Sunoikisis Digital Classics, Summer 2022, April 28th 2022

WordNets for Ancient Indo-European Languages

Erica Biagetti, University of Pavia
Chiara Zanchi, University of Pavia
William M. Short, University of Exeter

Outline of today's session

The seminar will consist of three main sections:

1. Introduction to WordNets and to their use in linguistic research
2. Description of the innovations we introduced or plan to introduce to the original architecture and of the utility of such innovations for studies in linguistic typology, historical linguistics, philology, and related fields
3. Demonstration on how data for Ancient Greek and Sanskrit were collected and integrated into a new annotation interface

+ a bonus: annotation exercise ☺

Part 1: Introduction to Wordnets

Chiara Zanchi, University of Pavia

What is WordNet?

- A large **lexical database or "electronic dictionary"** , originally developed at Princeton University for Modern English within a project led by George A. Miller and Christiane Fellbaum (http://wordnet.princeton.edu)
- Can be read by **humans and machines**
- Includes most English nouns, verbs, adjectives, adverbs ("**lexical PoS**")
- Princeton WordNet is for English only, but can be linked to WordNets in many other languages
  - Cfr. EuroWordNet, MultiWordNet projects
  - Global WordNet Association: http://globalwordnet.org/resources/wordnets-in-the-world/

Why is WordNet different from paper dictionaries?

- Traditional dictionaries are organized **alphabetically**: words on the same page are not related by *meaning*
- WordNet is organized **by meaning**: words in close proximity in the network are semantically similar
- WordNet contains information about two fundamental properties of human language: **synonymy** and **polysemy**
  - Synonymy = one : many mapping of meaning and form
  - Polysemy = one : many mapping of form and meaning

Synonymy and polysemy

Single concept, multiple words

- {01238728 hit against; come into sudden contact wit} (v) hit, strike, impinge on, run into, collide with
- {06443410 the sacred writings of the Christian religions} (n) Bible, Christian Bible, Book, Holy Scripture, Holy Write, Scripture, Word of God, Word

Single word, multiple concepts

- Bank
  - (n)sloping land
  - (n) financial institution
  - (n) a long ridge or pile
  - (v) tip laterally
  - (v) do business with a bank
  - Etc…
- Higly frequent words are highly polysemous

Synonymy and polysemy in WordNet

Synonymy

- WordNet groups (roughly) synonymous, denotationally equivalent words, into unordered sets of synonyms ("synsets")
- Each synset expresses a distinct concept
- Synsets are associated with a brief gloss, examples, and a unique ID number

Polysemy

- A word form that appears in *n* synsets is *n*-fold polysemous
- E.g., bank appears in 18 synsets (http://wordnetweb.princeton.edu/perl/webwn?s=bank&sub=Search+WordNet&o2=&o0=1&o8=1&o1=1&o7=&o5=&o9=&o6=&o3=&o4=&h=)
- Thus, bank has 18 distinct senses, it is associated with 18 concepts

Basic WordNet stats

WordNet 2.1

POS	Unique Strings	Synsets	Total Word-Sense Pairs
Noun	117097	81426	145104
Verb	11488	13650	24890
Adjective	22141	18877	31302
Adverb	4601	3644	5720
Totals	155327	117597	207016

➔ **OpenWordNet**:

Open English WordNet is a fork of the Princeton Wordnet developed under an open source methodology

(https://github.com/globalwordnet/english-wordnet)

What about the «net» part of WordNet?

