An Overview of the Coptic Wordnet Project

Laura Slaughter\textsuperscript{1}, So Miyagawa\textsuperscript{3}, Luis Morgado da Costa\textsuperscript{2}, Heike Behlmer\textsuperscript{3}, Amir Zeldes\textsuperscript{4}, Hugo Lundhaug\textsuperscript{1}

\textsuperscript{1}University of Oslo, Norway
\textsuperscript{2}Nanyang Technological University, Singapore
\textsuperscript{3}Georg-August-Universität Göttingen, Germany
\textsuperscript{4}Georgetown University, USA
Contents
20 minutes, 5 topics

• Brief overview of wordnets: What is a wordnet?
• Wordnets & CILI: How does a wordnet differ from a dictionary or lexicon?
• Coptic Wordnet: Where are the files and how do I get started?
• Applications of wordnets: What can I do with it?
• Improvements and future plans: Who is maintaining it and what are the future plans?
Brief Overview of Wordnets

What is a wordnet?

• Started in mid-1980s by a psychologist, George A. Miller
• Establish psycholinguistics as an independent field of research
Princeton WordNet (English)

The first WordNet - Princeton WordNet (PWN)

- large lexical database of English
- nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets)
- each synset expresses a distinct concept
- synsets are interlinked by means of conceptual-sematic and lexical relations
- covers over 117,000 concepts and over 150,000 English words
Lexical Categories

• Words are assigned to lexical categories: nouns, verbs, adjectives, adverbs

• Also other classes of concepts came later: determiners and other function words (pronouns, auxiliary words, conjunctions (coming soon))

• Prepositions (still out)- but point of discussion

• Lexical relations between word forms
Semantic Relationships

- Semantic relations hold between word meanings.
- Inferred relations - all the male children of my mother are my brothers; shark is a fish and fish is an animal, then shark is an animal
- Nouns
  - hypernyms (canine is a hypernym of dog)
  - hyponyms (dog is a hyponym of canine)
  - meronym (window is a meronym of building)
  - holonym (building is a holonym of window)
- Verbs
  - hypernyms (to perceive is a hypernym of to listen)
  - troponym (to stroll is a troponym of to walk)
  - entailment (to sleep is entailed by to snore)
WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: dance Search WordNet

Display Options: (Select option to change)  [ ] Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations
Display options for sense: <lexical filename > (gloss) "an example sentence"

Noun
- <noun.communication> S: (n) dance (an artistic form of nonverbal communication)
- <noun.group> S: (n) dance (a party of people assembled for dancing)
  - direct hypernym / full hypernym
    - <noun.group> S: (n) ball (the people assembled at a lavish formal dance) "the ball was already emptying out before the fire alarm sounded"
  - direct hypernym / inherited hypernym / sister term
  - derivationally related form
- <noun.act> S: (n) dancing, dance, terpsichore, saltation (taking a series of rhythmic steps and movements) in time to music
- <noun.event> S: (n) dance (a party for social dancing)

Verb
- <verb.motion> S: (v) dance (move in a graceful and rhythmic way) "The young girl danced into the room"
- <verb.creation> S: (v) dance, trip the light fantastic, trip the light fantastic (move in a pattern; usually to musical accompaniment; do or perform a dance) "My husband and I like to dance at home to the radio"
  - direct troponym / full troponym
  - verb group
    - <verb.motion> S: (v) dance (move in a graceful and rhythmic way) "The young girl danced into the room"
  - domain category
    - <noun.act> S: (n) dancing, dance, terpsichore, saltation (taking a series of rhythmic steps and movements) in time to music
  - direct hypernym / inherited hypernym / sister term
  - derivationally related form
  - sentence frame
- <verb.motion> S: (v) dance (skip, leap, or move up and down or sideways) "Dancing flames", "The children danced with joy"
Also:
lex. relations
Antonym, Pertainym, Participle

domain categories

simple sentence of verb use
e.g. Somebody ----s something
Readings & Resources


- Useful links:
  - Online Princeton Wordnet
    - https://wordnet.princeton.edu/
  - Joining Open Multilingual Wordnet
    - https://lr.soh.ntu.edu.sg/omw/join
  - Schemas
    - http://globalwordnet.github.io/schemas/
Wordnets and CILI

Wordnet vs Lexicon

#1 Wordnet introduces the concept of a concept (synset), synset IDs for a set of synonyms, meaning is not based on the definition written but through the synonym list

#2 Wordnet links these concepts and provides a hierarchy to traverse

#3 Wordnet provides the potential to link to many other languages beyond the translations in a dictionary
Other Wordnets: Multilingual Connection

- Wordnets are connected by linking concepts (synset IDs)
- Global Wordnet Association maintains a list (not complete)
  http://globalwordnet.org/wordnets-in-the-world/
- Ancient Greek
- Latin
- Arabic + Quranic Arabic WordNet
- Hebrew
- Middle Ancient Chinese
- Sanskrit
GWA: Global Wordnet Association

• To promote the standardization of the specification of wordnets for all languages in the world, including:
  
