An Experiment: Finding Parents for Parentless Synsets by Means of CILI

Ahti Lohk[1], **Martin Rebane**[2] and **Heili Orav**[3]

[1]Department of Software Science, Tallinn University of Technology, Tallinn, Estonia
[2]School of Engineering, University of Warwick, Coventry, United Kingdom
[3] Department of Computer Science, University of Tartu, Tartu, Estonia

<ahti.lohk@taltech.ee,
martin.rebane@warwick.ac.uk,
heili.orav@ut.ee>

Abstract

Identifying and correcting inconsistencies in wordnets is a natural part of their development. Focusing only on the sub-problem of *missing links*, we aim to find automatically possible parents for parentless synsets in IS-A hierarchies of a target wordnet by means of source wordnets where target and source wordnets are in XML-format and equipped with Collaborative Interlingual Index (CILI).
In this paper, we describe the algorithm and provide statistics on the possible parents of parentless synsets of the wordnets included in the study. Additionally, we investigate the suitability of the proposed potential parent synsets for correcting noun and verb synsets within the Estonian wordnet.

1 Introduction

One of the main goals of wordnet (Fellbaum 1998) development is to make it accessible while ensuring its correctness.

The developer must consider that wordnet errors can be formal, semantic, or structural, where **formal errors** are related to the source file structure or data presentation in it, **semantic errors** are related to wordnet semantics and **structural errors** are related to wordnet as a graph (Piasecki et al., 2013). The category of structural errors is set apart from formal and semantic errors in that it doesn't require any knowledge of the wordnet language, but correcting it requires the assistance of a lexicographer (Lohk 2015).

Structural errors often result in **missing links** between wordnet synsets, which is one of the most obvious problems. This type of problem can appear either as 1) synsets that are completely lacking semantic relationships, as 2) small separate hierarchies, or as 3) a big number of parentless synsets.

In identifying parentless synsets for noun and verb synsets, it must be considered that the synsets named as root concepts (or unique beginners) cannot have parents[1]. For example, only one root concept of the IS-A noun hierarchy - {entity} - has been considered correct for the Princeton WordNet. On the Dutch wordnet Cornetto (version 2)[2], however, the corresponding number is two. The same number of root concepts is also assigned to verb hierarchies of Cornetto. (Lohk, 2015). These three examples point to a situation where there is no problem with parentless synsets. Nevertheless, this problem is common to all wordnets tested by us in this experiment. For example, the verb hierarchies of Open English WordNet (OEWN) contain as many as 574 parentless synsets (see Table 1).

The fact that there are synsets with missing links in wordnet has been pointed out by other authors (Smrž, 2004, Richens, 2011). However, to the best of the authors' knowledge, no solution has been proposed that automatically provides a possible parent for a parentless synset. This article tries to partially fill this gap, focusing primarily on such missing links, where synset lacks a parent or higher-level concept (superordinate). An additional refinement of the proposed approach comes from

[1] An exception is synsets, which are labeled as nouns but are names in terms of content.

[2] http://www.cltl.nl/projects/previous-projects/cornetto/

the fact that we take advantage of the information available on wordnets equipped with Collaborative Interlingual Index (CILI) (Bond et al., 2016).

To conduct the experiment, we utilize the following wordnets: Estonian wordnet (EstWN, version 2.5)[3] (Orav et al., 2018), Open English WordNet (OEWN, version 2021) [4] and six wordnets downloaded from the Open Multilingual Wordnet [5] website: Open German WordNet (Odenet) [6], Open Dutch WordNet (ODWN) (Postma et al., 2016), Finnish WordNet (FinWN)

Wordnet (language)	Parentless synsets	
	noun	verb
OEWN (English)	8	574
EstWN (Estonian)	190	13
Odenet (German)	3 433	2 583
ODWN (Dutch)	0	87
FinWN (Finnish)	172	559
LSG (Irish)	6 000	1 468
OWN-PT (Portuguese)	18 577	7 143
NTU-JPN (Japanese)	5 766	420

Table 1: Number of parentless synsets in wordnets.

(Lindén et al., 2010), Irish Language Semantic Network (LSG)[7], Open Brazilian Wordnet (WN-PT) (de Paiva et al., 2012), Japanese Open Wordnet (NTU-JPN) (Isahara et al., 2008). All eight wordnets are in XML-format and many of their synsets are CILI- equipped.

