

EXTENDED CLUSTERS OF VERTICAL POLYSEMY: AN EXPLORATIVE STUDY OF ELEVEN WORDNETS

Ahti Lohk, Ene Vainik,
Geda Paulsen, Martin Rebane,
Francis Bond

Abstract. The purpose of this study is to identify and examine maximum-size polysemy clusters in semantic hierarchies of eleven wordnets that are included in Open Multilingual Wordnet. The clusters are extracted from both the super-subordinate hierarchies of the noun (IS-A) and the hierarchies of the verb (MANNER-OF). We focus on the substructures of semantic hierarchies that display vertical polysemy relationships (i.e. clusters containing words with the same spelling in their vertices). Besides the maximum-size polysemy clusters, the number of vertical polysemy relationships and the PoS distribution in these relations are examined. The findings of this study suggest that large polysemy clusters may indicate potential problems like inaccurate identification of a word's meaning in an NLP task. The algorithm detecting maximal clusters of polysemy has the potential to be applied to other wordnets than those focused on in this study. The analysis in this article is the first step in large-size polysemy clusters studies. We expect that polysemy clusters will help to evaluate the state of the wordnets' semantic hierarchies and to assess the suitability of these structures as background knowledge for solving NLP tasks.*

Keywords: vertical polysemy, polysemy patterns, semantic hierarchies, wordnet

1. Introduction

Every technical resource that attempts to model the semantic structures represented by natural language must deal with polysemy. The system of handling polysemy is an important aspect to assess while evaluating the quality of a lexical resource.

* This work was supported by Estonian Research Council grant PSG227. We are thankful to the anonymous referees for their valuable comments on this article.

For example, it is essential for the compilers of a lexical resource to search for the potential errors and internal inhomogeneity in order to improve the quality of a resource (Lohk 2015).

One of such resources is the lexical database Wordnet (Fellbaum 1998). The first wordnet (Princeton WordNet) was released in 1991. According to the website of the Global Wordnet Association, there are wordnet applications available for more than 60 languages now.¹ Wordnet enjoys several advantages, which originate from its specific design: it is a machine-readable dictionary with definitions and examples of concepts, and, at the same time, it is a network of concepts connected by semantic relations (Fellbaum 1998). It enables one to figure out the semantic distance of certain concepts (in terms of word senses) in the network. Similarly, a search for “vertical” and “horizontal” neighbors is possible (i.e. sub- and super-concepts and synonyms, respectively). Such lexical-semantic background knowledge is a valuable base for solving various NLP tasks, in particular for tasks that require semantic analysis (e.g. word sense disambiguation).

However, a wordnet is not a hundred-percent flawless solution for defining and describing the semantic network of senses and their relationships. One of the problems that can complicate the use of wordnets is lexical polysemy that occurs in its semantic hierarchies (Freihat et al. 2013, Mihalcea 2003). For example, identifying a lemma as being situated on multiple levels of semantic hierarchy would be misleading for the purpose of NLP tasks where sense disambiguation is needed. The meaning of two words with the same spelling can be indistinguishable if they have a near surrounding in the semantic hierarchy of the wordnet. The problem is even more acute when the concentration of words with the same spelling in the semantic hierarchy of the wordnet is high or even very high (e.g., 5–20). Such a situation may occur in cases where a polysemous word is its own hypernym (as e.g. *dog* meaning ‘canine’ and in addition ‘male canine’, a more specific sense contrasting with *bitch*). This phenomenon is defined as auto-hyponymy, auto-troponymy or vertical polysemy (Koskela 2011, Fellbaum 2002). In this article, we use the term vertical polysemy (for discussion on the terminology, see Cruse 2000, Koskela 2005).

The vertical polysemy relationships in the semantic hierarchies of wordnets have been studied mostly from the viewpoint of searching for a criterion for grouping meanings (Pociello et al. 2011, Pedersen 2018), and with the purpose to reduce polysemy (i.e. reducing the number of terms for their coarser distinction) (Mihalcea 2003). The general specifics of vertical polysemy structures are described also in (Lohk et al. 2019).

In this article, we will study the possible substructures of semantic hierarchies in different wordnets that can only be formed by vertical polysemy. Hereby, our purpose is to find the maximal-size polysemy clusters in eleven wordnets and visualize the largest ones. We expect that larger polysemy clusters will allow us to evaluate the state of semantic hierarchies in terms of polysemy and the suitability of those clusters as background knowledge for solving NLP tasks. We also mathematically formulate a polysemy cluster detection algorithm and analyze the distribution of PoSs forming the largest polysemy patterns.

For the detection of potentially too large or too complex structures formed by vertical polysemy, we apply an algorithm onto eleven different wordnets. In order

to reveal information about their condition in respect to the clusters of vertical polysemy, we utilize the advantage that the wordnets of different languages published on the Open Multilingual Wordnet (OMW) (Bond, Paik 2013, Bond, Foster 2013) website share the same format for presenting the data.

This explorative analysis is the first step in maximum-size polysemy cluster studies. The clusters found in this article are especially beneficial for wordnet developers and users, enabling them to discover the polysemy clusters caused by vertical polysemy.

The article is organized as follows: In Section 2, we give the background to the study. The algorithm is described in Section 3. The maximum-size polysemy clusters retrieved by the algorithm, the types and amount of vertical polysemy patterns, and the PoS distribution in these polysemic relations will be discussed in Section 4. The results are discussed, and the article is concluded, in Section 5. In the Appendix, we present the largest visualized polysemy clusters found by the algorithm in wordnets for German, Estonian, Finnish, Irish, Chinese, Japanese, Dutch, Portuguese, Polish, English, and Turkish.

