

An Overview of the Coptic Wordnet Project

**Laura Slaughter¹, So Miyagawa³, Luis Morgado da Costa²,
Heike Behlmer³, Amir Zeldes⁴, Hugo Lundhaug¹**

¹University of Oslo, Norway

²Nanyang Technological University, Singapore

³Georg-August-Universität Göttingen, Germany

⁴Georgetown 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?

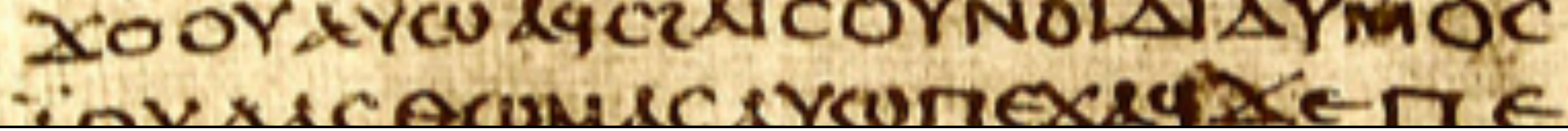
- Started in mid-1980s by a psychologist, George A. Miller
- Establish psycholinguistics as an independent field of research



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-semantic and lexical relations
- covers over 117,000 concepts and over 150,000 English words



Lexical Categories

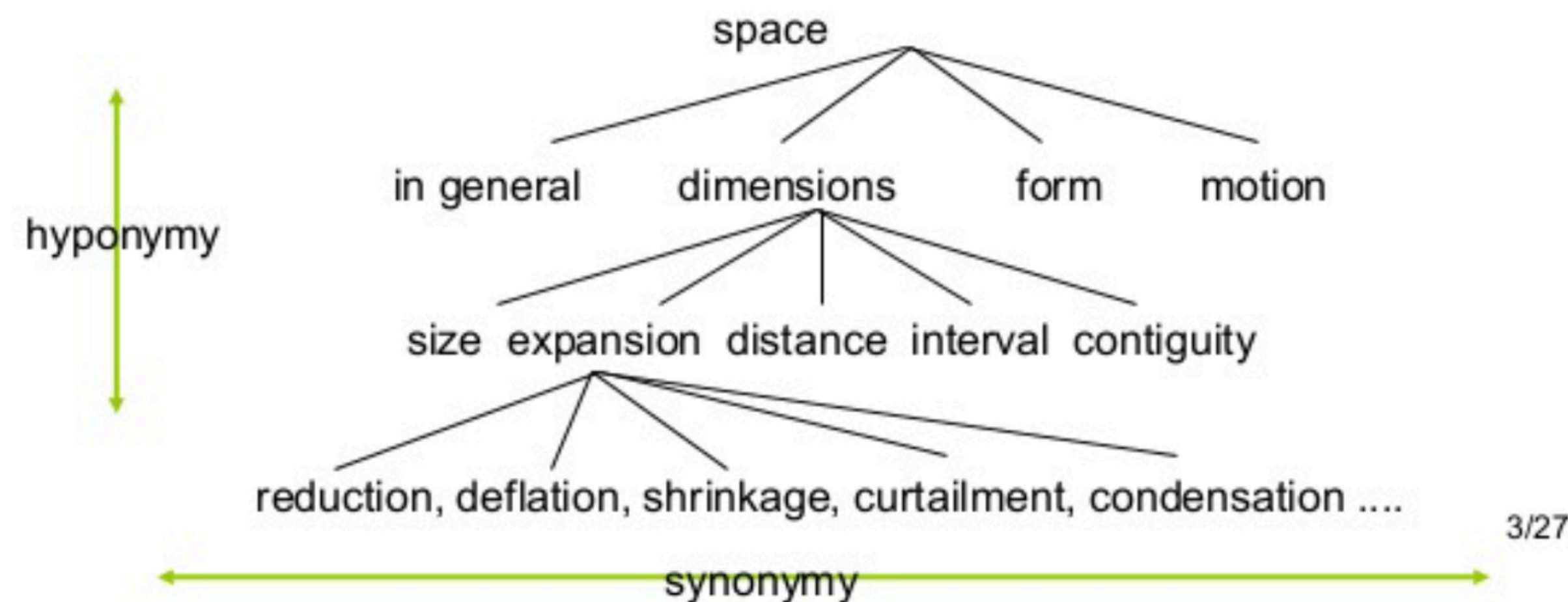
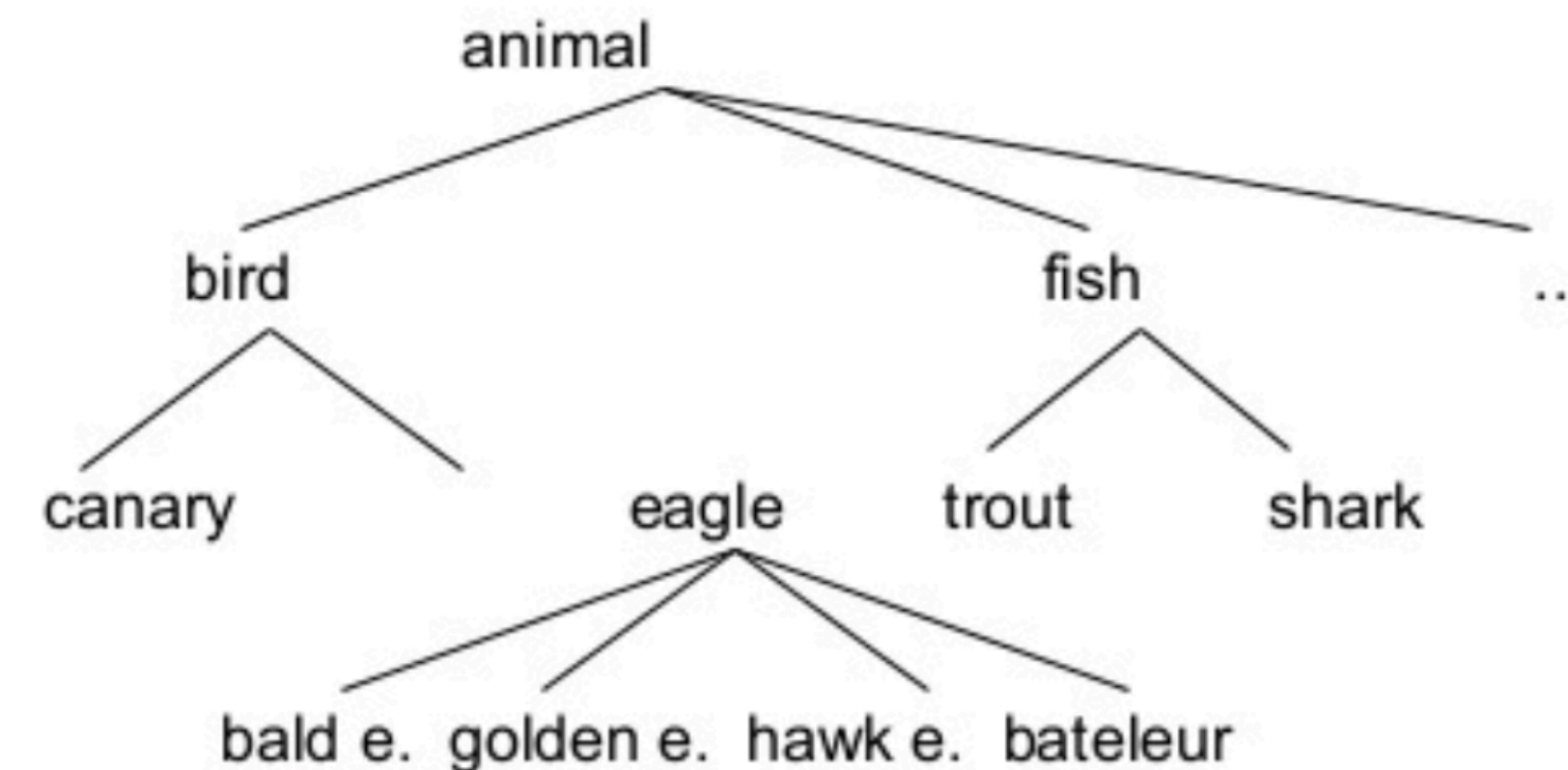
- Words are assigned to lexical categories: nouns, verbs, adjectives, adverbs
- Also other classes of concepts came later: determiners and other function words (pronouns, auxiliary words, conjunctions (coming soon))
- Prepositions (still out)- but point of discussion
- Lexical relations between word forms

