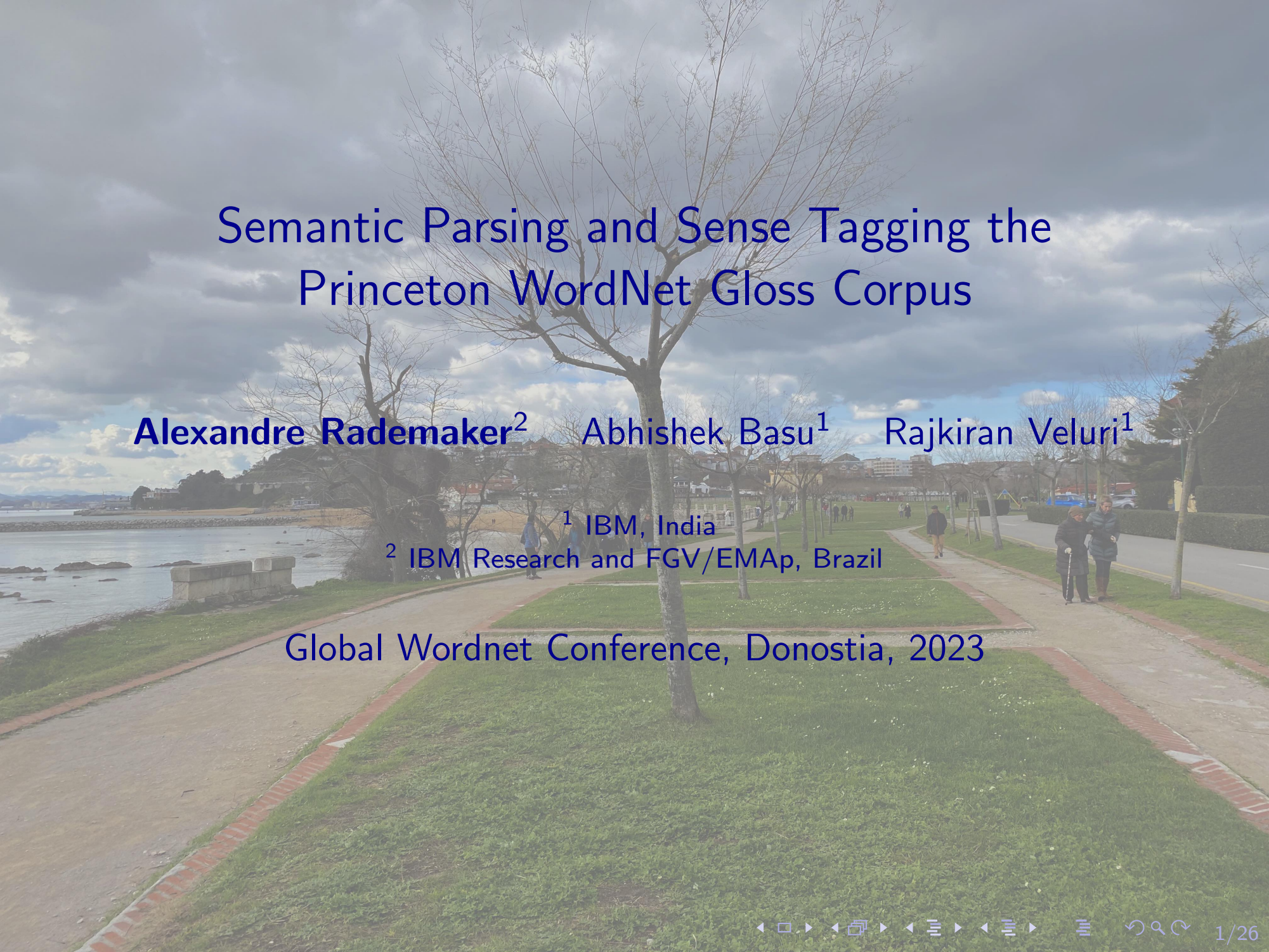


Semantic Parsing and Sense Tagging the Princeton WordNet Gloss Corpus

Alexandre Rademaker[2] Abhishek Basu[1] Rajkiran Veluri[1]

[1] IBM, India
[2] IBM Research and FGV/EMAp, Brazil

Global Wordnet Conference, Donostia, 2023

WordNet Glosses I

A synset gloss may contain a definition, one or more example sentences, or both. Glosses were introduced as redundancy to facilitate human understanding.

