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

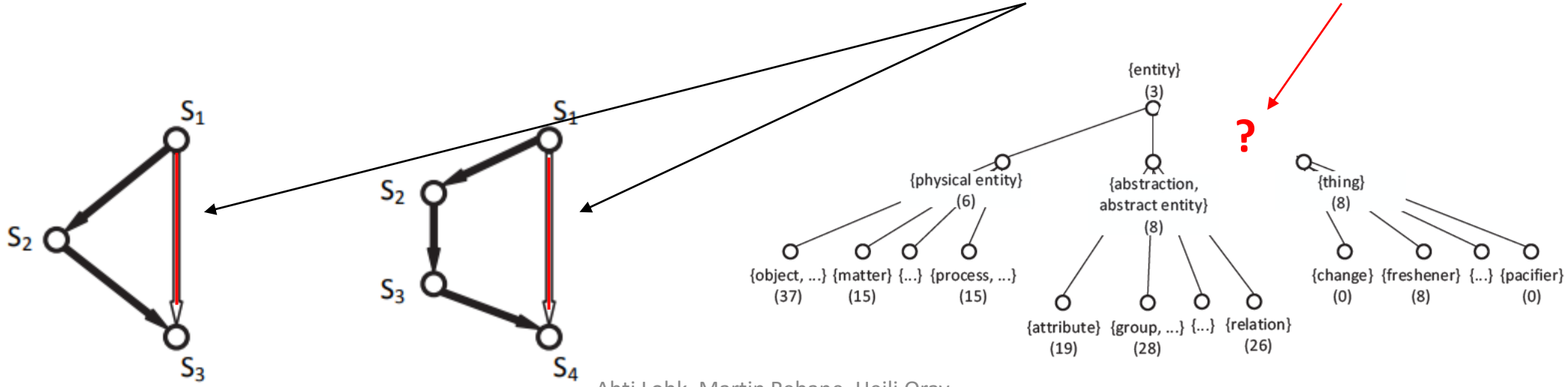
[2]School of Engineering, **University of Warwick**, Coventry, United Kingdom

[3] Department of Computer Science, **University of Tartu**, Tartu, Estonia

Background

- One of the main goals of wordnet development is to **make it accessible** while **en correctness**.

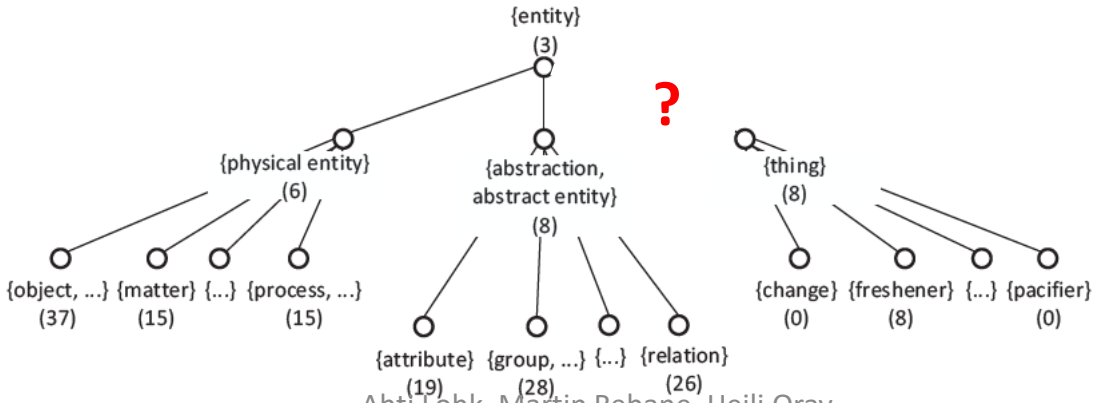
- Errors that require correction can be classified into **three categories**:
  - Formal – related to the source file structure or data presentation in it
  - Semantical – related to the wordnet semantics
  - **Structural** – related to the wordnet as a graph: **redundant** and **missing links**



Problem

- We focus only on **missing link problem**: specifically **parentless synsets.**
- That means that link is missing between the synset and its parent.

- **On the one hand,** the lexicographer may not know **which synset does not have a parent** and
- **on the other hand, which synset should be the parent** for the parentless synset.