- Words are connected through **lexical relations**
- Synsets (i.e. concepts) related through **semantic-conceptual relations**
- Bi-directional arcs express relations, resulting in a large semantic network (**graph**)
- Four semantic networks for the four lexical PoS included in the lexicon: nouns, verbs, adjectives, adverbs
- There are also relations connecting different PoS

Semantic-conceptual relations: examples

Hypo-/hypernymy relates noun synsets

{robin, redbreast} → {bird} → {animal, animate_being}

«Robin is a kind of bird», «bird is a kind of animal», «robin is a kind of animal», «The class of animals inclused birds», «the class of birds includes robins», «The class of animals includes robins»

- Relates more/less general concepts
- Creates hierarchies or “trees”; hierarchies can have up to 16 levels
- Properties: asymmetry and transitivity

Semantic-conceptual relations: examples

Meronymy/holonymy (part-whole relation) relates noun synsets

{finger} → {hand} → {harm} → {body}

«A hand has fingers», «a harm has hands», «a body has harms», ?? «a body has fingers»

- component/object {branch} → {tree}
- member/collection {tree} → {forest}
- stuff/object {aluminium} → {airplane}

Psycholinguistics bases and a bit of history

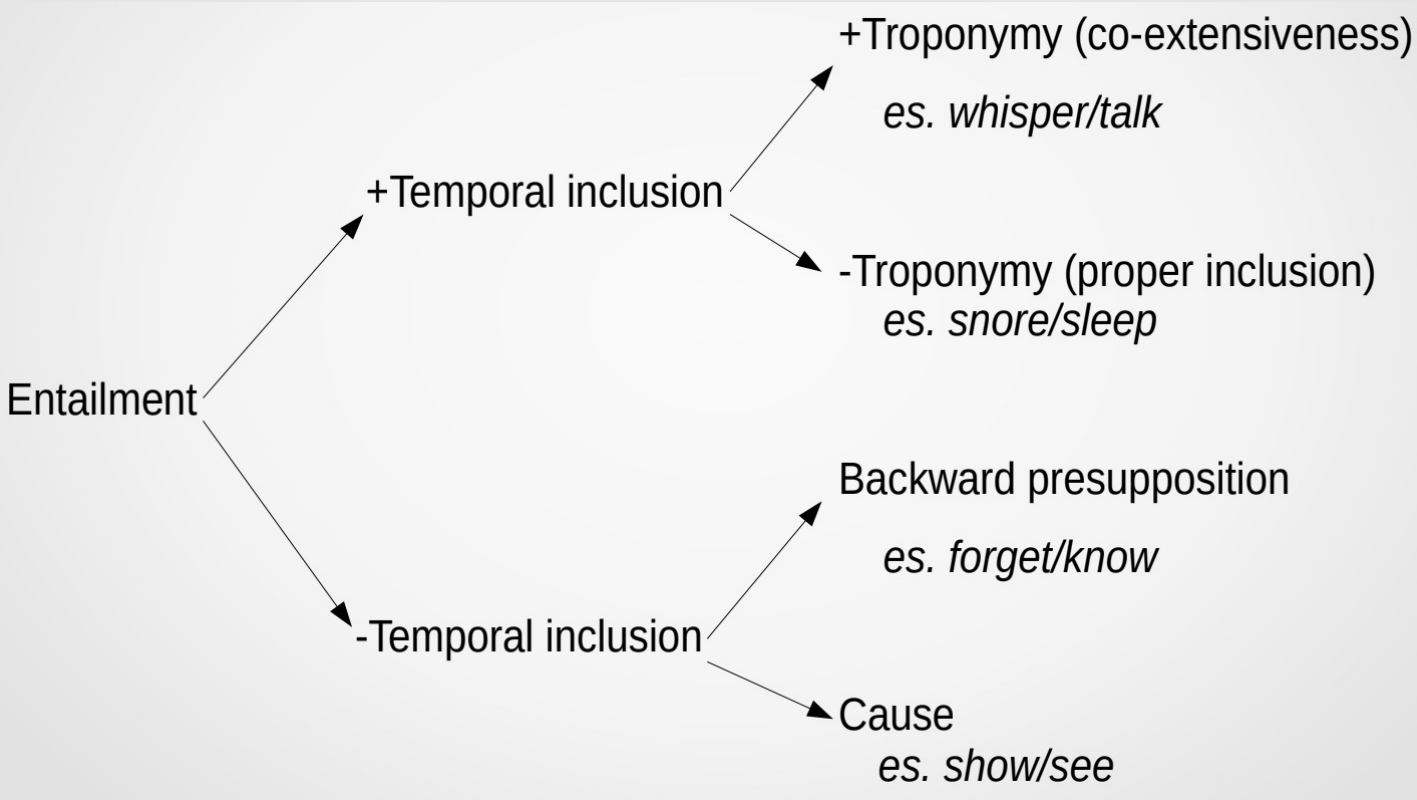
1980s Research in Artificial Intelligence:

