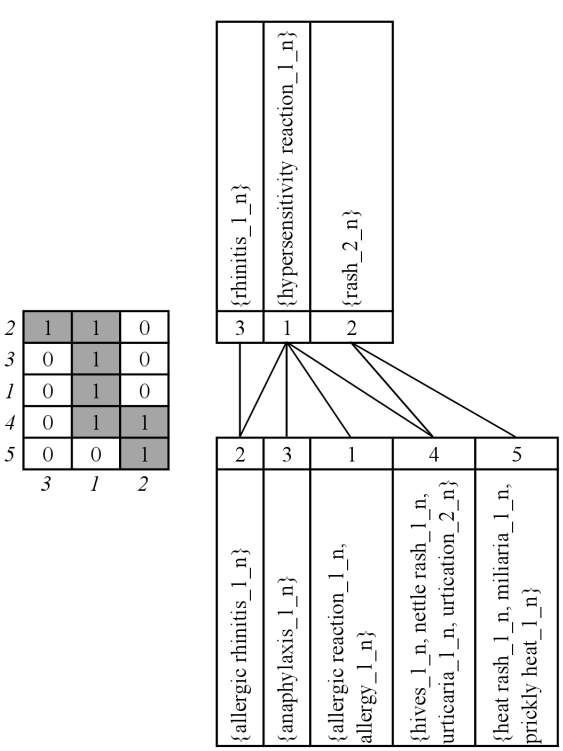


Figure 4. Rearranged bipartite graph, PrWN

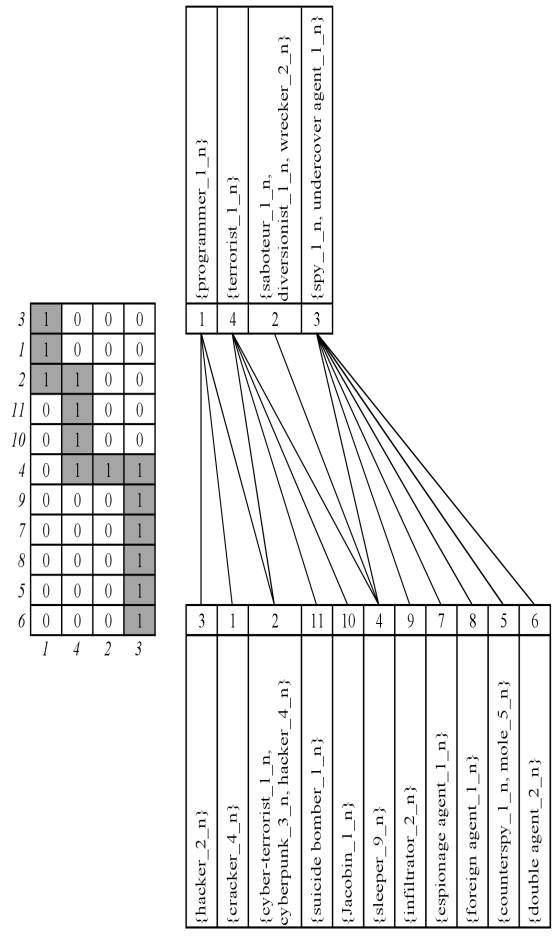


Figure 5. Rearranged bipartite graph, PrWN

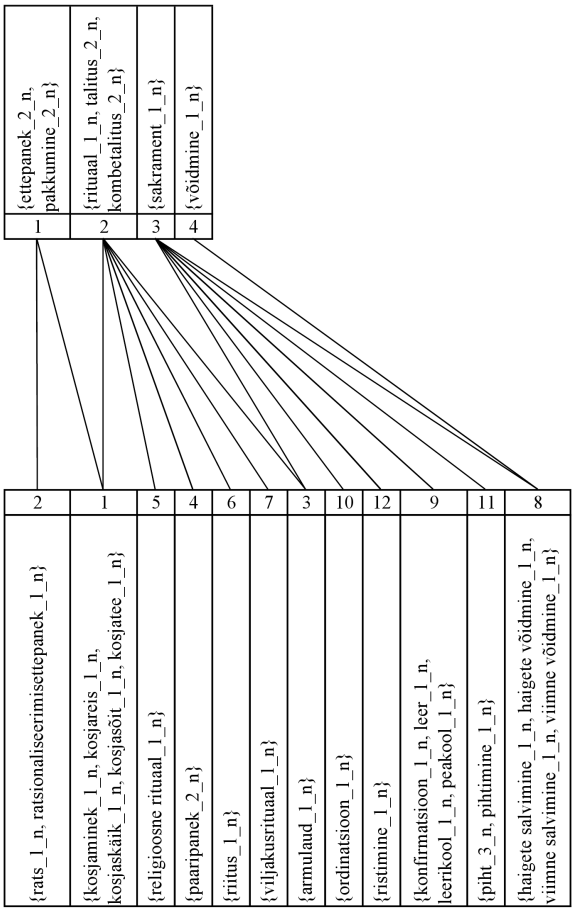


Figure 6. Rearranged bipartite graph, EstWN

Hyperonym-synsets:
1. ettepanek, pakkumine - proposal
2. rituaal, talitus, ... - rituaal
3. sakrament - sacrament
4. võidmine - unction, anoiting