Noun

- **S: (n) happiness, felicity** (state of well-being characterized by emotions ranging from contentment to intense joy)
 - [direct hyponym](#) / [full hyponym](#)
 - [attribute](#)
 - [direct hypernym](#) / [inherited hypernym](#) / [sister term](#)
 - [antonym](#)
 - [derivationally related form](#)
 - **W: (adj) happy** [Related to: [happiness](#)] (marked by good fortune) "a *felicitous life*"; "a *happy outcome*"
 - **W: (adj) happy** [Related to: [happiness](#)] (enjoying or showing or marked by joy or pleasure) "a *happy smile*"; "spent many *happy days on the beach*"; "a *happy marriage*"
 - **W: (adj) felicitous** [Related to: [felicity](#)] (marked by good fortune) "a *felicitous life*"; "a *happy outcome*"
- **S: (n) happiness** (emotions experienced when in a state of well-being)
 - [direct hyponym](#) / [full hyponym](#)
 - [attribute](#)
 - [direct hypernym](#) / [inherited hypernym](#) / [sister term](#)
 - [antonym](#)
 - [derivationally related form](#)
 - **W: (adj) happy** [Related to: [happiness](#)] (enjoying or showing or marked by joy or pleasure) "a *happy smile*"; "spent many *happy days on the beach*"; "a *happy marriage*"

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:

Display Options:

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 hyponym](#) / [full hyponym](#)
 - <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 rhythmical 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 rhythmical way) *"The young girl danced into the room"*
- <verb.creation> **S: (v) dance, trip the light fantastic, trip the light fantastic toe** (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 rhythmical way) *"The young girl danced into the room"*
 - [domain category](#)
 - <noun.act> **S: (n) dancing, dance, terpsichore, saltation** (taking a series of rhythmical 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"*

synset

gloss

lemmas

lexicographer's files

categories e.g.

noun.animal nouns denoting animals

noun.artifact nouns denoting man-made objects

Word to search for:

Display Options:

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- **S: (n) plant, works, industrial plant** (buildings for carrying on industrial labor) *"they built a large plant to manufacture automobiles"*
- **S: (n) plant, flora, plant life** ((botany) a living organism lacking the power of locomotion)
- **S: (n) plant** (an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience)
- **S: (n) plant** (something planted secretly for discovery by another) *"the police used a plant to trick the thieves"; "he claimed that the evidence against him was a plant"*