- **How do humans store and access knowledge about concepts?**
- Knowledge about concepts is huge, therefore must be stored in an efficient and economic fashion
- Hypothesis: **concepts are interconnected via meaningful relations**
- Related to this: knowledge about concepts is computed “on the fly” via access to general concepts: we know that “redbreasts fly” because “birds fly” and “redbreasts are a kind of bird”
- Knowledge is stored at the highest node *(animals move, birds fly, birds sing)*

Psycholinguistic bases and a bit of history

- Collins & Quillian (1969) measured reaction times to statements involving knowledge distributed across different “levels”
- People confirmed statements like (1) bird lays eggs faster than (2) redbreast lays eggs
- Hypothesis: (2) requires “look-up” at higher level, (1) does not
- **Collins’ & Quillian’s results are not compelling**
  - Reaction time to statements like “do canaries move?” are influenced e.g. by **prototypicality** (redbreasts are more typical birds than penguins) and **word frequency** (people recognize frequent words faster)
- **But the idea inspired WordNet**: Can most/all of the lexicon be represented as a semantic network? Or are there unconnectable words and concepts?

Semantic-conceptual relations: relations among verbs



General principle: **lexical entailment**

{to snore} V1 → {to sleep} V2

- **Unidirectional relation**: if V1 implies V2, V2 does not imply V1; if there is bidirectional implication, the verbs are synonyms

Relations among adjectives

(adj) **short**, little (low in stature; not tall) *"he was short and stocky"; "short in stature"; "a short smokestack"; "a little man"*

- **Attribute (S)**

(n) stature, height ((of a standing person) the distance from head to foot)

- **Antonym (W)**

(adj) tall [Opposed to: short] (great in vertical dimension; high in stature) *"tall people"; "tall buildings"; "tall trees"; "tall ships"*

- **Derivationally related form (W)**

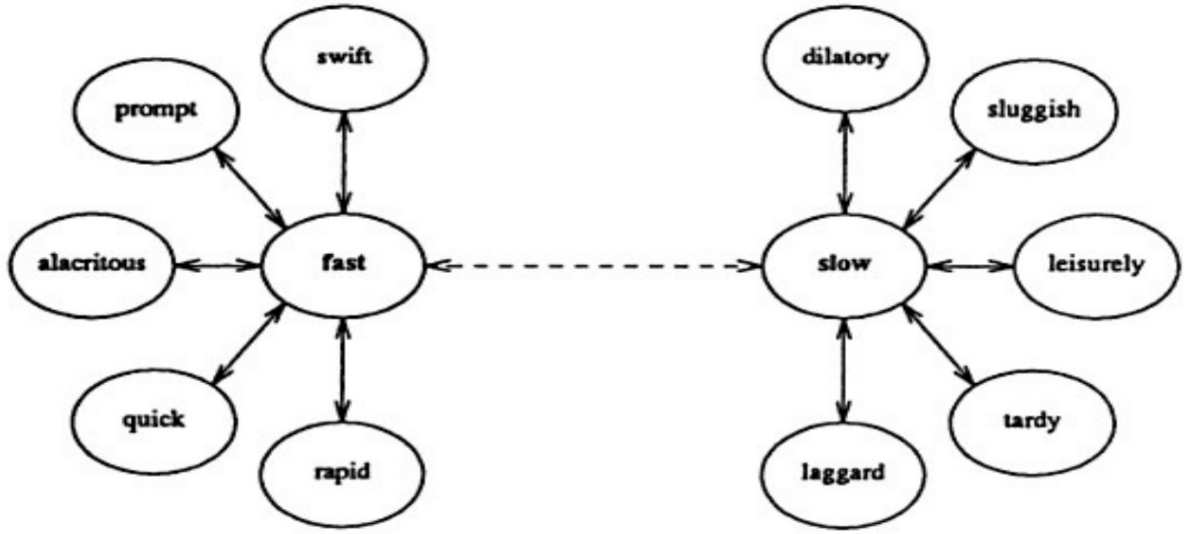
(n) shortness [Related to: short] (the property of being truncated or short)