What is a butter? A hyponym of solid food and dairy product. A hypernym of lemon butter, drawn butter, stick, yak butter and beurre noisette.

But butter is "an edible emulsion of fat globules made by churning milk or cream"

The tennis problem, "a game played with **rackets** by two or four players who hit a **ball** back and forth over a **net** that divides the court"

WordNet Glosses II

Redundancy has its price! We usually pay missing consistency

WordNet 3.0 Pluto is "a small planet and.."

WordNet 3.1 Pluto is "a large asteroid ..."

but Tombaugh is in both releases "the astronomer who discovered the **planet** Pluto".

Also, the Roman mythology vs the Greek mythology already mentioned, see here.

WordNet Glosses III

Completeness is also relevant.

A Japanese oyster is "a large oyster native to Japan and introduced along the Pacific coast of the United States; a **candidate** for introduction in Chesapeake Bay"

A Algeripithecus minutus is "tiny (150 to 300 grams) extinct primate of 46 to 50 million years ago; fossils found in Algeria; considered by some authorities the leading **candidate** for the first anthropoid"

but what is candidate?

WordNet Glosses IV

The angry machine learning approaches need data! We don't have many sense tagged corpora.

- SemCor 226,040 but with a lot of problems (Fellbaum talk!). 16% of the WordNet senses (here)
- OMSTI silver data, obtained from English-Chinese parallel corpus
- Senseval and SemEval tasks ($<$ 2K sentences)

See http://lcl.uniroma1.it/wsdeval/

Processing the glosses I

The GlossTag project (GlossCor) was started in Princeton. Some annotation but not completed, but numbers do not match 100%! **We call it GlossTag 2008**

Definitions and Examples are demarcated, tokenized and PoS tagged (only definitions). Some spans were annotated with semantic classes: dates, time, number, currency, math expressions, etc. Some spans are marked as auxiliary information (domain classification, verb arguments or contents that are secondary to the main sense of the synset (ignored to sense annotation).

Data has been used by tools like UKB (graph-based word sense disambiguation library).

Processing the glosses II

We also know about other projects that explore the knowledge from the glosses.

The eXtended WordNet from University of Texas at Dallas (website not available, based on WordNet 1.6 and 2.0). LF constructed from transformation rules applied to the syntactic analysis.

Standoff files from Priceton with logical forms from glosses. Generated by USC/ISI, California in 2006-2007. Also transformation rules, LFToolkit, applied to the output of Charniak syntactic parser.

Others?

Processing the glosses III

Our project started in 2019 (paper). The aim is to continue the annotation, fixing mistakes and adding extra layers to help on annotation. Thanks Fellbaum for the suggestion and directions!
We call it GlossTag 2019

We develop an annotation interface on top of Emacs sensetion.el

The annotation of verbs is hard, a syntactic/semantic parsing with an holistic interpretation of the sentence may help.

Processing the glosses IV

166,820	auto
664,175	ignore
334,533	man
449,967	un

We allow multiple senses whenever we can't distinguish the senses. We have 40% of the WordNet senses mentioned at least once.

MWE: $56,859$ tokens: $1,631,341$ sentences: $165,994$ (definitions: $117,658$ examples $48,336$).

Processing the glosses V

In this paper . . .

We revised tokenization issues and the demarcation of definitions and examples. We also revised the quoted examples, moving the source of the text (author or reference) to metadata.

We parsed the sentences with English Resource Grammar and combine the sense annotation with the semantic representation. Adding PoS to examples. Hopefully more consistent semantic representation of texts. We call it **GlossTag 2022**.

English Resource Grammar I

The English Resource Grammar (ERG) is a broad-coverage, general-purpose, linguistically precise HPSG computational grammar. It can map running English text to highly normalized logical-form representations of meaning (MRS).

After creating the profiles with 2000 sentences each, we processed them with the Ace parser in a cluster in 30 minutes. For each sentence, we asked for the top-best analysis of ERG. From 165,976 sentences; only 5,282 (2%) were not parsed by ERG. Using some heuristics (e.g. 'get the votes of X'), 600 more sentences.