The main idea behind our approach is to provide possible parents for parentless synsets in target wordnet using other wordnets. More specifically, this means that a possible parent can only be provided if both the target wordnet synset and its possible parent are equipped with a CILI, and so are the synsets from other wordnets corresponding to the same CILIs.

The paper is organized in the following manner: Section 2 formulates the algorithm to find parents for parentless synsets by means of CILI. Next, Section 3 describes the format for reporting the results and provides descriptive statistics about the results obtained. Section 4 focuses on the case study of Estonian Wordnet. Section 5 concludes the paper and its findings.

2 Algorithm

Each wordnet w contains a set of synsets S. Each synset $s \in S$ has a unique ID number i and might have an optional Collaborative Inter Lingual Index (CILI) $c \in C$ where C is a set of all CILIs. ID i is unique with a wordnet, but each wordnet uses its own set of ID numbers. CILI c is also unique within a wordnet but all wordnets use the same c for equivalent synsets. Additionally, most (but not all) synsets have hierarchical parent-child relationship structure. However, such relationship might exist in a language but be missing in the wordnet. CILIs make it possible to use one wordnet w_{src} as a source to estimate the hierarchical parent-child relationship of the other wordnet w_{tgt} (target). The algorithm in this Section does this by using a set of CILIs C_{src} of w_{src}, a set of CILIs C_{tgt} of w_{tgt}, parent-child relationship map M_{src} Msrc of w_{src} for computing a parent-child relationship map M_{tgt} for w_{tgt}. Each CILI c in C_{src} has an associated set of parent CILIs P_c. Each CILI $c \in C$ also has an associated ID number i. Therefore, it is possible to construct a map M_{tgt} that represents estimated parent-child relationships for target wordnet w_{tgt} based on similar relations in w_{src}. We introduce Algorithm 1 to construct such map.

Algorithm 1 Synset Sync

Input: $map : M_{src}, set : C_{src}, set : C_{tgt}$
$M_{tgt} = map : \emptyset$
for all $c \in C_{tgt}$ **do**
  $P_c \leftarrow getParentCilis(c, M_{src})$
  for all $c_{parent} \in P_c$ **do**
    $i_{child} = getSynsetIdByCili(c, C_{tgt})$
    $i_{parent} = getSynsetIdByCili(c_{parent}, C_{tgt})$
    $relation: \ r =< i_{child}, i_{parent} >$
    $M_{tgt} = M_{tgt} \cup \{r \mid r \notin M_{tgt}\}$
  end for
end for
return M_{tgt}

We assume that it is trivial to map CILI c to a corresponding ID i and will represent this operation as a function $getSynsetIdByCili(cili : c, set : C)$. Finding parents using a map data structure is also a standard procedure in every programming language, hence we represent this as

[3] https://gitlab.keeleressursid.ee/avalik/data/-/tree/master/estwn/estwn-et-2.5
[4] https://en-word.net/
[5] https://github.com/globalwordnet/OMW
[6] https://ikum.mediencampus.h-da.de/projekt/open-de-wordnet-initiative
[7] https://cadhan.com/lsg/index-en.html

a function $getParentCilis(cili : c, map : M)$. The resulting map M_{tgt} contains all found parent-child relations for w_{tgt} based on similar relations in w_{src}. Therefore, it does not limit the depth of hierarchy, i.e., the algorithm is able to find and store complex and deep hierarchical relations.

3 Results

Within this section, we describe the format for reporting the results and provide descriptive statistics about the results obtained.

3.1 Presentation format of results

The examples presented in **Appendices A-D** give an idea of the format for presenting the results. Here we provide a detailed overview of the structure of the presentation format.

Appendices A-D represent four categories of results, which are explained in more details in Section 4.

Each case of a parentless synset begins with a **sequence number**. The rest of the information is distributed among five fields. We will explain their content in more thoroughly below.