2. The basics of the theory

No handbook of (lexical) semantics can avoid the notions of polysemy, homonymy and synonymy (Lyons 1995, Cruse 2000, Saeed 2001, Pajusalu 2009). The terms designate different combinations of word forms with their meanings in terms of one-to-many/many-to-one relationships. A single word form expressing several related meanings is a case of polysemy (Fellbaum 2015). It is a case of multiple reference: one and the same linguistic form refers to multiple concepts. As such, it is a classical one-to-many relationship. The meanings must be semantically related (e.g. *table* ‘a piece of furniture’ and ‘a set of facts or figures displayed’). It is claimed that there is twice as much polysemy (at the lexical level) among verbs as among nouns (Fellbaum 1998). The situation of classical polysemy is depicted in Figure 1.

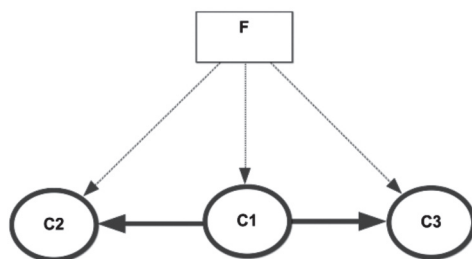


Figure 1. Classical polysemy: one linguistic form (F) refers to different but related meanings/concepts (C)

Polysemy used to be strictly distinguished from homonymy, which is a similar situation, except that the meanings are not expected to be related (e.g. *bank* ‘river bank’ and ‘financial institution’). In recent years, however, the distinction line between polysemy and homonymy is no longer drawn. Even the two meanings of *bank* have been traced back to the same etymological root (Fellbaum 2015). Both polysemy and homonymy are instances of a broader category of lexical ambiguity.

The central notion of this study – vertical polysemy – can only be explained in relation to hierarchically organized knowledge systems like wordnets. A wordnet is a hierarchical network of concepts paired with their linguistic expressions (Figure 2). A single concept is not only linked to other concepts by specific semantic relations (such as hyponymy, meronymy etc.) but is also linked to a (set of) lemma(s) by a general relation of reference. This set of lemmas is called a synset.

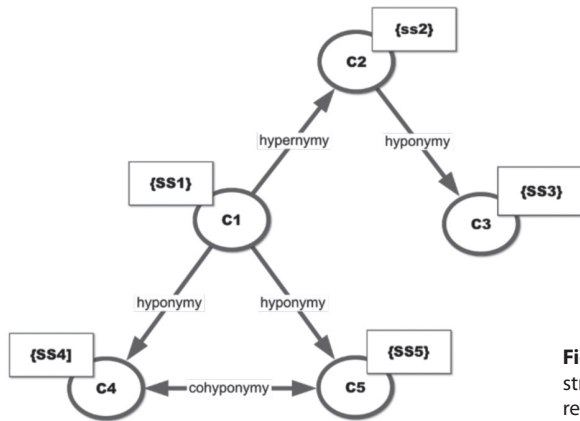


Figure 2. A fragment of the general structure of wordnet (ovals = concepts, rectangles = synsets)

Vertical polysemy is a situation where the same word form refers to concepts that occupy positions on different levels of the hierarchy (Figure 3). Vertical polysemy is also a one-to-many relationship by nature. It looks a little more complicated, however, because the word forms rarely stand alone as synsets. Usually, a vertically polysemous word form is one of many members in a synset. The members of a synset stand in a lexical relation of synonymy, which is a many-to-one relation (many forms for the same meaning (concept)). This does not mean, however, that vertical polysemy applies to all members of a synset. Vertical polysemy is a feature of a particular word/lexeme, which has both broader and narrower or more specific meanings. In the structure of a wordnet, vertical polysemy means in practice that a word form appears as a member of synsets at different locations in the conceptual hierarchy.

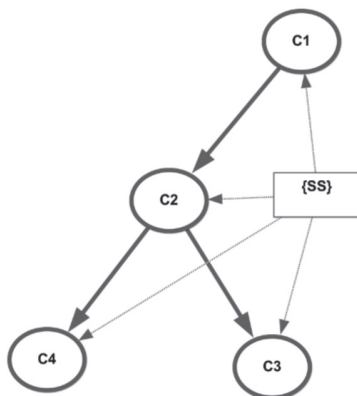


Figure 3. Vertical polysemy

As mentioned in the introduction, the notion of vertical polysemy refers to a lexical relationship where a term occurs in the entry of both a subordinate and a superordinate concept at the same time, indicating broader and narrower meaning of a concept (Koskela 2011). In wordnet formalism, there is the characteristic IS-A relation between these two levels in the case of nouns and the IS-A / MANNER-OF relation in the case of verbs.

3. The basics of the patterns: previous work

Previous work in relation to vertical polysemy (by using the terms auto-hyponymy and auto-troponymy) aimed to find sense clusters of polysemous words (Peters et al. 1998, Lin et al. 2002) and reduce the structures caused by polysemy in the semantic hierarchy of wordnets in order to transform the word senses from fine-grained ones to coarse-grained ones (Mihalcea 2003).

In the present study, we search the wordnets for structures similar to basic polysemic patterns (as elementary polysemy structures) (introduced by Lin et al. 2002). These patterns subsume cases where two or more synsets contain the same term and, at the same time, the synsets are related to each other in the semantic network. The basic patterns are called *sisters*, *twins*, *child*, *chain*, and *triangle*; the three last ones display vertical polysemy. In those cases, the same word form occurs both on the subordinate and superordinate level of the semantic hierarchy. Practically, this covers the cases where the same word is used in both a general and a more specific meaning(s). The *child*, *chain* and *triangle* patterns are illustrated in Figures 4–6. The illustrations have been extracted from the hierarchies of verbs and nouns in Princeton WordNet (Lohk 2015). In our explorative study, we search for more patterns implying the relation of vertical polysemy.