Sentences are typically ambiguous, we had hundreds or thousands of readings for some sentences. Preliminar evaluation gives us F1 80% for the first analysis be the expected one, future work aim to manually treebanking all sentences using FFTB tool.

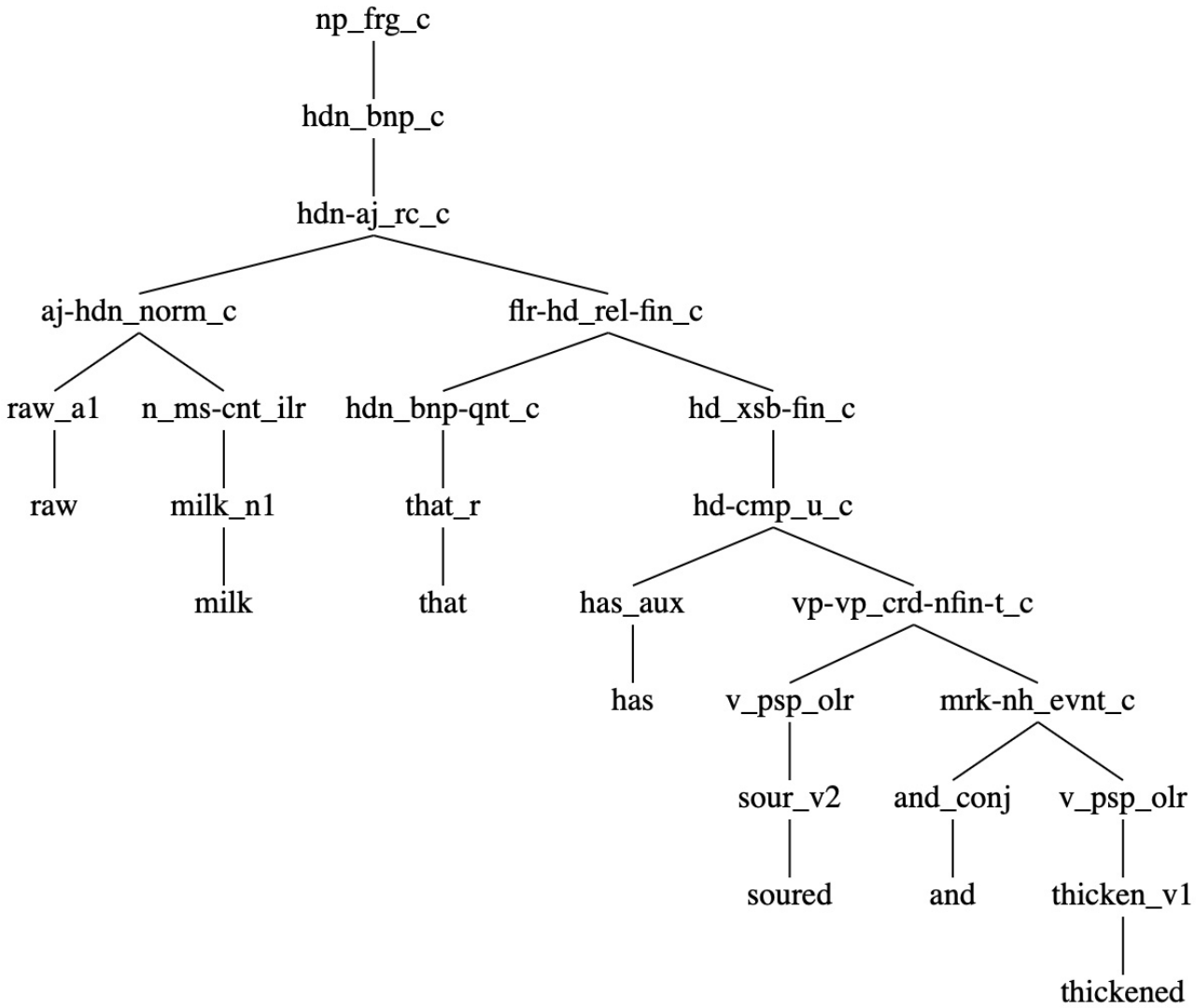
An Example I

clabber: raw milk that has soured and thickened

```
# text = raw milk that has soured and thickened
# id = 07850219-n-1
# type = def
1       _       glob|a  _       raw_milk%1      raw_milk%1:13:00::      auto    _
2       raw     cf|a    JJ      raw%1|raw%3     _       un      0:3
3       milk    cf|a    NN      milk%1|milk%2   _       un      4:8
4       that    wf      WDT     that    _       ignore  9:13
5       has     wf      VBZ     have%2  _       un      14:17
6       soured  wf      VBN     sour%2|soured%3 soured%3:00:00::        man     18:24
7       and     wf      CC      and     _       ignore  25:28
8       thickened       wf      VBN     thicken%2|thickened%3   thicken%2:30:00::       man     29:38
```