Verb

- **S: (v) plant, set** (put or set (seeds, seedlings, or plants) into the ground) *"Let's plant flowers in the garden"*
 - [direct troponym](#) / [full troponym](#)
 - [direct hypernym](#) / [inherited hypernym](#) / [sister term](#)
 - [derivationally related form](#)
- **W: (n) plant** [Related to: [plant](#)] ((botany) a living organism lacking the power of locomotion)
 - **W: (n) planter** [Related to: [plant](#)] (a decorative pot for house plants)
 - **W: (n) planting** [Related to: [plant](#)] (putting seeds or young plants in the ground to grow) *"the planting of corn is hard work"*
 - **W: (n) set** [Related to: [set](#)] (the act of putting something in position) *"he gave a final set to his hat"*
- [sentence frame](#)
- **S: (v) implant, engraft, embed, imbed, plant** (fix or set securely or deeply) *"He planted a knee in the back of his opponent"; "The dentist implanted a tooth in the gum"*
- **S: (v) establish, found, plant, constitute, institute** (set up or lay the groundwork for) *"establish a new department"*
- **S: (v) plant** (place into a river) *"plant fish"*
 - [domain category](#)
 - **S: (n) animal husbandry** (breeding and caring for farm animals)
 - [direct hypernym](#) / [inherited hypernym](#) / [sister term](#)
 - [sentence frame](#)
- **S: (v) plant** (place something or someone in a certain position in order to secretly observe or deceive) *"Plant a spy in Moscow"; "plant bugs in the dissident's apartment"*
- **S: (v) plant, implant** (put firmly in the mind) *"Plant a thought in the students' minds"*