(n) shortness [Related to: short] (the property of being shorter than average stature)

N.B.: S = relation among synsets, W = relation among words

Relations among adjectives

- Adjectival similar synsets are grouped in clusters
- These clusters are organized around a prototype
- Prototypes are direct antonyms; other members of the clusters are indirect antonyms
- Experimental evidence: reaction time measurements for semantic judgments (it takes less time to confirm that *fast* and *slow* are opposites than that *fast* and *sluggish* are)



SOS how do we select these prototypes for ancient languages (no native speakers)?

Other relations among words

(adj) **stellar**, astral (being or relating to or resembling or emanating from stars) *"an astral body"; "stellar light"*

- **Pertainym (W)**

(n) star [Related to: stellar, astral] ((astronomy) a celestial body of hot gases that radiates energy derived from thermonuclear reactions in the interior)

- It is the same lexical relation holding between {quick} → {quickly}

Other relations among words

(v) **bring**, convey, take (take something or somebody with oneself somewhere) *"Bring me the box from the other room"; "Take these letters to the boss"; "This brings me to the main point"*

- **Phrasal verb (W)**
  - (v) bring forward [Related to: bring] (cause to move forward) *"Can you move the car seat forward?"*
  - (v) bring down [Related to: bring] (cause to come to the ground) *"the pilot managed to land the airplane safely"*
  - (v) bring down [Related to: bring] (move something or somebody to a lower position) *"take down the vase from the shelf"*
  - (v) bring up [Related to: bring] (raise from a lower to a higher position) *"Raise your hands"; "Lift a load"*
  - (v) bring on [Related to: bring] (bring onto the market or release) *"produce a movie"; "bring out a book"; "produce a new play"*
  - (v) bring out [Related to: bring] (bring onto the market or release) *"produce a movie"; "bring out a book"; "produce a new play"*

Sentence frames

- Syntactic information in the Princeton WordNet is very limited
- **Sentence frames**

(v) **bring** (cause to come into a particular state or condition) *"Long hard years of on the job training had brought them to their competence"; "bring water to the boiling point"*

- **Sentence frame**
  - Somebody ----s something
  - Somebody ----s somebody
  - Something ----s somebody
  - Something ----s something

- Animacy information on events participants
- Aspect
- Other WordNets may contain different information: e.g., German WordNet gives information about case marking (GermaNet (https://weblicht.sfs.uni-tuebingen.de/rover/)

What is not in WordNet?

- IPA, prosody
- Morphological information
- Etymologies
- Collocations
- Information about semantic change → but the lexicon is dynamic!
- Information about genres
- **The "tennis" problem**: no relation indicates that two concepts belong with the same topic in discourse: racquet – ball – net

Applications

- WordNet has been incorporated into many **online dictionaries**
- **Semantic annotation** → semantic + syntactic annotation: great potential
- **Word Sense Identification – Word Sense Disambiguation**: crucial tasks for information retrival, text mining, document sorting, machine translation

WordNet based approach to WSD

Example: ***John needed cash so he walked over to the* bank**

Which *bank*?
- Money institution? (building/institution?)
- Sloping land by the water?