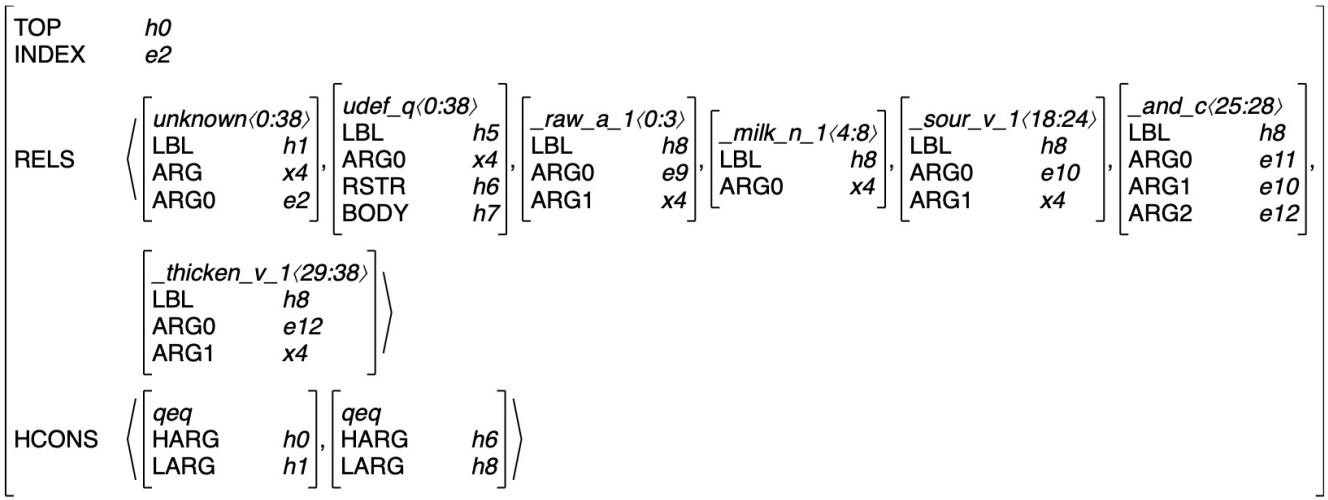
An Example II

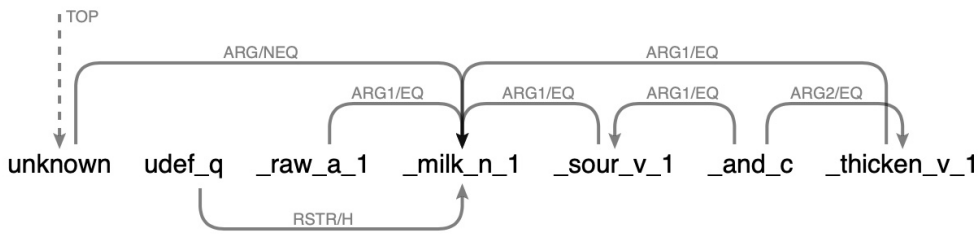
clabber: raw milk that has soured and thickened



An Example III

clabber: raw milk that has soured and thickened





See here and here from palmatifid.

Speeding up the annotations I

Manual word sense disambiguation (WSD) is an arduous task. One non-native speaker annotator doing it manually from the last 4 years (not full-time).

Many techniques for automatic WSD are being investigated: graph-based (or knowledge-based), supervised and unsupervised machine learning methods.

Automatic annotation would allow us to provide intermediary releases of the data (silver versions).

Two-ways . . . Remember that GlossTag 2008 was already used by UKB tool (graph-based WSD) and to training supervised WSD algorithms replacing the SemCor.