Also:
lex. relations
Antonym,
Pertainym,
Participle

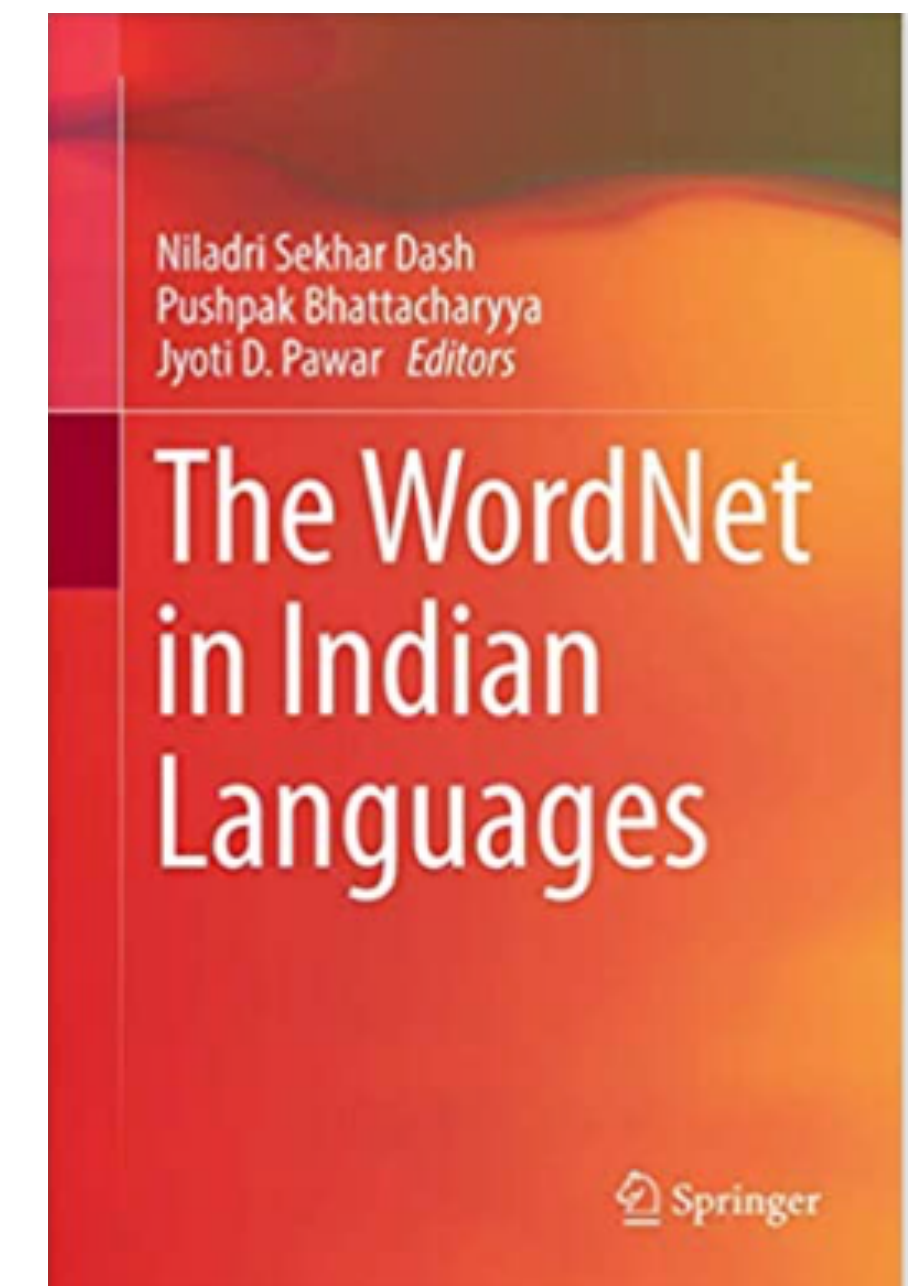
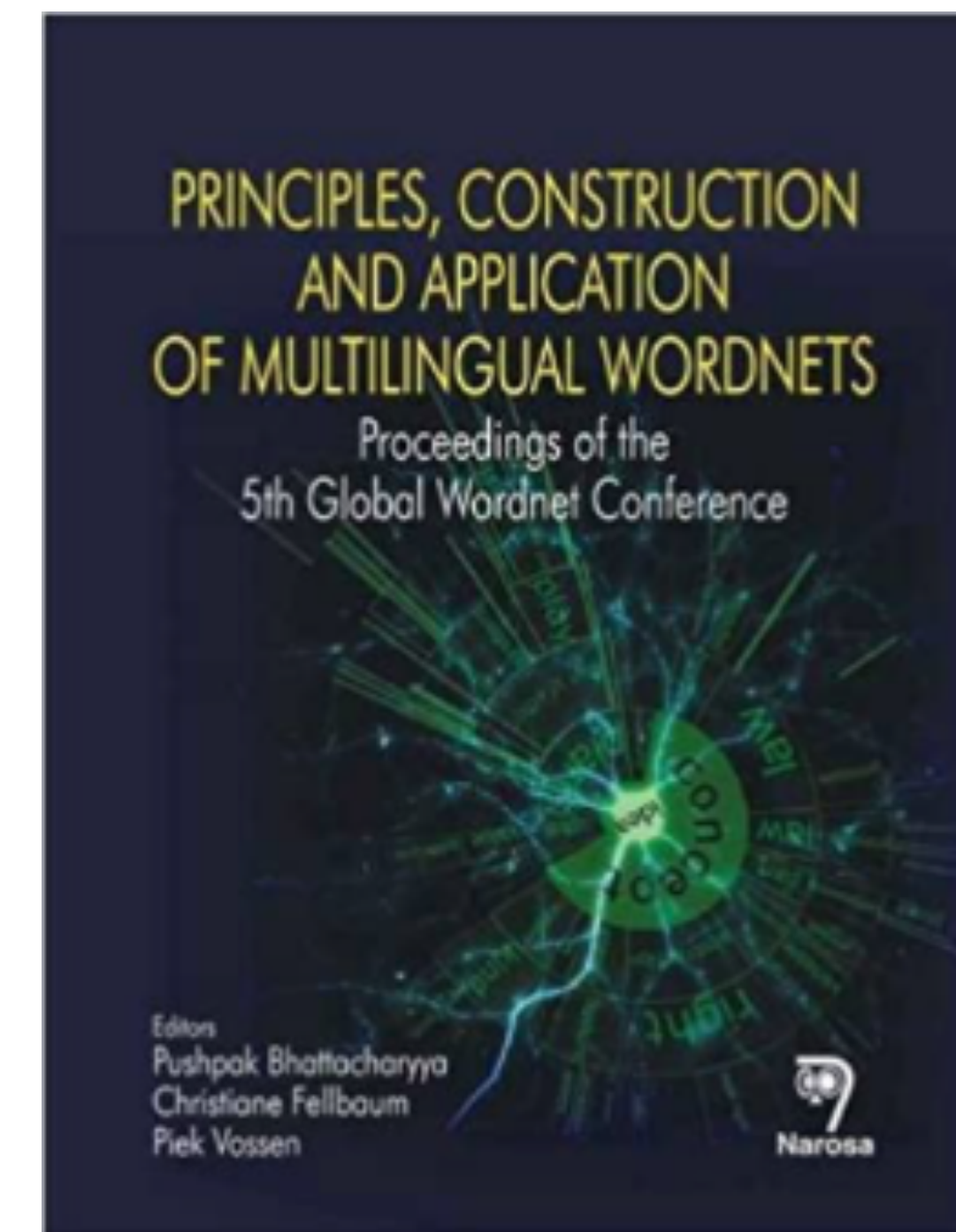