- Look for words in the vicinity (context) of the target word
- Find that sense of the target word that is related to the context words in WordNet (shared superordinate, parts, definition, etc.)
- Shortest path among candidate words often shows the intended sense

Bank/money institution and *cash* are linked (share words in their definitions), so if *cash* and *bank* co-occur in a context, *bank* likely has the "financial institution" sense

Crosslinguistic WordNets

- Starting in late 1990s, WordNets were built for languages other than English
- Genetically and typologically unrelated languages (Turkish, Hindi, Chinese, Korean, Basque, Zulu, Arabic,…currently >70) mapped to Princeton WordNet
- Great potential for crosslinguistic applications, e.g. translation, but also language typology
- First set of foreign-language WNs (“EuroWordNet” by Piek Vossen) were built with reference to Princeton WordNet; each synset in each WN was linked to a “record” (PWN synset identifier) in the index; crosslingual mapping of words and synsets proceeds via the index

Mismatches in multilingual WordNets

- Concepts not lexicalized in English require new synsets: e.g., Latin has words for maternal and paternal uncles and aunts, English has not
- Some languages lack equivalents of English words: e.g., Dutch lacks a word for *container* but has kinds (hyponyms) of *container* (*box, bag, bucket..*)
- Proper names lacking in the original WordNet
- Anachronistic synsets, e.g., Athens = “the capital and largest city of Greece; named after Athena (its patron goddess)”; referentially correct, but Ancient Greek *Athênai* could also metonymically refer to the entire region; the gloss refers to the city in a way that is incompatible with its ancient conceptualization (“capital”)

WordNets of ancient languages: previous attempts

(Minozzi 2009, Bizzoni et al 2014, Boschetti 2019)

- The first **Latin WordNet (LWN)** was created using Vossen's (1998) "Expand Method": parts of the Italian and English sections of the MultiWordNet were automatically translated into Latin with the aid of bilingual dictionaries.
- The resulting dataset of 9,378 lemmas and 8,973 synsets was both largely incomplete (e.g. common verbs such as "amo" 'to love' were lacking many synsets) and inaccurate, especially due to anachronistic senses extracted from the MultiWordNet (e.g. by association with the meanings of "chiamare" and "call", Latin "voco" is assigned the synset "send a message or attempt to reach someone by radio, phone, etc [...]").

WordNets of Ancient languages: previous attempts

- The **Sanskrit WordNet (SWN)** is based on Oliver Hellwig's work for the Digital Corpus of Sanskrit (DCS: http://www.sanskrit-linguistics.org/dcs/index.php; Hellwig 2017)
- The SWN was built by annotating an ontology, i.e. the OpenCyc knowledge base (Lenat 1995), which contains concepts with English descriptions and relational knowledge about them. Appropriate concepts were used to manually annotate selected texts in the DCS for lexical semantics
- Ca. 600,609 tokens and 32,227 lemmas have been semantically annotated, and the current semantic network consists of 124,040 concepts and 194,092 relations
- Annotating with this method, synsets were meanwhile populated by the Sanskrit words assigned to the same ontological concept. Subsequently, a large subset of OpenCyp was aligned with the PWN

WordNets of Ancient languages: previous attempts

- Drawing from a previous collaboration with the University of Pavia, the first version of the Ancient Greek WordNet (AGWN) was born in 2014 as an international effort of the Institute of Computational Linguistics “Antonio Zampolli” (Pisa), the Perseus Project, the Open Philology Project, and the Alpheios Project
- The AGWN was initially constructed using Greek-English digitized lexica at the Perseus Project using the "Expand Method", as in the case of the first Latin WordNet.
- In addition, synsets in the AGWN were aligned to the Italian section of the MultiWordNet, to another Italian WordNet (ItalWordNet: http://www.ilc.cnr.it/it/content/italwordnet) and to the Latin WordNet. A subset of synsets was used to evaluate the automatic extraction, and erroneous alignments were eliminated by filtering anachronistic domains