Speeding up the annotations II

We used UKB, data was transformed into UKB input for: (1) evaluation UKB performance; (2) complete annotation. From palmatifid

```
# text = of a leaf shape; palmately cleft rather than lobed
# id = 02173264-a
# type = def
1       wf      ignore  0:2     IN      of      of      _
2       wf      ignore  3:4     DT      a       a       _
3       glob|a  auto    _       _       _       leaf_shape%1    leaf_shape%1:25:00::
4       cf|a    un      5:9     NN      leaf    leaf%1|leaf%2   _
5       cf|a    un      10:15   NN      shape   shape%1|shape%2 _
6       wf      ignore  15:16   :       ;       _       _
7       wf      auto    17:26   RB      palmately       palmately%4     palmately%4:02:00::
8       wf      man     27:32   VBN     cleft   cleft%1|cleave%2|cleft%3        cleft%5:00:00:compound:00
9       wf      un      33:39   RB      rather  rather%4        _
10      wf      ignore  40:44   IN      than    than    _
11      wf      man     45:50   JJ      lobed   lob%2|lobed%3   lob%2:35:00::
```

```
ctx-02173264-a/a
leaf_shape#n#w4#1#1 palmately#r#w7#1#1 cleave#v#w8#1#1 rather#r#w9#1#1 lob#v#w11#1#1 02173264-a#a#fake1#2#1

ctx-02173264-a/b
leaf_shape#n#w4#1#1 palmately#r#w7#1#1 cleave#v#w8#1#1 rather#r#w9#1#1 lob#v#w11#1#1
```

We try with and without extra context. We compare results for annotated words with the annotations.

Speeding up the annotations III

Results

	Total	# (a)	# (b)	% (a)	% (b)
All	442782	413546	374648	93.39	84.61
Noun	329692	308245	287033	93.49	87.06
Adj	64298	60591	52008	94.23	80.89
Verb	41520	37832	29529	91.11	71.12
Adv	7272	6878	6078	94.58	83.58

For automatically complete annotation, words already annotated would increase UKB performance.

Projecting annotations to WordNet 3.1 I

WordNet 3.1 contains minor fixes in the texts of the glosses and removed many newly considered offensive words.

676 senses were added and 382 removed, some WordNet 3.0 senses have moved between synsets, or the corresponding synsets were changed in WordNet 3.1.

Mapping the GlossTag to Wordnet 3.1 would allow mappings to other lexical resources like VerbNet and PropBank, enrich information about verb valences. Also allowing extra mapping from the ERG lexicon.

Syntactic restrictions and VN classes can facilitate sense annotation of the verbs too.

Projecting annotations to WordNet 3.1 II

We need to identify which definitions and examples are removed, preserved or created from WN 3.0 to 3.1.

New sentences need to be processed by ERG and prepared for manual annotation from scratch. Removed sentences are just removed (or not!)

Next we need to check the annotations for sentences preserved from 3.0 to 3.1.

We need to consider the annotated words only. We found cases where a given sense key got a different meaning in WordNet 3.1.

Projecting annotations to WordNet 3.1 III

The sense `pluto%1:17:00::` for the word 'Pluto' has changed.

In 3.1 it is part of the synset "a large asteroid that was once thought to be the farthest known planet from the sun; it has an elliptical orbit"

In 3.0 it was "a small planet and the farthest known planet from the sun; it has the most elliptical orbit of all the planets"

since the definition changed, the relations also changed. This is a case of a sense key that should not be reused.

Projecting annotations to WordNet 3.1 IV

Another sense of ‘Pluto’ in WordNet 3.0 is part of the synset “(Greek mythology) the god of the underworld in ancient mythology; brother of Zeus and husband of Persephone”

In WordNet 3.0, Pluto was defined as a synonym of Hades, but WordNet 3.1 revised that definition making it part of Roman mythology and a counterpart of Hades.

There are eight occurrences of ‘pluto’ in the WordNet 3.0 sentences.

Projecting annotations to WordNet 3.1 V

Another challenge arises when a new sense is introduced in WordNet 3.1, and some words in the sentences could be better annotated with the new sense.

