WordNets and TEI–LEX

John P. McCrae, Thierry Declerck
Global WordNet Grid
Princeton WordNet

EuroWordNet

BalkaNet

Multi WordNet

Indo WordNet

Open WordNet PT

Asian WordNet

Open Multilingual WordNet

3.0

2012
Problems

- Different relations used
- Different interpretations of relations
- Different definitions of synonymy (tight, loose)
- Anglo–saxon view of Princeton WordNet
- Updates to Princeton break all other wordnets
- Different coverage

“25 synsets shared from 117,677 (0%)” - Open Multilingual WordNet (Bond)
All open source wordnets linked to a single ILI

- Merge of concepts across all languages
- Princeton WordNet versions linked to ILI
- Adaptable by the open source wordnet community.
- Available as LOD and downloadable with open source license: CC–BY, CC–BY–SA.
How to define a concept

1. A unique, permanent URI

2. A proper (!) English gloss

3. Linked to at least one other synset
Cross-lingual Mapping

Lehrer

Lehrerin

Teacher

Professore

Professoressa
Interlingual Mapping

Lehrer

Lehrerin

Teacher

i123456

Professore

Professoressa

i123457
WordNet LMF is ‘LMF–like’ XML format

Converter/validator
http://server1.nlp.insight-centre.org/gwn-converter/

WordNet JSON is JSON–LD based representation

Isomorphic

RDF/XML

Turtle

SPARQL

A profile of OntoLex
The GWN WordNet Formats
GlobalWordNet Structure

LexicalResource

Lexicon+

LexicalEntry+

Lemma

Form*

Sense*

SenseRelation*

Synset*

Definition*

SynsetRelation*
<xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE LexicalResource SYSTEM "http://globalwordnet.github.io/schemas/ WN-LMF-1.0.dtd">
<LexicalResource xmlns:dc="http://purl.org/dc/elements/1.1/">
  Root element
</LexicalResource>

XML Declaration

DTD for validation

Dublin Core Namespace
<Lexicon id="example-en"
label="Example wordnet (English)"
version="1.0"
language="en"
email="john@mccr.ae"
citation="CILI: the Collaborative ...
license="https://creativecommons.org/publicdomain/zero/1.0/
url="http://globalwordnet.github.io/schemas/
dc:publisher="Global Wordnet Association">
<LexicalEntry id="w1">

  <Lemma writtenForm="paternal grandfather" partOfSpeech="n"/>

  <Sense id="example-en-1-n-1" synset="example-en-1-n">
    <SenseRelation relType="derivation" target="example-en-10161911-n-1"/>
  </Sense>

</LexicalEntry>
<Synset id="example-en-10161911-n"
    ili="i90287"
    partOfSpeech="n">  
    <Definition language="en">
        the father of your father or mother
    </Definition>
</Synset>

<SynsetRelation relType="hyponym"
    target="example-en-10162692-n"/>
**Lexical vs. Encyclopedic**

**Yellow** (in a dictionary)
- Is a verb, noun and adjective
- Secondary synonyms: cowardly, warning (especially in soccer)

**Yellow** (in an encyclopedia)
- A colour
- 2 books
- 8 Films or TV shows
- 4 songs
- A butterfly
Princeton WordNet

- Is a **lexical** resource with some **encyclopedic** information
- This information is quite biased to Anglo-Saxon, American and even North Eastern US context.

Paterson, NJ, USA
- Pop: 147,000
- In Princeton WordNet

Kawasaki, Japan
- Pop: 1,500,000
- Not In Princeton WordNet
Wikipedia

- Wikipedia is open and most-widely used encyclopedia
- Many lexical concepts are included, e.g.,
  - Play (activity)
Overlap between lexical and encyclopedic resources
Linking between resource types

Lexical

Encyclopedic

Gold-standard, manual linking
All but 77 (1.0%) synsets have at least one candidate
Average of 21.6 Wikipedia articles per synset
Category Matches (based on Suchanek’s YAGO)

Paris, capital of France — the capital and largest city of France...

Prague, Praha, Czech capital — the capital and largest city of the Czech Republic
Length based matches

Longer matching strings are less ambiguous
\[ s(h,c) = \sum_{(i,a) \in P(h,c)} \sigma(i,a) \]

\[ \sigma(i,a) = l(i,a) - \alpha \text{ if } (i,a) \text{ is unambiguous} \]

\[ \sigma(i,a) = -\beta \text{ otherwise} \]
Manually checked linking of 7,687 synsets to Wikipedia available at

https://github.com/jmccrae/wn–wiki–instances
Beyond Princeton WordNet: Colloquial WordNet
Princeton WordNet is infrequently released

- Version 2.1: Mar 2005
- Version 3.0: Dec 2006
- Version 3.1: Nov 2012 (fewer synsets)
Tweeting  

Spreading  

Exit
Tweeting

Manspreading

Brexit
Trump and his cronies attempt to delegitimate the media should concern everyone.
Now will you admit goal has been to delegitimate 1st Black president?
Spicer delegitimizes the press. But can you delegitimate 2.5 million protesters?
Prior to taking office Trump has managed to delegitimate his own presidency far more than he ever
how the very media trump is attempting to delegitimate gave trump all the coverage in the world
RT On the radio ABC News tries to delegitimate Our President even after he says in two
RT The democrats continually try to delegimize President Trump but they never fail to end
yes they want to delegimize Palestinians and actually stole their heritage identity to
shows are fun and all but they ultimately delegimize dance as a career
start planting forged emails as an attempt to delegimize the leaks