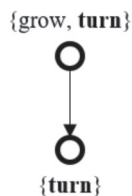


Figure 4. Child is a pattern where the same term occurs in a synset and its superordinate synset

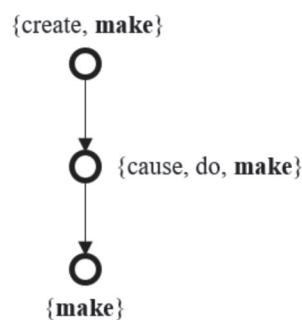


Figure 5. Chain is a pattern where the same word form appears in sub- and superordinate synsets for three or more times, sequentially

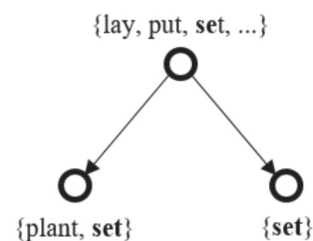


Figure 6. Triangle is a pattern where the same term appears simultaneously in three synsets: in two co-hyponym synsets and their superordinate synset

4. Algorithm

The core idea of the algorithm is to find all pairs of synsets that are adjacent to each other in a wordnet and share at least one lemma. After finding such pairs, we will test if they are directly connected to any other such pair in Wordnet. By connecting pairs of synsets, we find patterns of polysemy.

4.1. Algorithm description

A wordnet contains a set of synsets W , set of senses S , and a set of lemmas L . For each sense $s \in S$ there is exactly one corresponding synset $w \in W$ which in turn contains one or more lemmas $l \in L$. All lemmas of a synset w_i are in $L_{w_i} \subset L$. Therefore, each sense s is represented by one or more lemmas. Each lemma has one or more senses and therefore may belong to one or more synsets.

All synsets in W form a directed graph G which represents a semantic hierarchy of senses. Each synset w is a vertex in G and the edges represent either an IS-A or a MANNER-OF relationship.

Step 1. Find a set P of all possible pairs of synsets $p \in P : p = (w_1, w_2)$ where w_1 and w_2 are adjacent to each other in G and contain the same lemma l_p such that $l_p \in L_{w_1} \wedge l_p \in L_{w_2}$. So, each synset pair p has a common lemma l_p .

Step 2. Find all subsets $P_i \subset P$ that share a common lemma. For this, define an equivalence relation $p_1 \sim p_2 \Leftrightarrow l_{p_1} = l_{p_2}$.

Step 3. From each subset P_i from Step 2 extract all synsets from each $p_j \in P_i : p_j = (w_{j_1}, w_{j_2})$ into U_i . From each U_i remove duplicate synsets, so all $w_j \in U_i$ are unique.

Step 4. For each U_i extract a subgraph C_i from a semantic hierarchy graph G such that C_i contains only the vertices $w_j \in U_i$, and the set of edges E_i between them.

Step 5. View C_i as an undirected graph and extract all connected components. Output these components for visualization.

4.2. Trivial algorithm

While the algorithm described in previous section is a preferred approach as it traverses the full graph G only once for Step 1 and is therefore faster, we will also present a trivial version, which is easier to understand and implement. It uses the same notation as in the full algorithm.

Algorithm 1 Vertical Polysemy

Input: $graph : G \langle V, E \rangle$, $synsets : W$, $lemmas : L$
 $out = \emptyset$

for all lemma $l \in L$ **do**

$P \leftarrow findAdjacentPairs(G, l)$

$U_l = \emptyset$

for all pair $p \in P$ **do**

$w_1, w_2 = p[0], p[1]$

$U_l = U_l \cup \{w_1 \mid w_1 \notin U_l\}$

$U_l = U_l \cup \{w_2 \mid w_2 \notin U_l\}$

end for

$C_i = \langle U_l, \{w_u w_v \mid w_u w_v \in E \wedge w_u, w_v \in U_l\} \rangle$

$out = out \cup connectedComponents(C_i)$

end for

return out

5. Polysemy structures in wordnets

We applied the algorithm onto the 11 wordnets available on the Open Multilingual Wordnet website. The algorithm retrieved all the patterns that matched the set criteria. As a result, we discovered novel shapes of related items besides the basic patterns of vertical polysemy (*child*, *chain*, and *triangle*) mentioned above (Figures 4–6). These patterns, though not completely new in nature, have not yet been reflected in the context of vertical polysemy. The patterns are labelled *multiple inheritance*, *shortcut*, and *shortcut with cycle*.

The next figures (Figures 7–11) originate from four different wordnets. Every node label contains only the term common in all nodes of its substructure and synonym set ID. The most basic structure here is the pattern of polysemy with *multiple inheritance* (Figure 7). The next one (Figure 8) is known as a *shortcut*. Here, it seems that one additional link is added to the *multiple inheritance* structure. The next two (Figures 9–10) are *shortcut* structures with an additional connection that causes the cycle. In Figure 11, purely two *shortcut* structures are together with an additional link that again causes the *cycle*.

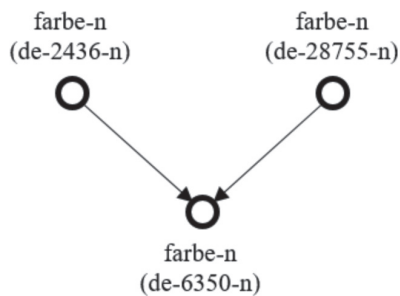


Figure 7. Pattern of multiple inheritance caused by vertical polysemy. Example of German *farbe* 'color' from Odenet

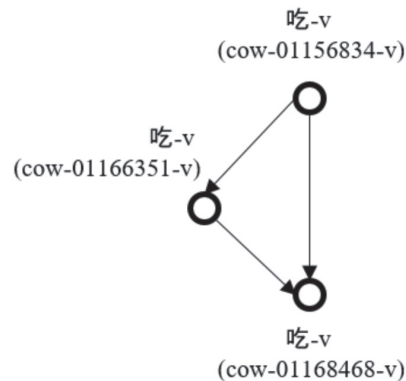


Figure 8. Pattern of shortcut caused by vertical polysemy. Example of Chinese 吃 'eat' from Chinese Open WordNet