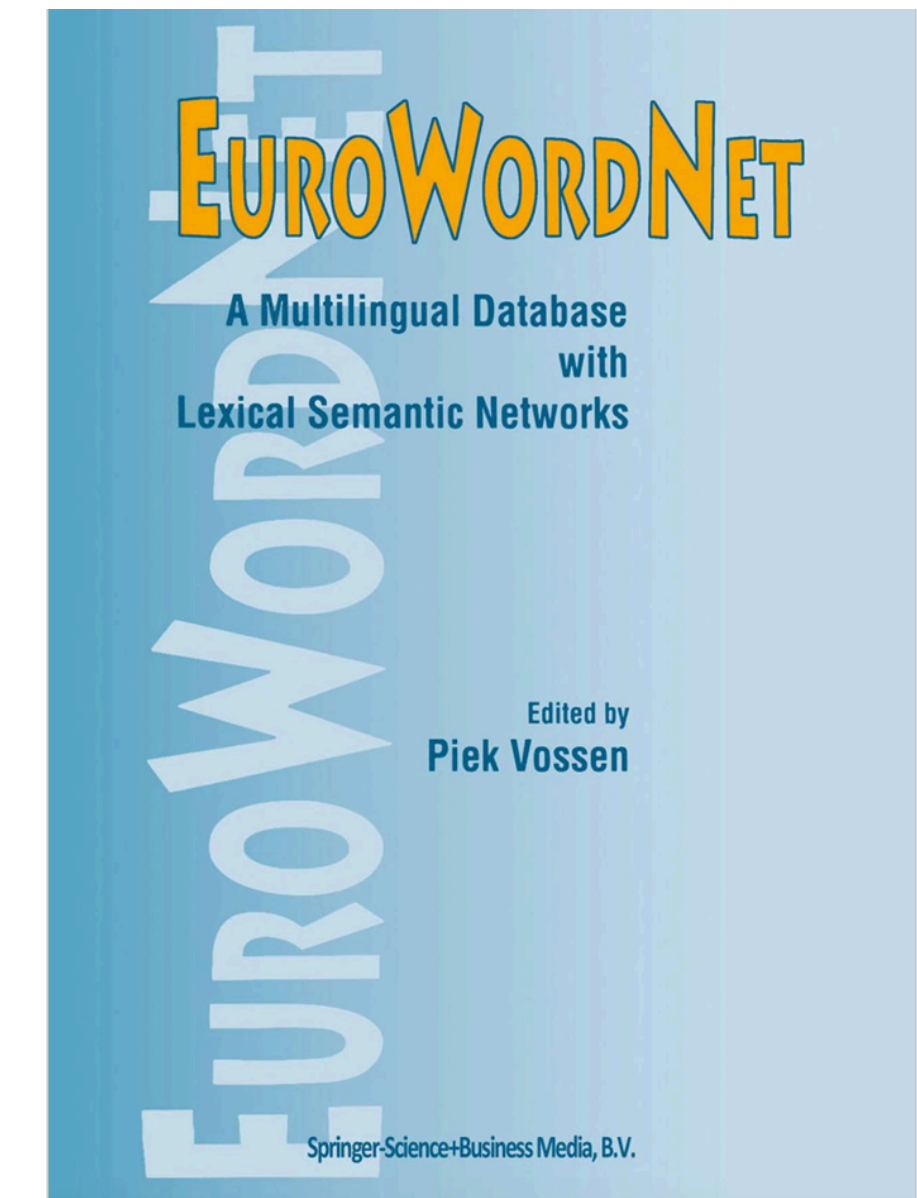
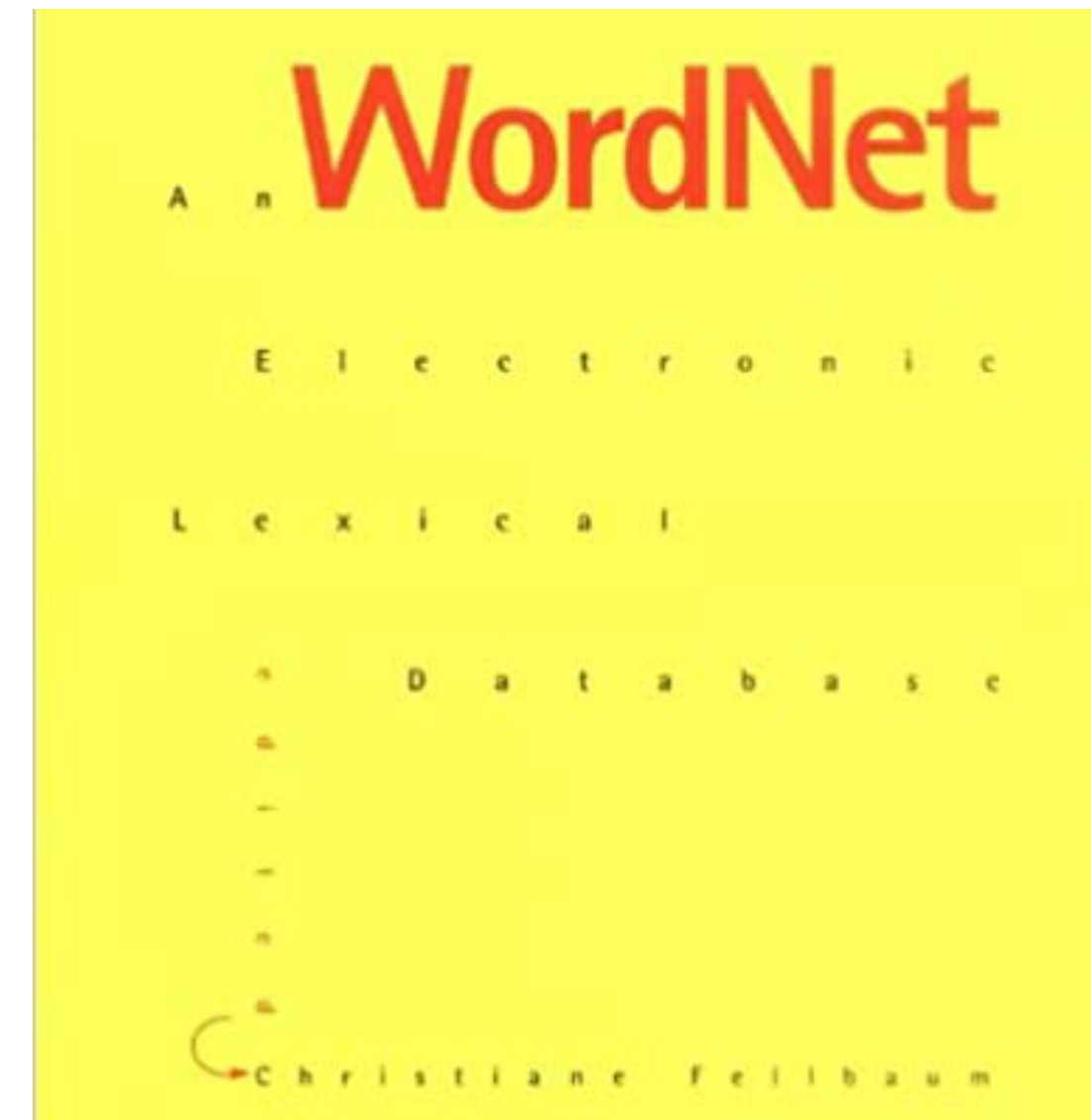
simple sentence
of verb use