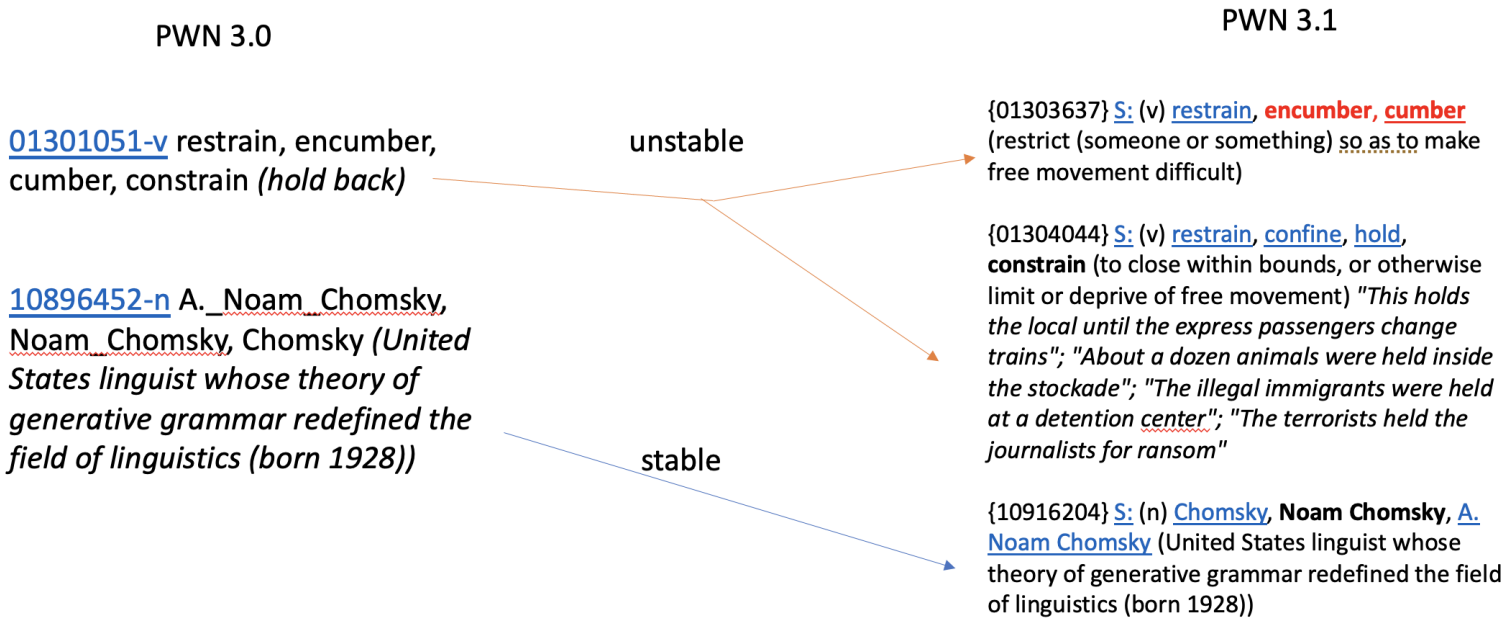
The word 'technology' in 3.1 has the sense "machinery and equipment developed from engineering or other applied sciences" (here). But in 3.0 we only have two senses (here)

We found 53 instances of the word 'technology' in sentences (definitions and examples), and the new sense from WordNet 3.1 may be more appropriate for some of them.

The synset 08343534-n "has procured nuclear **technology** and delivery capabilities" is one example.

Projecting annotations to WordNet 3.1 VI

We are refining the idea of sense stability.



https://github.com/globalwordnet/cili/pull/17

Final Thoughts I

The project is hosted in the https://github.com/own-pt/glosstag

We aim to build a web interface to browse the data, possible improvement in the http://openwordnet-pt.org.

We need to make code available to the reproducibility of the experiments presented here.

We need to improve our annotation tool, fixed dependencies.

We need to finish the annotation and treebank the ERG analyses. Ongoing work. Define a proper workflow to combine the two layers of annotation.

We plan to experiment with alternative WSD methods.

Final Thoughts II

This work is part of our effort in expanding and improving WordNet-like resources in an application-driven and domain-specific way.

Finally, we need to finish the migration to WordNet 3.1 before forking it from the Princeton official release (or further mapping to Open English Wordnet, http://en-word.net) for changes driven by the annotation.

How WSD methods could benefit from the new GlossTag 2022?


Thank You !