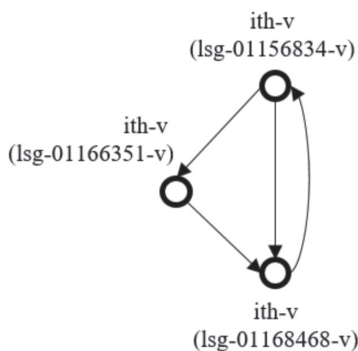


Figure 9. Shortcut structure with a cycle. The example of Gaelic *ith* 'eat' from Gaelic Wordnet

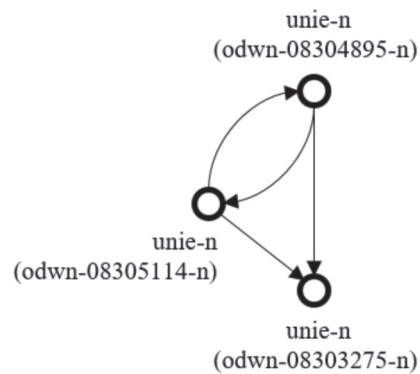


Figure 10. Shortcut structure with a cycle. The example of Dutch *unie* 'union' from Open Dutch Wordnet

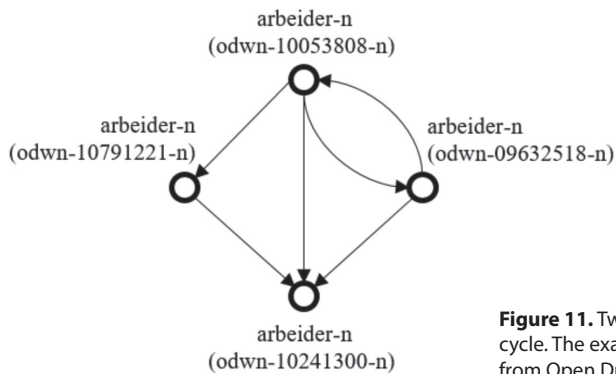


Figure 11. Two shortcut structures with a cycle. The example of Dutch *arbeider* ‘worker’ from Open Dutch Wordnet

Table 1 presents the vertical polysemy relations in the 11 selected wordnets in alphabetical order (according to the name). We restrict our overview to the relations of nouns and verbs. In particular columns, three of the most extreme values are presented in bold font.

The second column in Table 1 includes the data on hyponymy relations in each wordnet as a basis for comparison. The third column includes the number of relations of vertical polysemy. The relative frequency of vertical polysemy as compared to hyponymy is presented on diagram in Figure 12. The statistics reveal the highest relative prominence of vertical polysemy in Irish LSG and the second highest in Finnish FinWN.

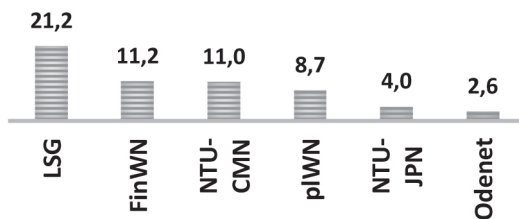


Figure 12. Proportions of hyponymy relations associated with vertical polysemy (%; the abbreviations are explained in Table 1)

In the fourth column, the distribution of vertically polysemous relations across the PoS is presented. There is a trend that nouns prevail in the polysemic relations, except for the Polish plWN-eng. The fifth column presents the number of clusters, and the sixth their distribution across PoS. The proportions are similar to those of polysemic relations, except for the Japanese NTU-JPN. In the latter case we see a more balanced noun to verb ratio. The seventh column provides statistics about the number of extended clusters of vertical polysemy. We set the baseline at five relations. The Finnish, Chinese and Irish wordnets include the largest numbers of such clusters.

It is characteristic that the patterns of polysemy combine and create complex clusters whose size can extend up to 20 relations. We present the largest clusters of each wordnet in the Appendix (see Figures 13–23). When there are multiple clusters sharing the same maximum size, only one is presented as an illustration.

Table 1. Statistical indicators related to vertical polysemy and polysemy structures

(1) Wordnet	(2) Relations of hyponymy (only nouns and verbs)	(3) Relations of hyponymy in connection to VP (vertical polysemy)	(4) Distribution of PoS-s among the relations of VP	(5) Nr of polysemy clusters (only nouns and verbs)	(6) Distribution of PoS-s among the clusters	(7) Nr of clusters >=5
Odenet (German)	1 524	42	n : 42 (85,7%) v : 7 (14,3%)	47	n : 42 (89,4%) v : 5 (10,6%)	0
EstWN (Estonian)	79 711	254	n : 191 (72%) v : 74 (28%)	227	n : 164 (72,2%) v : 63 (27,7%)	3
FinWN (Finnish)	91 879	10281	n : 8044 (69,8%) v : 3485 (30,2%)	7478	n : 5570 (74,5 %) v : 1908 (25,5%)	435
LSG (Irish)	19 117	4062	n : 4711 (73%) v : 1713 (27%)	4752	n : 3570 (75,1%) v : 1182 (24,9%)	118
NTU-CMN (Chinese)	89 371	9806	n : 9799 (75%) v : 3312 (25%)	9313	n : 7212 (77,4%) v : 2101 (22,6%)	336
NTU-JPN (Japanese)	89 371	3544	n : 4611 (92%) v : 385 (8%)	6704	n : 3824 (57,0%) v : 2880 (43,0%)	104
ODWN (Dutch)	102 789	1814	n : 3134 (84%) v : 600 (16%)	2176	n : 1832 (84,2%) v : 344 (15,8%)	6
OWN-PT (Portuguese)	8 577	4	n : 4 (100%)	4	n : 4 (100%)	0
pIWN (Polish)	167 463	1743	n : 1190 (78%) v : 301 (22%)	1357	n : 1089 (80,3%) v : 268 (19,7%)	9
pIWN-eng (English)	97 596	352	n : 88 (23%) v : 292 (77%)	355	n : 84 (23,7%) v : 271 (76,3%)	1
TrWN (Turkish)	4 647	9	n : 9 (90%) v : 1 (10%)	11	n : 10 (90,9) v : 1 (9,1%)	0

6. Discussion

The algorithm enabled us to detect some novel patterns of vertical polysemy – *multiple inheritance*, *shortcut* (with *cycle*) –, which contribute to the ones known previously as *child* and *chain* (Lin 2002).