Hyponym-synsets:
4. paaripanek - marriage ritual
6. riitus - rite
7. viljakusrituaal - fertility rite
3. armulaud - Holy Communion
10. ordinatsioon - ordination
12. ristimine - baptism
9. konformatsioon, ... - confirmation
11. piht, pihtimine - confession
8. haigete salvimine, ... - extreme unction
2. rats, ratsionaliseerimisettepanek - proposal for rationalization
1. kosjaminek, kosjareis, ... - a visit to bride's house to make a marriage proposal
5. religioosne rituaal - religious ritual

From EstWN many problematic synsets and/or semantic relations were discovered by using this method. In Figure 6, for example, from EstWN there is an example of a closed set for nouns. It can be seen that the word *ratsionaliseerimisettepanek* (’proposal to rationalization’) does not belong to this semantic field (this semantic field can be named ‘different kinds of rituals’ for example). It is strange that words *ratsionaliseerimisettepanek* (‘proposal to rationalization’) and *kosjakäik* (‘a visit to bride’s house to make a marriage proposal’) belong to the same closed set. Both these synsets share a hyperonym *ettepanek* (‘proposal’), but *kosjakäik* should be connected to *ettepanek* (‘proposal’) by is_involved relation and the hyperonym to *kosjakäik* should be ‘ritual’ instead.

Also the relation of hyperonyms *võidmine* ('unction') and *sakrament* ('sacrament'). should be interesting. It can be seen that all the semantic relations of hyperonym *võidmine* ('unction') belong actually to *sakrament* ('sacrament'). So it is possible to state that sacrament should be hyperonym to unction. Another question arises with the word *armulaud* ('Holy Communion'). In principle, this word is correctly connected to both sacrament and ritual, but still – all of the hyponyms of sacrament are some sorts of services. These connections are probably missing from the system.

In addition, a minor detail – although *abielu* ('marriage') belongs to sacrament, it is in EstWN categorized only as a ritual and not even directly but implicitly by the word *paaripanek* ('marriage ritual')

5 Conclusion

In order to find mistakes from closed sets it is not necessary to use a bipartite graph. In some cases only the relation-matrix will be enough (Figure 1,2 left side). Clear created groupings can be considered as an advantage of bipartite graphs, which present the hyponym synsets connecting the hyperonym synsets. Often these connections can turn out as the problematic ones. Sometimes it is necessary to use the wordnet database in order to move a level up to understand the meaning of a synset.

Out of the 20 arbitrarily extracted closed sets 6 seemed to have some problems. And in PrWN there were 185 closed sets with hyperonym synsets having at least three hyperonyms. This seems to be a promising start towards using visual pictures. The situation is similar in EstWN, and since EstWN is far from “being completed” then this method has already

proven useful for lexicographers in the revision work.

To conclude, the structured bipartite figures are informative in following ways:

- It is possible to use different kinds of semantic relations to create closed sets.
- It is possible to detect subgroups.
- It is possible to detect wrong and missing semantic relations.

Acknowledgments

In this paper Kadri Vare is supported by META-NORD project (CIP-ICT-PSP.2010-4 Theme 6: Multilingual Web: Machine translation for the multilingual web); Estonian Ministry of Education and Research (Target financed research theme SF0180078s08, "Development and implementation of formalisms and efficient algorithms of natural language processing for the Estonian language") and National Programme for Estonian Language Technology.

References

Ashok K. Chandra, Dexter C. Kozen, and Larry J.Stockmeyer. 1981. Alternation. Journal of the Association for Computing Machinery, 28(1):114-133.

Association for Computing Machinery. 1983. Computing Reviews, 24(11):503-512.

Dan Gusfield. 1997. Algorithms on Strings, Trees and Sequences. Cambridge University Press, Cambridge, UK.

Donald E. Knuth. 1968, *Fundamental Algorithms, vol. 1 of Art of Computer Programming* (Reading, MA, Addison-Wesley), §2.3.3.

Frank Lin and William W. Cohen. 2010. *Power Iteration Clustering* in ICML-2010.

George Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross and Kathrine Miller. 1990. *Introduction to WordNet: An On-line Lexical database.* – International Journal of Lexicography 3, 235-312.

Kadri Kerner, Heili Orav and Sirli Parm. 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.) Bhattacha-ryya, P.; Fellbaum, Ch.; Vossen, P. Mumbai, India: Narosa Publishing House, pp 198-202.

Kadri Vider. 2001. Eesti keele tesaurus - teooria ja tegelikkus Leksikograafiaseminar *"Sõna tänapäeva maailmas" Leksikografinen seminaari "Sanat nykymaailmassa". Ettekannete kogumik*. Toim. M. Langemets. Eesti Keele Instituudi toimetised 9. Tallinn, lk 134-156.

Michael Louw. 1998. *Polaris User's Guide*. Technical report, Lernout & Hauspie . Antwerp, Belgium.

Maciej Piasecki, Michal Marcinczuk, Adam Musial, Radoslav Ramocki and Marek Maziarz. 2010. *WordnetLoom: a Graph-based Visual Wordnet Development Framework*. In Proceedings of IMCSIT, 469-476.

Stefan Niermann. 2005. Optimizing the Ordering of Tables With Evolutionary Computation. *The American Statistician*, 59(1):41-46.