e.g. Somebody
----s something

domain categories

Readings & Resources

- Fellbaum, Christiane. "WordNet." Theory and applications of ontology: computer applications. Springer, Dordrecht, 2010. 231-243.
- Global Wordnet Association (GWA) proceedings: <http://globalwordnet.org/global-wordnet-conferences-2/>
- 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

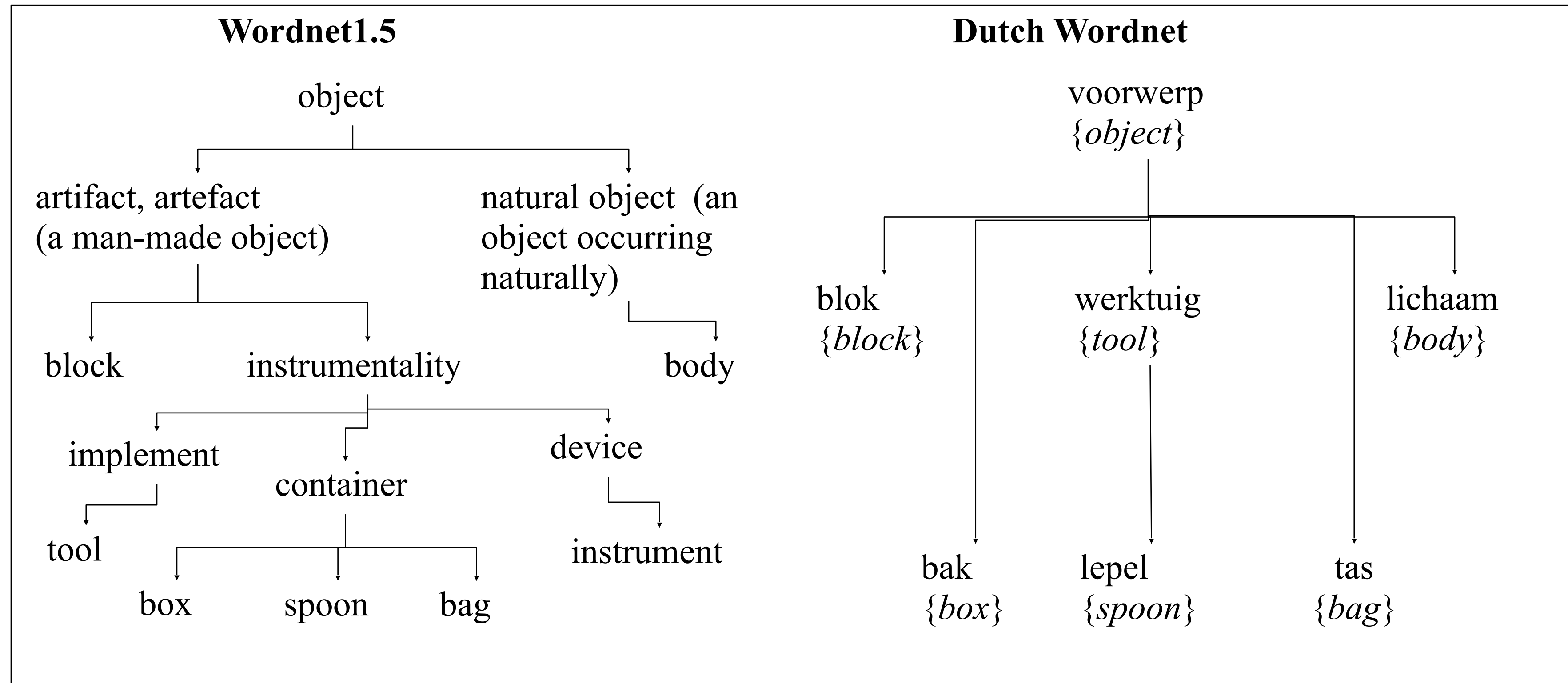
- To promote the standardization of the specification of wordnets for all languages in the world, including:
 - the standardization of the **Common Inter-Lingual-Index (CILI)** for **inter-linking the wordnets of different languages, as a universal index of meaning**
 - the development of a common representation 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

Bond, F., Vossen, P. T. J. M., McCrae, J., & Fellbaum, C. (2016). CILI: the Collaborative Interlingual Index. In V. Barbu Mititelu, C. Forăscu, P. T. J. M. Vossen, & C. Fellbaum (Eds.), *Proceeding of the 8th Global WordNet Conference 2016*.



- Artificial Classes versus Lexicalized Classes:

instrumentality; natural object

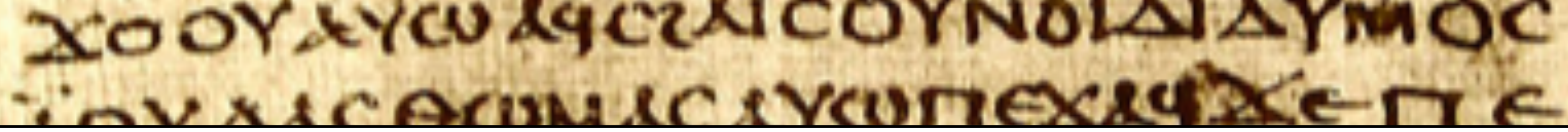
- Lexicalization differences of classes:

container and artifact (object) are not lexicalized in Dutch



Religious/Theology Related Concepts— CILI

- Slaughter, Laura, Wenjie Wang, Luis Morgado da Costa, and Francis Bond. "Enhancing the Collaborative Interlingual Index for Digital Humanities: Cross-linguistic analysis in the domain of theology." (2018)
 - <http://compling.hss.ntu.edu.sg/events/2018-gwc/pdfs/gwc-2018-proceedings.pdf>
- 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.