The distribution along PoS demonstrated that nouns prevail in polysemic relations (the proportion being ca 75–90%), in general, except for the Polish p1WN-eng, which demonstrates an inverse proportion. This is an interesting finding, considering that it is claimed that there is twice as much polysemy (at the lexical level) among verbs as among nouns (Fellbaum 1998). The comparison of wordnets demonstrated that this claim may well hold true for English but its validity about for other languages has yet to be proved. It is also noteworthy that the Polish p1WN shows an exceptionally large amount of vertical polysemy among adjectival relationships (in total 259 cases), while PoSs other than nouns and verbs are generally rather marginal in this sample of wordnets considering vertical polysemy.

Two wordnets – the Chinese NTU-CMN and the Japanese NTU-JPN – show an identical number of hyponymy relations (because they were developed in parallel), while the proportion of their vertical polysemy varies dramatically. Chinese shows a stronger tendency for vertical polysemy than Japanese does.

Another pair with a close number of hyponym relations is the Finnish FinWN and the Polish p1WN-eng. This is because the Finnish FinWN has been compiled by translating the Polish PWN with the help of professional translators. However, the amount of vertical polysemy in the Finnish wordnet is much higher. Apparently, there is a tendency to use the same terms in both broad and narrow meanings in Finnish.

There are some objective reasons for the patterns of vertical polysemy and the extended clusters to occur. Many of the structures are caused by the economy principle of languages (i.e. general meaning is transferred to a more specific meaning or to a domain of terminology). Yet another reason for their emergence results from the subjective choices of the lexicographers (e.g. whether to present the two related meanings as a configuration of a *sister* structure or as two *children* instead). A third possible reason emerges from practical work. In some cases, the material for a wordnet is acquired, at least partly, by translating another wordnet and mapping its relations, as is the case with the Finnish wordnet. In such a case, the synsets contain material that originates in the “logic” of two languages and some redundancy and redundant complexity is only to be expected.

The wordnets can be compared in respect of the wealth of the structures of vertical polysemy and maximal size clusters. The biggest proportion of vertical polysemy occurs in the Irish LSG, and the largest clusters in the Polish p1WN-eng, the Chinese NTU-CMN and the Finnish FinWN. One must consider that wordnets with a higher proportion of such structures might have disadvantages in supporting NLP tasks where sense disambiguation is required. We suggest that the primary goal for developers should be to eliminate the *cycles*, in particular, but also the *shortcuts* need reconsideration as a violation of the economy principle (two concepts are connected both directly and indirectly). However, more research is needed in order to determine the optimal tolerance rate of a wordnet for vertical polysemy and maximum size clusters.

7. Conclusion

The study of polysemy occurring in the semantic hierarchies of wordnets reveals one of the central problems which needs to be considered in the case of distinctive NLP tasks that require semantic analysis. The developed algorithm captures the vertically polysemic structures in the 11 wordnets uploaded to Open Multilingual Wordnet. We were also able to find the extended patterns that occur in the wordnets and to provide comparative statistics. Besides the maximum-size polysemy clusters, also the number of vertical polysemy relationships and the PoS distribution in polysemic relations were examined. The analysis showed that at least in the studied wordnets, most cases of vertical polysemy relations are related to nouns, whereas the largest clusters are formed around verbs.

In the future perspective, the goal of the analysis of large polysemy clusters is to develop a way to detect possible errors in the semantic network and possibly an application that can be used for any other wordnet. We are planning to study these large polysemy clusters from three perspectives: single-member synonym sets, the existence of synonym sets' definitions, and the semantic similarity of definitions.

References

- Bond, Francis; Foster, Ryan 2013. Linking and extending an open multilingual wordnet. – 51st Annual Meeting of the Association for Computational Linguistics: ACL-2013. Sofia, 1352–1362.
- Bond, Francis; Paik, Kyonghee 2012. A survey of wordnets and their licenses. – Proceedings of the 6th Global WordNet Conference (GWC 2012). Matsue, 64–71.
- Cruse, Alan 2000. *Meaning in Language: An Introduction to Semantics and Pragmatics*. Oxford: Oxford University Press.
- Fellbaum, Christiane 1998. *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press. <https://doi.org/10.7551/mitpress/7287.001.0001>
- Fellbaum, Christiane 2002. On the semantics of troponymy. – Rebecca Green, Carol A. Bean, Sung Hyong Myaeng (Eds.), *The Semantics of Relationships: An Interdisciplinary Perspective*. Springer, 23–34. https://doi.org/10.1007/978-94-017-0073-3_2
- Fellbaum, Christiane 2015. Lexical relations. – John R. Taylor (Ed.), *The Oxford Handbook of the Word*. Oxford: Oxford University Press, 350–363. <https://doi.org/10.1093/oxfordhb/9780199641604.013.028>
- Freihat, Abed Alhakim; Giunchiglia, Fausto; Dutta, Biswanath 2013. Approaching regular polysemy in WordNet. – *The Fifth International Conference on Information, Process, and Knowledge Management (eKNOW)*. Nice, France: IARIA XPS Press, 63–69.
- Koskela, Anu 2005. On the Distinction Between Metonymy and Vertical Polysemy in Encyclopaedic Semantics. *University of Sussex Working Papers in Linguistics and English Language*, University of Sussex, Falmer.
- Koskela, Anu 2011. Metonymy, category broadening and narrowing, and vertical polysemy. – Réka Benczes, Antonio Barcelona, Francisco José de Mendoza Ibáñez (Eds.), *Defining Metonymy in Cognitive Linguistics: Towards a Consensus View*. *Human Cognitive Processing* 28. Amsterdam: John Benjamins Publishing Co., 125–146. <https://doi.org/10.1075/hcp.28.07kos>
- Lin, Jen-Yi; Yang, Chang-Hua; Tseng, Shu-Chuan; Huang, Chu-Ren 2002. The structure of polysemy: A study of multi-sense words based on WordNet. – *Proceedings of the*