Classification of parentless synsets

Parentless synset can be:

1. **Root synsets** (or unique beginners or top concepts)

   *{**entity:** n}, {**do**: v, **execute**: v, **perform**:v}, {**exist**: v, **be**: v} in OEWN (v 2.0)*

2. **Noun synsets with name content**

   *{**Washington**: n, **Washington DC:** n, ...} OEWN (v 2.0)*

3. **Other parentless synsets (with and without subordinates)**

   *{**separate**: v, **divide**: v}, {**globe**: n, **earth**: n, ...} OEWN (v 2.0)*

NB! We look only cases where the **synset does not have any parent**:
No multiple inheritance cases!

Wordnets and their parentless synsets

From their webpages

From Open Multilingual Wordnet webpage

Wordnet (language)	Parentless synsets	
	noun	verb
OEWN (English)	8	574
EstWN (Estonian)	190	13
Odenet (German)	3433	2583
ODWN (Dutch)	**0***	87
FinWN (Finnish)	172	559
LSG (Irish)	6000	1468
OWN-PT (Portuguese)	18577	7143
NTU-JPN (Japanese)	5766	420

*odwn-02532028-n | {**haring**}) **->** odwn- odwn-00001740-n | {**wezen, wezenlijkheid , iets, ...**}
oewn-02534659-n | {**herring, Clupeaharangus**} **->** oewn-00001740-n | {**entity**}

Purpose

Our purpose was to

find automatically

possible parents for **parentless synsets**

in **IS-A hierarchies** of

a **target wordnet**

by means of **source wordnets**.

Both target and source wordnets are in

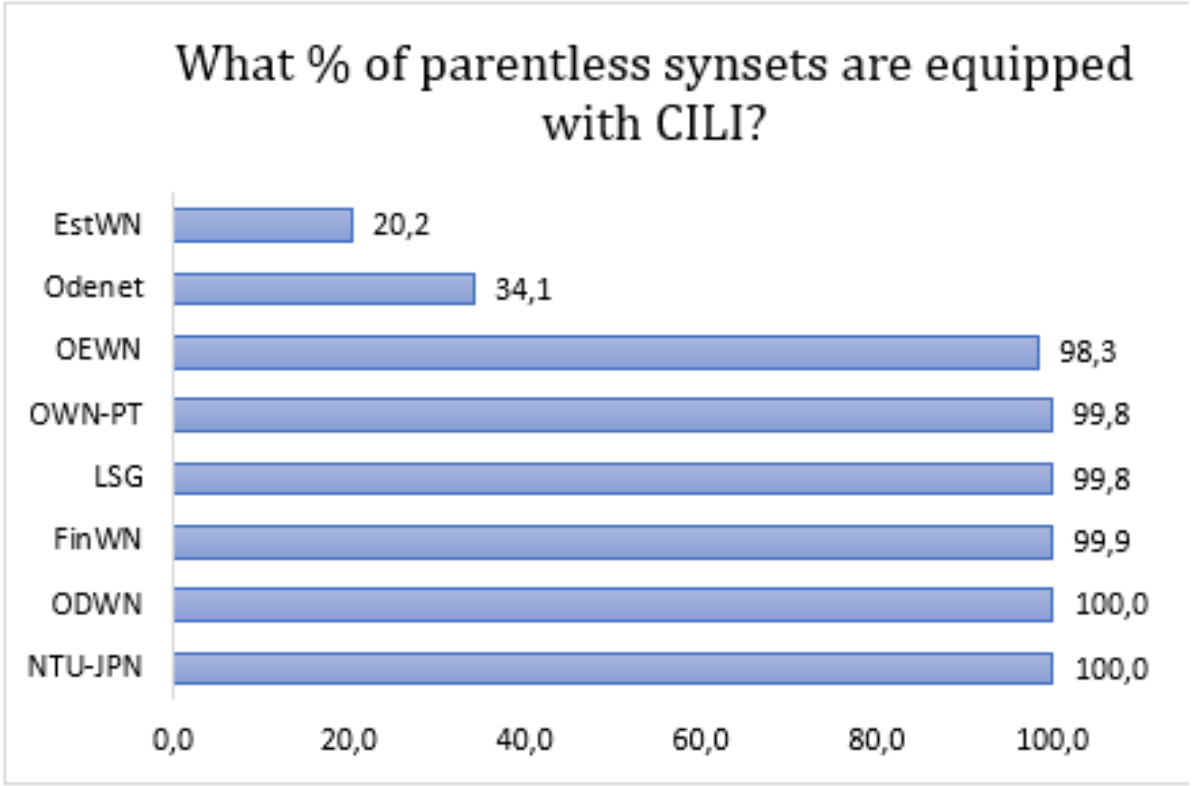
XML-format and

equipped with **Collaborative Interlingual Index** (CILI).