1) **Without parent**
   A target wordnet synset with no parent is displayed under "WITHOUT PARENT". Along with the synset, synset ID, and the OEWN equivalent synset are presented through CILI.
2) **Possible parent(s)**
   Finding possible parents is based on the CILIs identified under "PARENTS FROM OTHER WORDNET(S):" which refers to parents in other wordnets. "POSSIBLE PARENT(S)" is presented above because this information is more important to the lexicographer. If no parent is found for the target wordnet parentless synset through CILI, the text "No possible parent(s) through CILI" is returned.
3) **Parents from other wordnet(s)**
   This structure field gets its content based on the CILI given in the "WITHOUT PARENT" field. There are as many lines in this field as there are wordnets among the source wordnets that input CILI finds parent with CILI. Each line contains information about the CILI of the synset without a parent, the CILI of the synset with its corresponding parent, the synset ID given to the parent of the CILI in a particular wordnet, and the equivalent synset in the OEWN. The latter is added so that the content of synsets can be quickly captured. If no parent is found in source wordnets, the text "No possible parent(s) through CILI" is returned.
4) **Possible grandparent(s)**
   To get a broader background of the problem situation, we added possible grandparents in addition to possible parents. Finding possible grandparents is based on the CILIs identified under "GRANDPARENTS FROM OTHER WORDNET(S):" which refer to grandparents in other wordnets. If no grandparent is found for the target wordnet parentless synset through CILI, the text "No possible parent(s) through CILI" is returned.
5) **Grandparents from other wordnet(s)**
   The content of this field is derived like the "PARENTS FROM OTHER WORDNET(S)" field. The difference is that the CILIs used as input are the same as those given in the "Possible parents" field.

3.2 Statistics

Just as in Table 1, only synsets whose first, last and second member (lexical unit) does not start with a capital letter are considered in Table 2 to avoid synsets, which are defined by nouns, but which are names in terms of content. In the last column of the Table 2, the first two numbers represent cases where parents were found in the source wordnets regardless of whether a parent was also found in the target wordnet. Many of the figures seen in the table are very large. One reason for this is that the result contains both synsets with subordinates and those without (so called orb synsets). In the case of the EstWN, the number of synsets without parents is low as expected, since its structure has been validated with various graph methods in the last ten years (Lohk, 2015).

For our study, it is important to know in how many cases it is possible to obtain additional information for parentless synsets. This information can be obtained by dividing the last number in the third column (possible simultaneous absence of parents and grandparents) by the first number in the second column (number of synsets equipped with CILI and without parents). The resulting quotient gives an idea of how large amount of synsets lack a parent and/or grandparent. Parents and grandparents found through CILI seem to benefit the most in the case of the OWN-PT, where possible parents/grandparents information is

missing only in 0.8% of the cases (194/25660 x 100). It is followed by the Irish wordnet and EstWN, where these numbers are 1.5% and 12.2% respectively.

By comparing the number of parentless synsets in Tables 1 and 2, we can see in Figure 1 the extent to which parentless synsets are endowed with CILI.

Wordnet/ Language	Nr of parentless synsets with CILI total|noun|verb	Nr of possible parents| grandparents| no parents & grandparents
OEWN (English)	572|7|565	268|250|300
EstWN (Estonian)	41|35|6	36|36|5
Odenet (German)	2052|1313|739	1178|1140|874
ODWN (Dutch)	87|0|87	38|31|49
FinWN (Finnish)	730|171|559	410|390|319
LSG (Irish)	7454|5989|1465	7337|7258|114
OWN-PT (Portuguese)	25660|18517|7143	25457|25020|194
NTU-JPN (Japanese)	5950|5530|420|	5211|5192|739

Table 2: Descriptive statistics of results

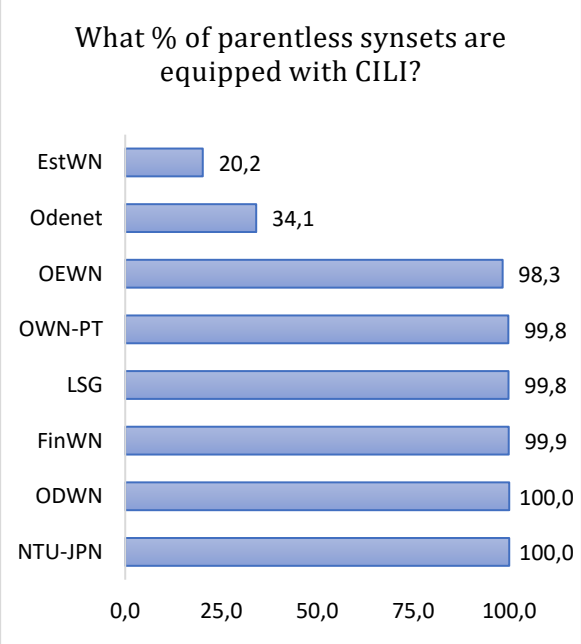

Figure 1: CILI proportions in parentless synsets in different wordnets

4 Case Study of Estonian Wordnet (EstWN)

In the EstWN analysis, our program found 41 parentless and CILI-equipped synsets. The noun synsets were represented 35 times and the verb synsets 6 times. In 7 cases out of 41 it was not necessary to determine a parent, as the synsets represented root concept.