- 16th Pacific Asia Conference on Language, Information and Computation. Jeju: The Korean Society for Language and Information, 320–329.
- Lohk, Ahti 2015. A System of Test Patterns to Check and Validate the Semantic Hierarchies of Wordnet-type Dictionaries. Dissertation. Tallinn: TUT Press.
- Lohk, Ahti; Orav, Heili; Vare, Kadri; Bond, Francis; Vaik, Rasmus 2019. New polysemy structures in wordnets induced by vertical polysemy. – Proceedings of the Tenth Global Wordnet Conference. Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej, 394–403.
- Lyons, John 1995. Linguistic Semantics. An Introduction. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511810213>
- Mihalcea, Rada 2003. Turning wordnet into an information retrieval resource: Systematic polysemy and conversion to hierarchical codes. – *International Journal of Pattern Recognition and Artificial Intelligence*, 17 (5), 689–704. <https://doi.org/10.1142/S0218001403002605>
- Pedersen, Bolette; Agirrezabal, Manex; Nimb, Sanni; Olsen, Sussi; Rørmann, Ida 2018. Towards a principled approach to sense clustering: A case study of wordnet and dictionary senses in Danish. – The 9th Global WordNet Conference. Singapore: Global WordNet Association, 183–190.
- Peters, Wim; Peters, Ivone; Vossen, Piek 1998. Automatic sense clustering in EuroWordNet. – Proceedings of the 1st International Conference on Language Resources and Evaluation. Granada, Spain: European Language Resources, 409–416.
- Pociello, Elisabete; Agirre, Eneko; Aldezabal, Izaskun 2011. Methodology and construction of the Basque WordNet. – *Language Resources & Evaluation*, 45 (2), 121–142. <https://doi.org/10.1007/s10579-010-9131-y>
- Pajusalu, Renate 2009. Sõna ja tähendus ['Word and Meaning']. Tallinn: Eesti Keele Sihtasutus.
- Saeed, John 2001. Semantics. Cambridge: Blackwell. <https://doi.org/10.1016/B0-08-043076-7/02989-2>

Appendix. The maximum-size clusters of polysemy in 11 wordnets²

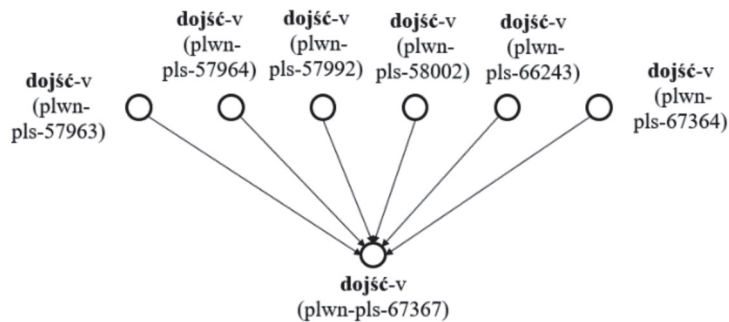


Figure 13. The largest polysemy structure in Polish wordnet (plWordNet). Lexical unit *dojść*

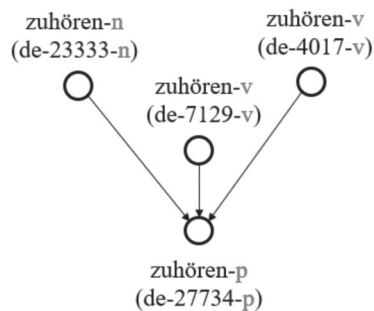


Figure 14. The largest polysemy structure in German wordnet. Lexical unit *zuhören*

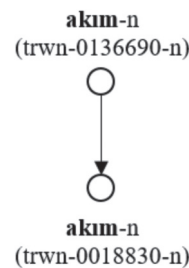


Figure 15. The largest polysemy structure in Turkish wordnet. Lexical unit *akım*

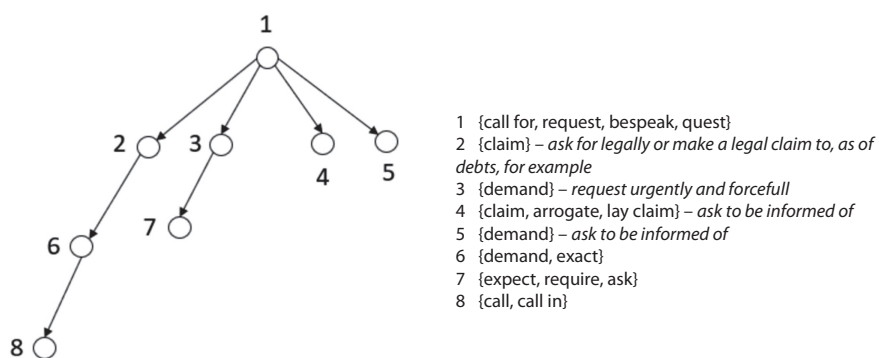


Figure 16. The largest polysemy structure of Japanese wordnet. Lexical unit 要求+する

² Node numbering has been used for larger patterns of polysemy (> 7). In this case, the nodes are labeled with the English equivalent terms from the Princeton wordnet.