What % of parentless synsets are equipped with CILI?

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.



Simplified algorithm

for *synset* **in** ***target***_*wn.parentless_synsets*: *#verb and noun synsets*

If not *synset.content_is_name*:

for *wordnet* **in** ***source***_*wordnets*:

if *target_wn* **is not** *wordnet*:

parents = find_parents_via_CILI_from_other_wordnets(*synset*, *wordnet*)
grandparents = find_parents_via_CILI_from_other_wordnets(*synset*, *wordnet*)

Results presentation format


```
--------------------------------------------------------------------------------
26
--------------------------------------------------------------------------------
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 enwn-ens-367184 oewn-00040659-v|['makeup']
       (i21979)->i21972 ewn-00040659-v oewn-00040659-v|['makeup']
       (i21979)->i21972 fiwn-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 enwn-ens-367182 oewn-00040084-v|['neaten', 'groom']
       (i21972)->i30124 estwn-et-70-v oewn-01679858-v|['decorate','adorn', '...']
       (i21972)->i21970 ewn-00040084-v oewn-00040084-v|['neaten', 'groom']
       (i21972)->i21970 fiwn-00040353-v oewn-00040084-v|['neaten', 'groom']
       (i21972)->i21970 odwn-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']
```

Results

Wordnet (Language)	What % of parentless synsets are CILI-equipped	Nr of parentless synsets	Found parents	Found grand-parents	No parent, no grand-parent	% of no possible parent & grandpar ent
EstWN (Estonian)	20,2%	41	36	36	5	12,2%
Odenet (German)	34,1%	2 052	1 178	1 140	874	42,6%
OEWN (English)	98,3%	572	268	250	300	52,4%
OWN-PT (Portuguese)	99,8%	25 660	25 457	25 020	194	0,76%
LSG (Irish)	99,8%	7 454	7 337	7 258	114	1,53%
FinWN (Finnish)	99,9%	730	410	390	319	43,7%
ODWN (Dutch)	100,0%	87	38	31	49	56,30%
NTU-JPN (Japanese)	100,0%	5 950	5 211	5 192	739	12,4%

Case Study of Estonian Wordnet

- 41 parentless and CILI-equipped synsets (20,2% of parentless synsets)
  - 35 noun + 6 verb synsets
- After a closer examination, we found that the decisions that had to be made in solving them fell into four categories:

1. The **parentless synset turned out to be a root concept.** **7** **cases** (3 nouns + 4 verbs).
2. The **suggested possible parent was suitable** for the parentless synset. **10** **cases**.
3. The **suggested possible grandparent was suitable** for the parentless synset. **4** **cases**.
4. A parentless synset receives a parent that was not present in either the possible parents or grandparents. **20** **cases**.

Case study of Estonian wordnet

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 4:

Parentless synset:

{‘foneetika', '...'} *({' phonetics'})*

Suggested parent:

{'akustika', 'heliõpetus'} *({'acoustics'})*

Correct parent:

{lingvistika, '...'} *({'linguistics'})*

Argument:

The authoritative dictionary of the Estonian language (Sõnaveeb: https://sonaveeb.ee/) declares that phonetics is a part of linguistics.

Final thoughts

We proposed an approach to **automatically detect** the **possible parent for** cili-equipped **parentless synsets** using source wordnets.

1. The proposed **approach is universal**. That is, it can be applied to wordnets of all languages (with specific xml-format and cili-equipped synsets).
2. **Such an approach will definitely help increase the quality of wordnets**.
3. On the basis of the Estonian wordnet, we can say that the **correction requires the intervention of a lexicographer** and
4. that the **synsets** connected through CILI **in different languages may differ in their broader meaning**.
5. **In the future**, the proposed approach **could also be applied to synsets that have parents to detect possible inconsistencies**.

Questions?