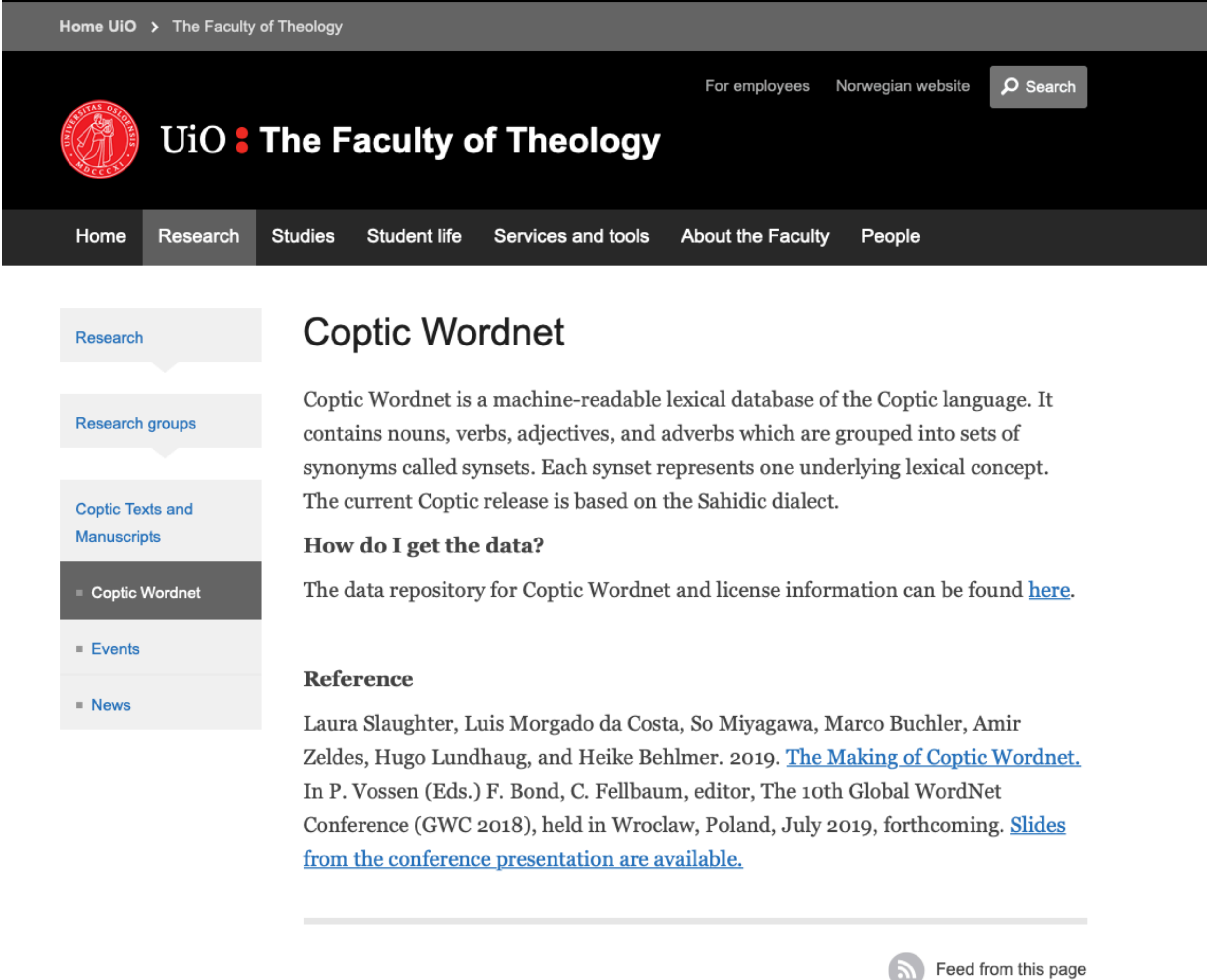
```
<LexicalEntry id="w613347">
  <Lemma writtenForm="动物沟通" partOfSpeech="n" script="Hans"/>
  <Form writtenForm="dòngwùgōutōng" script="Latn-pinyin">
    <Tag category="transliteration">pīnyīn</Tag>
    <Tag category="confidence">0.77</Tag>
  </Form>
  <Form writtenForm="dong4wu4gou1tong1" script="Latn-pinyin">
    <Tag category="transliteration">pin1yin1</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/>

Laura Slaughter, Luis Morgado da Costa, So Miyagawa, Marco Buchler, Amir Zeldes, Hugo Lundhaug, and Heike Behlmer. 2019. [The Making of Coptic Wordnet](#). In P. Vossen (Eds.) F. Bond, C. Fellbaum, editor, The 10th Global WordNet Conference (GWC 2018), held in Wroclaw, Poland, July 2019



The screenshot shows the website for UiO The Faculty of Theology. The header includes the university logo and navigation links for Home, Research, Studies, Student life, Services and tools, About the Faculty, and People. A search bar is also present. The main content area is titled "Coptic Wordnet" and contains the following text:

Coptic Wordnet is a machine-readable lexical database of the Coptic language. It contains nouns, verbs, adjectives, and adverbs which are grouped into sets of synonyms called synsets. Each synset represents one underlying lexical concept. The current Coptic release is based on the Sahidic dialect.

How do I get the data?

The data repository for Coptic Wordnet and license information can be found [here](#).

Reference

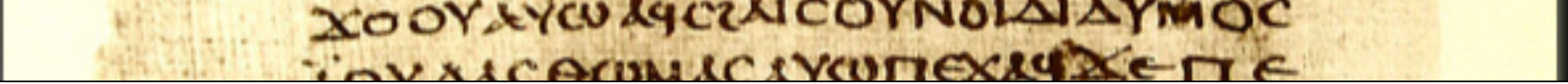