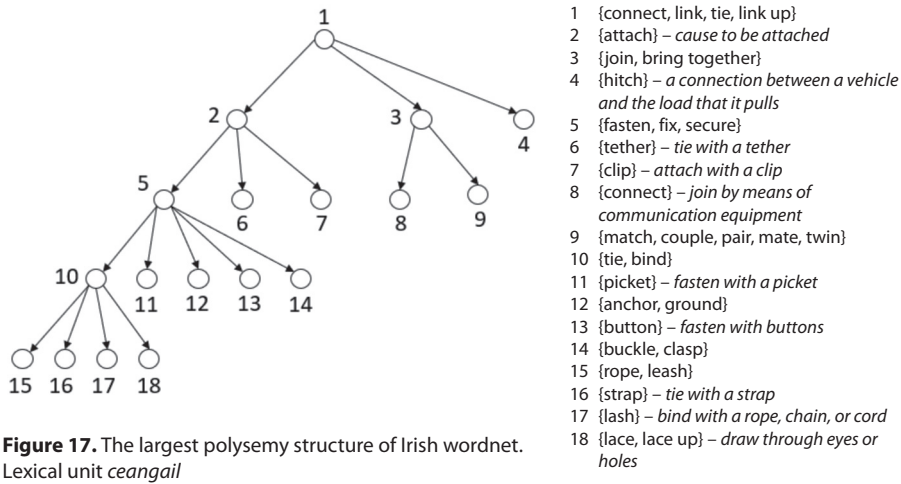


Figure 17. The largest polysemy structure of Irish wordnet. Lexical unit *ceangail*

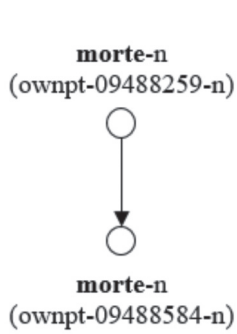


Figure 18. The largest pattern of polysemy in Portuguese wordnet. Lexical unit *morte*

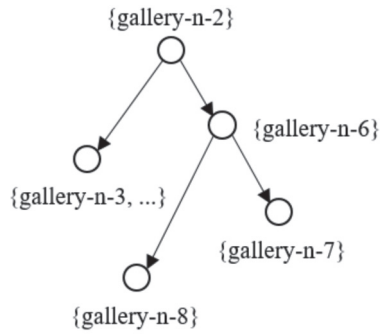


Figure 19. The largest polysemy structure of Estonian Wordnet. Lexical unit *galerii* 'gallery'

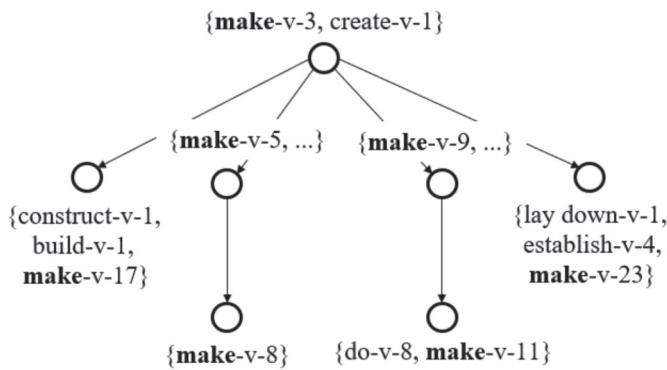


Figure 20. The largest polysemy structure in Princeton WordNet. Lexical unit *make*

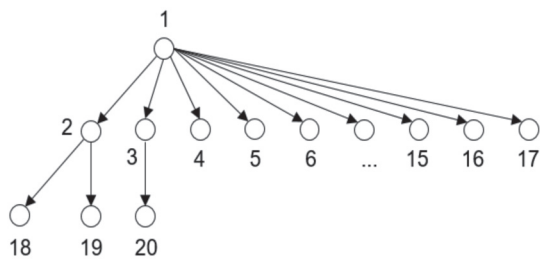


Figure 21. The largest polysemy structure in Finnish wordnet. Lexical unit *poistaa*

- 1 {remove, take, take away}
- 2 {take away, take out}
- 3 {delete, cancel}
- 4 {wipe off, wipe away}
- 5 {extirpate, excise}
- 6 {eliminate} – *remove (an unknown variable) from two or more equations*
- 7 {pull} – *take away*
- 8 {clean, strip}
- 9 {bur, burr}
- 10 {tusk, detusk}
- 11 {expectorate, clear out, drive out}
- 12 {ablate} – *remove an organ or bodily structure*
- 13 {stem} – *remove the stem from*
- 14 {harvest} – *remove from the culture or ...*
- 15 {draw off, pull off, draw away}
- 16 {unburden, disburden}
- 17 {strip} – *remove a constituent from a liquid*
- 18 {delete, erase}
- 19 {obliterate, kill, wipe out}
- 20 {strike, excise, expunge, scratch}

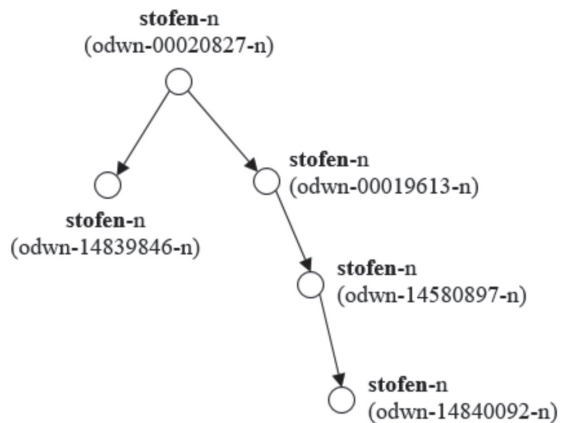


Figure 22. The largest polysemy structure in Dutch wordnet. Lexical unit *stofen*