  • the standardization of the **Common Inter-Lingual-Index (CILI)** for inter-linking the wordnets of different languages, as a universal index of meaning
  
  • the development of a common representation for wordnet data
  
  • studies in lexical semantics- wordnet is output of this worldwide research

• To promote the development of guidelines and methodologies for building wordnets in new languages

• To promote the development of explicit criteria and definitions for verifying the relations in any language
Common Inter-lingual Index (CILI)

• A flat list of concepts, no lemmas and no structure, officially has no parts of speech; exists in relation with other wordnets

• Doesn’t presuppose a hierarchical structure, or that a concept exists in all languages

• CILI IDs are persistent, only deprecate or supercede, never change the meaning of a concept

- Artificial Classes versus Lexicalized Classes:
  *instrumentality; natural object*
- Lexicalization differences of classes:
  *container and artifact (object) are not lexicalized in Dutch*
Religious/Theology Related Concepts— CILI


- Concepts related to theology in the PWN are Anglo- Christian skewed. I worked with Wenjie - his main interests were in Buddhism, to examine concepts in PWN

  - Examine available synsets in PWN that are relevant to support scholarly work on sacred texts
    - *Sunyata* in sanskrit wordnet, *emptiness* in PWN: *offering*

  - Could be useful to catalog specific named entities that are found within sacred texts and compare them with those available in existing wordnets- places, individuals, supernatural beings, mythical beasts, or objects with magical properties; WordNet is linked to DBpedia

  - Should annotate with sacred texts, PWN never had theological texts so it is incomplete
Extensions - more info the join OMW page

You can add variations of lemmas, including orthographic variations and transliterations, as shown below. You can have various classes of transliteration, and if they are automatically generated, you can give them a confidence score.

<LexicalEntry id="w613347">
  <Lemma writtenForm="动物沟通" partOfSpeech="n" script="Hans"/>
  <Form writtenForm="dòngwùgōutōng" script="Latn-pinyin">
    <Tag category="transliteration">pinyin</Tag>
    <Tag category="confidence">0.77</Tag>
  </Form>
  <Form writtenForm="dong4wu4goultong1" script="Latn-pinyin">
    <Tag category="transliteration">pinyin1</Tag>
    <Tag category="confidence">0.77</Tag>
  </Form>
  <Form writtenForm="dongwugoutong" script="Latn-pinyin">
    <Tag category="transliteration">pinyin</Tag>
    <Tag category="confidence">0.77</Tag>
  </Form>
</LexicalEntry>

- Tags for orthographic variation and transliteration, regional dialects
- Diachronic meaning change tagging
- Wordnet structures are not pre-defined, so we can contribute to add additional tags as needed, for example “region” or “period”
Coptic Wordnet

- Slides and link to GitHub are on the Coptic Wordnet page at UiO

- https://www.tf.uio.no/english/research/groups/coptic-texts-and-manuscripts/coptic-wordnet/

Release

- Creative Commons Attribution 4.0 International License (CC BY 4.0)
- GitHub (coptic-wordnet):
  - OMW tsv files to be used with Python NLTK
  - WN-LMF format (for CILI)
Automated Construction Approach

- Dictionaries (Crum, 1939. Oxford)
  - Coptic Dictionary Online (English, French, German)
  - MARCION (English, French, German, Czech, Greek)
  - Dictionary of Greek Loan Words in Coptic (DDGLC) (Ancient Greek, English)
- Use word-aligned dictionaries between Coptic and the five other languages: English, Greek (modern + ancient), French, German, and Czech
- Naive algorithm, Multilingual Sense Intersection (Luis’s project)
**Multilingual Sense Intersection**

- Ranking procedure: What is the likelihood that a coptic lemma $x$ belongs to each of the meanings (concept/synset) provided by the aligned lemmas in each language.

- Concepts suggested by more languages have a higher chance of being correct; the more matching lemmas in the set, the more likely it is correct.

- Within concepts having same # of language matches, metrics to rank (ranking score):
  - # of individual lemmas matched in each language
  - Part-of-speech congruency,
  - Lemma-concept saturation level (i.e. for each concept being suggested, what percentage of lemmas was seen to inform the same concept, per language)

- Full description of this will be reported in a future paper (author: Luis Morgado da Costa)
Evaluation: Manual Checking

<table>
<thead>
<tr>
<th>0/1/7</th>
<th>No. Langs</th>
<th>Candidate Lemma</th>
<th>Matched Translations</th>
<th>Candidate Synset</th>
<th>English Lemmas, Definitions and Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>丢失</td>
<td>‘fra</td>
<td>saisir</td>
<td>n’, ‘fra</td>
</tr>
</tbody>
</table>

- **0,1,**?