Laura Slaughter, Luis Morgado da Costa, So Miyagawa, Marco Buchler, Amir Zeldes, Hugo Lundhaug, and Heike Behlmer. 2019. [The Making of Coptic Wordnet](#). In P. Vossen (Eds.) F. Bond, C. Fellbaum, editor, The 10th Global WordNet Conference (GWC 2018), held in Wroclaw, Poland, July 2019, forthcoming. [Slides from the conference presentation are available](#).

At the bottom right of the page, there is a "Feed from this page" icon.



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)

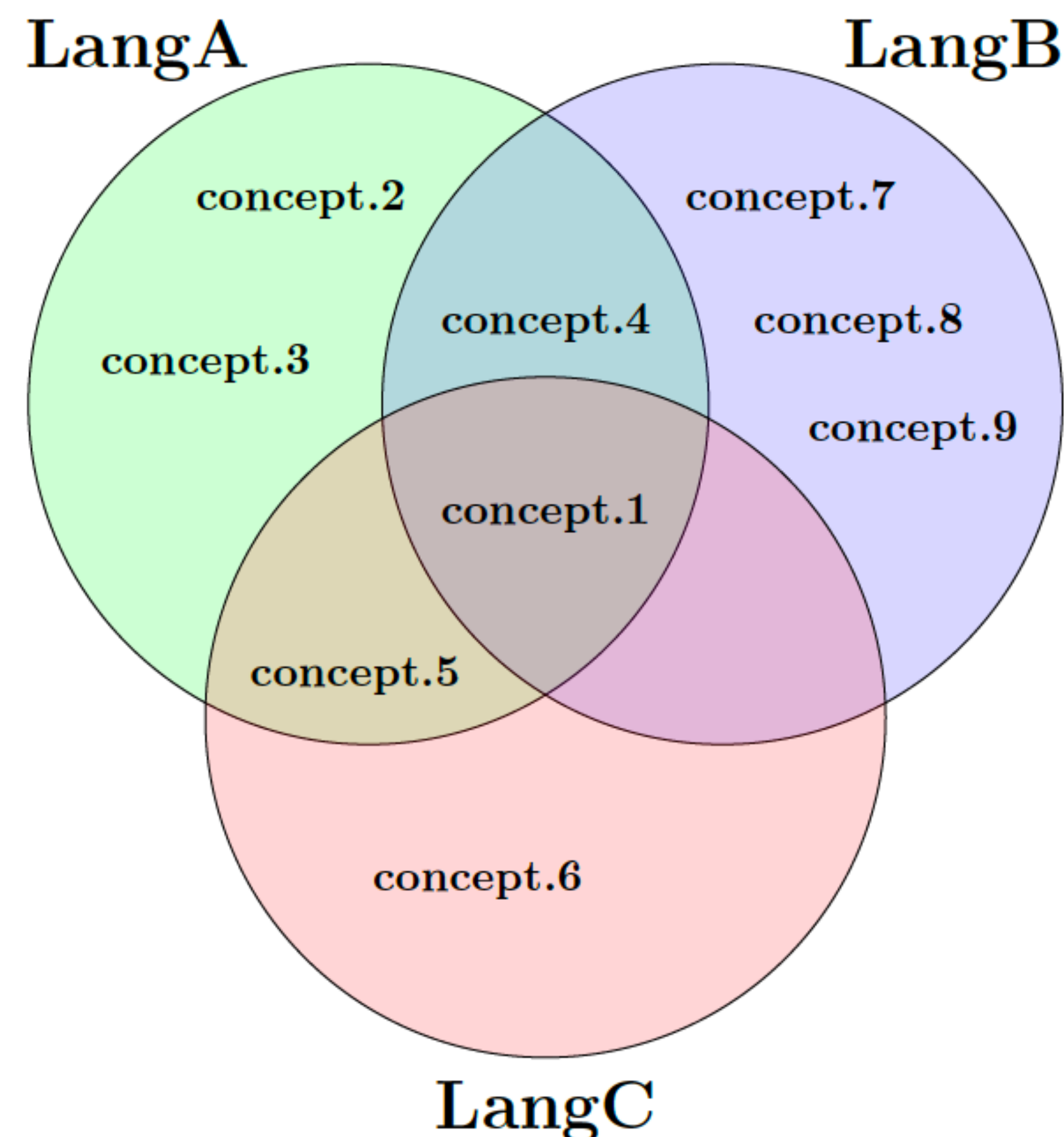


Automated Construction Approach

- 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)



**1 coptic lemma had 9 possible concepts
concept 1 has highest ranking**



Evaluation: Manual Checking

0/1/?	No. Langs	Candidate Lemma	Matched Translations	Candidate Synset	English Lemmas, Definitions and Examples
1	2	βωπε	'fra saisir n', 'fra saisir v', 'eng seize v', 'eng seize n'	02273293-v	confiscate; attach; impound; seize; sequester [take temporary possession of as a security, by legal authority] The police confiscated the stolen artwork

- 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”



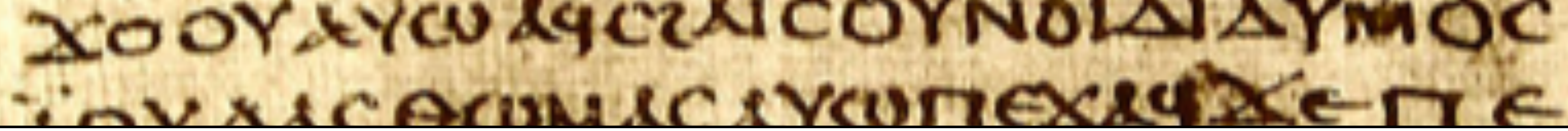
Coptic Wordnet Coverage

POS	No. synsets	No. senses
nouns	13,904	97,527
verbs	7,491	92,019
adjective	3,488	20,723
satellite adj	229	587
adverb	737	7,373
non-referential	22	448
Total	25,871	218,677



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”

No. Langs	Correct (%) Union	Correct (%) Intersect.
1	(n=119) 25%	7%
2	(n=134) 89%	49%
3	(n=40) 98%	63%
4	(n=7) 100%	100%
Total	62% (n=300)	34% (n=300)



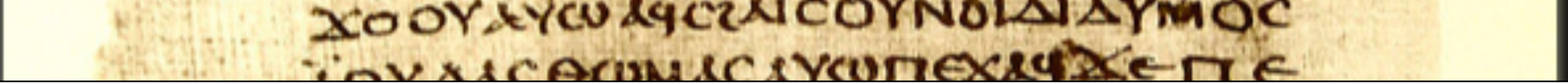
Applications of Wordnets

- 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 CLI 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 teuse 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