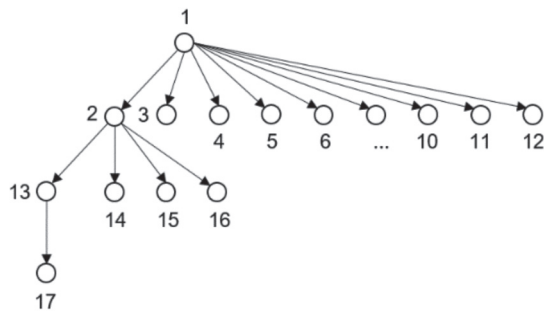


Figure 23. The largest polysemy structure in Chinese wordnet. Lexical unit 获得

- 1 {get, acquire}
- 2 {find, line up, come up, get hold}
- 3 {buy} – *acquire by trade or sacrifice or exchange*
- 4 {come upon, luck into, enter upon}
- 5 {pick up} – *get in addition, as an increase*
- 6 {earn, make, bring, realize, gain}
- 7 {win, gain, acquire, obtain, ...}
- 8 {get} – *acquire as a result of some effort or action*
- 9 {buy, purchase} – *obtain by purchase*
- 10 {come by, come into} – *obtain, especially accidentally*
- 11 {earn, garner} – *acquire or deserve by one's efforts or actions*
- 12 {obtain} – *come into possession of*
- 13 {source} – *get (a product) from another country or business*
- 14 {secure, procure} – *get by special effort*
- 15 {extract} – *get despite difficulties or obstacles*
- 16 {gain, derive, obtain}
- 11 {earn, garner} – *acquire or deserve by one's efforts or actions*
- 12 {obtain} – *come into possession of*
- 13 {source} – *get (a product) from another country or business*
- 14 {secure, procure} – *get by special effort*
- 15 {extract} – *get despite difficulties or obstacles*
- 16 {gain, derive, obtain}
- 17 {draw, reap}

VERTIKAALSE POLÜSEEMIA LAIENDATUD KLASTRID: ÜHETEISTKÜMNE *WORDNET* VAATLUS

Ahti Lohk^{1,2}, Ene Vainik²,
Geda Paulsen^{2,3}, Martin Rebane¹,
Francis Bond⁴

Tallinna Tehnikaülikool¹, Eesti Keele Instituut²,
Uppsala Ülikool³, Nanyangi Tehnikaülikool⁴

Käesoleva uurimistöö eesmärk on tuvastada ja analüüsida maksimaalse suurusega polüseemiaklastreid veebilehelt *Open Multilingual Wordnet* välja valitud üheteistkümneme suurema *wordnet*'i semantilistest hierarhiatest. Polüseemiaklastrid on ekstraheeritud nii nimisõna kui ka verbi semantilistest hierarhiatest. Uurimuses keskendutakse sellistele semantilise hierarhia alamstruktuuridele, mis sisaldavad vaid vertikaalse polüseemia seoseid (s.o sellistele klastritele, mille tipud sisaldavad sama kirjpildiga sõnu). Peale maksimaalse suurusega klastrite käsitleme vertikaalse polüseemia seoste osakaalu hierarhiliste seoste hulgas ning sõnaliikide jaotust. Uuringu tulemused näitavad, et suured polüseemiaklastrid võivad viidata võimalikele probleemidele loomuliku keele masintöötlemise ülesannetes, nagu näiteks sõna tähenduse ebatäpne tuvastamine. Maksimaalse suurusega klastrite tuvastamiseks loodud algoritm on rakendatav ka neile *wordnet*'idele, mida selles töös ei käsitletud. Artiklis esitatud meetod on esimene samm suuremõõtmeliste polüseemiaklastrite analüüsis. Artikli autorid on seisukohal, et polüseemiaklastrid aitavad hinnata *wordnet*'ide semantiliste hierarhiate seisundit ja nende sobivust loomuliku masintöötlemise ülesannete lahendamise teadmusbasesina.

Võtmesõnad: vertikaalne polüseemia, polüseemiamallid, Wordneti semantilised hierarhiad

Ahti Lohk's (Tallinn University of Technology, Institute of Estonian Language) research areas are validation of wordnet semantic hierarchies with graph-based methods and text mining algorithms for extraction of useful, new, and applicable information from unstructured texts.
Akadeemia tee 15A, 12618 Tallinn, Estonia
ahti.lohk@taltech.ee

Ene Vainik is a cognitive linguist working at the Institute of the Estonian Language as a senior researcher. Her field of studies covers semantics, folk-psychology and, in particular, conceptualisations of emotions. Another related field has been affective computing (e.g detecting emotions in written Estonian text and detecting emotional valence). Currently, she works for a project on part of speeches from the perspective of lexicography.
Roosikrantsi 6, 10119 Tallinn, Estonia
ene.vainik@eki.ee

Geda Paulsen's (Institute of the Estonian Language, University of Uppsala) research interests include lexical semantics, lexicography, morphology, word formation, corpus linguistics and contrastive linguistics.

Roosikrantsi 6, 10119 Tallinn, Estonia
geda.paulsen@eki.ee, geda.paulsen@moderna.uu.se

Martin Rebane (Tallinn University of Technology) is exploring and developing autonomous navigation algorithms based on deep learning but is also happy to work on natural language processing and humor research issues.

Akadeemia tee 15A, 12618 Tallinn, Estonia
martin.rebane@taltech.ee

Francis Bond's (Nanyang Technological University) main research areas are lexical semantics and word sense disambiguation.

14 Nanyang Avenue, 639818 Singapore
fcbond@ntu.edu.sg