After a closer examination of each of the 41 cases by the lexicographer, it was found that the decisions that had to be made in solving them fell into four categories (for each category, one example is given in the appendices):

1) The suggested possible parent was suitable for the parentless synset. **10 cases**. (See Appendix A)
2) The suggested possible grandparent was suitable for the parentless synset. **4 cases**. (See Appendix C)
3) The parentless synset turned out to be a root concept. **7 cases** (3 nouns + 4 verbs). (See Appendix B).
4) A parentless synset receives a parent that was not present in either the possible parents or grandparents. **21 cases**. (See Appendix D)

All root concepts classified under category 3 are not root concepts in any other language. This becomes obvious when comparing the EstWN root concepts with the corresponding synsets of OEWN. It turns out that two out of three EstWN noun root concepts have parent in the OEWN. That means, if the EstWN has ['existence', 'existence', ...] (['existence', 'being', '...']) as a root concept, then in the OEWN its parent is ['state']. Also, if the EstWN has ['fenomen', 'ilming', '...'] (['phenomenon'] as a root concept, then in the OEWN its parent is ['process', 'physical process'].

With four EstWN verb root concepts, it is noteworthy that no single source wordnet (including OEWN) offers any parents for them.

In the EstWN, root concepts for nouns and verbs are as follows:

1) (n) ['olev'] (['entity'], oewn-00001740-n)
2) (n) ['eksisteerimine', 'eksistents', 'olelu', '...'] (['existence', 'being', 'beingness', '...'], oewn-13977471-n)
3) (n) ['fenomen', 'ilming', 'nähe', '...'] (['phenomenon'], oewn-00034512-n)
4) (v) ['modifitseeruma', 'muutuma', '...'] (['switch', 'change', …], oewn-00551194-v)

5) (v) ['sooritama', 'tegema'] (['do', 'execute', 'perform'], oewn-01716563-v)
6) (v) ['eksisteerima', 'olema', '...'] (['exist', 'be'], oewn-02609706-v)
7) (v) ['olema'] (['be'], oewn-02610777-v)

Summarizing the results of the four categories, it is easy to decide about a parentless synset in approximately half of the cases. Such cases belong to categories 1 to 3. Most efforts should be made to resolve Category 4 cases where possible parents and/or grandparents have been suggested but are not suitable.

Hereby we give some examples where the suggested parent was unsuitable for the EstWN. Briefly, these cases can be summarized on the grounds that, although a parentless synset is related via CILI to synsets in other language wordnets, its semantic field is sufficiently different to be assigned the same parents as in other wordnets.

Example 1:
Parentless synset:
['smugeldamine', '...'] (['smuggling'])
Suggested parent:
['import', '...'] (['importation', 'importing'])
Correct parent:
[transport, '...'] ([['transport', 'transfer', '…'])
Argument:
smuggling in Estonian does not mean only import but also export

Example 2:
Parentless synset:
['möirataja', '…'] (['screamer', 'shouter', '...'])
Suggested parent:
['suhtleja'] (['communicator'])
Correct parent:
['hääletegija'] (*voice maker*). No corresponding CILI.
Argument:
'screamer' is not necessarily only a person in Estonian.

Example 3:
Parentless synset:
['amatörism'] (['amateurism'])
Suggested parent:
['conviction', 'articleoffaith', 'strongbelief']
Correct parent:
['harrastus'] (['avocation', 'by-line', 'hobby', '…']
Argument:
'amateurism' in Estonian is more of a hobby than conviction.

Example 4:
Parentless synset:
['foneetika', '...'] ([' phonetics'])
Suggested parent:
['akustika', 'heliõpetus'] (['acoustics'])
Correct parent:
[lingvistika, '...'] ([['linguistics'])
Argument:
The authoritative dictionary of the Estonian language (Sõnaveeb [8]) declares that phonetics is a part of linguistics.

5 Conclusion

The present study proposed an approach for identifying potential parents for parentless synsets equipped with the Collaborative Interlingual Index (CILI) feature using source wordnets. The method is applicable to all wordnets with different languages that have specific XML formats and CILI-equipped synsets and has the potential to enhance the quality of wordnets.