1: attesting the existence of the candidate sense, (i.e. the lemma was known to include the meaning proposed by the candidate synset)

0: rejecting the possibility that the candidate lemma could be used in the candidate sense,

and the question mark ? “uncertain”
## Coptic Wordnet Coverage

<table>
<thead>
<tr>
<th>POS</th>
<th>No. synsets</th>
<th>No. senses</th>
</tr>
</thead>
<tbody>
<tr>
<td>nouns</td>
<td>13,904</td>
<td>97,527</td>
</tr>
<tr>
<td>verbs</td>
<td>7,491</td>
<td>92,019</td>
</tr>
<tr>
<td>adjective</td>
<td>3,488</td>
<td>20,723</td>
</tr>
<tr>
<td>satellite adj</td>
<td>229</td>
<td>587</td>
</tr>
<tr>
<td>adverb</td>
<td>737</td>
<td>7,373</td>
</tr>
<tr>
<td>non-referential</td>
<td>22</td>
<td>448</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>25,871</strong></td>
<td><strong>218,677</strong></td>
</tr>
</tbody>
</table>
Notes About Coverage

- Senses distributed among 25,871 synsets, fairly well across different parts of speech
  - Cover about 77.4% of the list of 5000 core word senses in PWN, the usual measure for coverage for a WN
- 7 senses per nominal synset
- 12.2 senses per verbal synset
- May have so many senses because a single lemma can take many forms and spelling variation (future work to explore)
Union and Intersection

- **Union**: Percentage of senses accepted by either of the reviewers (either said “1”)
  - Always rewards the reviewer who claims to know the existence of a sense
- **Intersection**: Percentage of senses accepted by both reviewers (both said “1”)
  - When one said “uncertain”, we looked at the other reviewer’s responses and counted it if it was “1”

<table>
<thead>
<tr>
<th>No. Langs</th>
<th>Correct (%) Union</th>
<th>Correct (%) Intersect.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(n=119) 25%</td>
<td>7%</td>
</tr>
<tr>
<td>2</td>
<td>(n=134) 89%</td>
<td>49%</td>
</tr>
<tr>
<td>3</td>
<td>(n=40) 98%</td>
<td>63%</td>
</tr>
<tr>
<td>4</td>
<td>(n=7) 100%</td>
<td>100%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>62% (n=300)</strong></td>
<td><strong>34% (n=300)</strong></td>
</tr>
</tbody>
</table>
Applications of Wordnets

• What can you do with a wordnet that you cannot do with a lexicon?

• #1 Multilingual possibilities- can generate a new lexicon to another language

• #2 Robust language documentation, language study from linguistic perspective- 30+ years in the making of wordnet structure; big return on investment to use the structure provided; though making a wordnet is a long term commitment
WordNet Applications (in general)

• A wordnet should be designed to facilitate automatic text analysis
  • #1 use is Word Sense Disambiguation (WSD)
    • e.g. bass is (1) a fish, (2) tones of low frequency
  • Information retrieval, text mining, summarization, machine translation, sentiment analysis, question-answering, language generation
TRACER- Text Reuse, WN Hierarchy

• Text reuse, of interest to So because of his dissertation
  • Traverse the hierarchy to find replacements
• Uses BabelNet (WordNets are used)
• Except for Ancient Greek, uses a flat file result of query to the AGWN database
• Results: found that it isn’t enough to use hypernym to expand query, but needs to traverse the graph of relationships to find some optimal distance
Word-Sense Tagging

• Multilingual Comparisons: Use CILI to create new contextualized concepts that would be interesting for other languages

• e.g. What is a “bank” for Coptic speakers?

• Diachronic change/conceptual shift: Concept alignments between other languages can be interesting for Coptic’s context, along with diachronic change information, looking towards measuring “concept shift” or flow of ideas over time
Work-in-Progress (Immediate Plans)

• Link Coptic WN to the Coptic Dictionary Online (CDO), following practices of Linked Open Data connected to CDO entries

• Larger evaluation where we have balanced the sample (w/number of languages)
  • Determine a confidence score for each sense, filter the size of the wordnet depending on the task

• Text Reuse: TRACER application to find text reuse, replacing words, the WN provides synonyms, hypernyms, hyponyms, co-hyponyms (future paper, hierarchical traversal)

• Create and annotate a sense-tagged corpus, alongside the wordnet, gain word frequency information, test for coverage and review concepts in context

• Needs Funding! - #1 Priority, currently maintained at UiO, Oslo (me) and with the help of Luis (NTU, Singapore) and So (Kansai University, Japan)
Wish List

• Additional work with TRACER and other text reuse applications

• Digital Humanities work within the domain of Theology, Digital Philology, CILI theology concepts, create better resource for working on scholarly texts

• Provide a tool for the study of Coptic-related language evolution, including the problems of concept drift (how concepts travel through space and time)

• Multilingual projects. Coptic WN, AGWN- these are places to start, though AGWN seems to be abandoned; or Egyptian

• Instructions for implementation, providing Python scripts for defined tasks, user-friendly visual way to work with the WN
Thanks!

- Thanks go to:
- So Miyagawa and Amir Zeldes for organizing the workshop
- Tonio Sebastian Richter and Katrin John from DDGLC
- Adam Rambousek for access to in-progress Czech Wordnet.
- Milan Konvicka for MARCION