The experiment revealed that seven out of the eight wordnets analyzed contained a significant number of parentless synsets. However, the majority of these synsets, six out of eight wordnets, had 98% or more parentless synsets that were also equipped with a designated CILI. Possible parents and grandparents were automatically found for 43% to 99% of the parentless synsets across different wordnets, with 87% or more of the synsets having possible parents in half of the cases.

The study indicates that lexicographer involvement may be necessary to correct the identified inconsistencies (missing parents), and that synsets connected through CILI in different languages may have different meanings.

The proposed approach could also be applied to detect inconsistencies in synsets that already have parents in the future. Overall, the method presented in this study provides a useful tool for improving the quality of wordnets across various languages.

References

Fellbaum, D. C. 1998. *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press

Bond, F., Vossen, P. Th. J. M., McCrae, J. P. and Fellbaum, C. D., 2016. CILI: the collaborative interlingual index. In *Proceedings of the 8th Global WordNet Conference (GWC2016),* pp. 50-57. https://aclanthology.org/2016.gwc-1.9/

[8] https://sonaveeb.ee/

Isahara, H., Bond, F., Uchimoto, K., Utiyama, M. and Kanzaki, K., 2008. Development of the Japanese WordNet. *In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08),* pp. 2420-2423. https://aclanthology.org/L08-1269/

Lindén, K. and Carlson, L., 2010. *FinnWordNet–finnish wordnet by translation*. LexicoNordica–Nordic Journal of Lexicography, 17, pp.119-140.

Lohk, A. 2015. *A System of Test Patterns to Check and Validate the Semantic Hierarchies of Wordnet-type Dictionaries*. Tallinn, Estonia: TalTech Press.

McCrae, J.P., Rademaker, A., Bond, F., Rudnicka, E. and Fellbaum, C., 2019, July. English WordNet 2019 – An Open-Source WordNet for English. *In Proceedings of the 10th Global WordNet Conference (GWC2019),* pp 245-252.

de Paiva, V., Rademaker, A., de Melo, G. 2012. OpenWordNet-PT: An Open Brazilian Wordnet for Reasoning. *In Proceedings of COLING 2012: Demonstration Papers*. Mumbai, India, pp. 353-360.

Orav, H., Vare, K. and Zupping, S., 2018, January. Estonian Wordnet: Current State and Future Prospects. *In Proceedings of the 9th Global Wordnet Conference*, pp. 347-351.

Piasecki, M., Burdka, L., Maziarz, M., 2013. Wordnet Diagnostics in Development, *In Human Language Technologies as a Challenge for Computer Science and Linguistics*. Poznań, Poland, pp. 268–272.

Postma, M., Van Miltenburg, E., Segers, R., Schoen, A. and Vossen, P., 2016. Open Dutch WordNet. *In Proceedings of the 8th Global WordNet Conference (GWC2016)*, pp. 302-310.

Richens, T., 2008. Anomalies in the Wordnet Verb Hierarchy, *In Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. Association for Computational Linguistics (ACL)*, pp. 729–736.

Smrž, P., 2004. Quality control for wordnet development. *In Proceedings of the Second International WordNet Conference (GWC2004),* pp. 206-212.

Appendix A. The suggested possible parent which was suitable for parentless synset

```
--------------------------------------------------------------------------------
26 # Correct parent for ['puuderdama'] (['powder']) is ['maalima', 'meikima', 'minkima',
'...'] (['makeup'])
--------------------------------------------------------------------------------
WITHOUT PARENT
       i21979 estwn-et-19703-v|['puuderdama']
       (OEWN equivalent: oewn-00041904-v|['powder'])

POSSIBLE PARENT(S):
       i21972 estwn-et-5410-v|['maalima', 'meikima', 'minkima', '...']

PARENTS FROM OTHER WORDNET(S):
       (i21979)->i21972 cow-00040928-v oewn-00040659-v|['makeup']
        ...
       (i21979)->i21972 lsg-00040928-v oewn-00040659-v|['makeup']
       (i21979)->i21972 oewn-00040659-v oewn-00040659-v|['makeup']

POSSIBLE GRANDPARENT(S):
       i30124 estwn-et-70-v|['dekoorima', 'dekoreerima', 'ehtima', '...']
       i21970 estwn-et-173-v|['kohendama', 'kordaseadma', 'korrastama']

GRANDPARENTS FROM OTHER WORDNET(S):
       (i21972)->i21970 cow-00040353-v oewn-00040084-v|['neaten', 'groom']
        …
       (i21972)->i21970 oewn-00040084-v oewn-00040084-v|['neaten', 'groom']
       (i21972)->i21970 slownet-eng-30-00040353-v oewn-00040084-v|['neaten', 'groom']
```

Appendix B. The parentless synset which turned out to be the root concept

```
--------------------------------------------------------------------------------
31 # Synset ['olema'] (['be']) is a root concept
--------------------------------------------------------------------------------
WITHOUT PARENT
     i34713 estwn-et-148-v|['olema']
     (OEWN equivalent: oewn-02610777-v|['be'])

POSSIBLE PARENT(S):
     No possible parent(s) through CILI

PARENTS FROM OTHER WORDNET(S):
     No possible parent(s) through CILI

POSSIBLE GRANDPARENT(S):
     No possible grandparent(s) through CILI

GRANDPARENTS FROM OTHER WORDNET(S):
     No possible grandparent(s) through CILI
```

Appendix C. The suggested possible grandparent which was suitable for parentless synset

```
--------------------------------------------------------------------------------
8 # Correct parent for ['akkommodatsioon'] (['accommodation']) is possible
  grandparent ['acquisition', 'learning'] as ['developmentallearning'] is not use in
  Estonian
--------------------------------------------------------------------------------
WITHOUT PARENT
     i67146 estwn-et-51697-n|['akkommodatsioon']
     (OEWN equivalent: oewn-05763483-n|['accommodation'])

POSSIBLE PARENT(S):
     i67135 not in ESTWN

PARENTS FROM OTHER WORDNET(S):
     (i67146)->i67135 cow-05753207-n oewn-05761204-n|['developmentallearning']
      …
     (i67146)->i67135 fiwn-05753207-n oewn-05761204-n|['developmentallearning']
     (i67146)->i67135 oewn-05761204-n oewn-05761204-n|['developmentallearning']

POSSIBLE GRANDPARENT(S):
     i67133 not in ESTWN

GRANDPARENTS FROM OTHER WORDNET(S):
     (i67135)->i67133 cow-05752544-n oewn-05760541-n|['acquisition', 'learning']
      …
     (i67135)->i67133 oewn-05760541-n oewn-05760541-n|['acquisition', 'learning']
     (i67135)->i67133 wnja-05752544-n oewn-05760541-n|['acquisition', 'learning']
```

Appendix D. A parentless synset which received a parent that was not present in either the possible parents or grandparents

```
--------------------------------------------------------------------------------
13 # Correct parents for ['hüpnopeedia'] (['hypnopedia', 'sleep-learning']) is
   ['õppimine', 'tudeerimine', 'õpe'] (['acquisition', 'learning'])
--------------------------------------------------------------------------------
WITHOUT PARENT
     i40094 estwn-et-31344-n|['hüpnopeedia']
     (OEWN equivalent: oewn-00894218-n|['hypnopedia', 'sleep-learning'])

POSSIBLE PARENT(S):
     i40057 estwn-et-9263-n|['haridustegevus', 'õpetamine']

PARENTS FROM OTHER WORDNET(S):
     (i40094)->i40057cow-00887081-n oewn-00888759-n|['pedagogy','teaching', '…']
      …
     (i40094)->i40057 oewn-00888759-n oewn-00888759-n|['pedagogy', 'teaching', '…']
     (i40094)->i40057 wnja-00887081-n oewn-00888759-n|['pedagogy', 'teaching', '…']

POSSIBLE GRANDPARENT(S):
     i37550 estwn-et-677-n|['talitlus', 'tegevus', 'tegutsemine', '...']
     i68339 not in ESTWN
     i38639 not in ESTWN
     i36822 not in ESTWN

GRANDPARENTS FROM OTHER WORDNET(S):
     (i40057)->i38639 cow-00611433-n oewn-00612720-n|['education']
     (i40057)->i37550 estwn-et-677-n oewn-00408356-n|['activity']
      …
     (i40057)->i36822 plwn-pls-27941 oewn-00271644-n|['coaching', 'coachingjob']
     (i40057)->i68339 trwn-0103020 oewn-06008975-n|['science', 'scientificdiscipline']
```