WordNet

What is WordNet?

- A large lexical database, semantic resource, “electronic dictionary,” developed and maintained at Princeton University

  http://wordnet.princeton.edu
- Includes most English nouns, verbs, adjectives, adverbs
- Electronic format makes it accessible and useful for automatic systems
- Used in many Natural Language Processing applications requiring semantic analysis (information retrieval, text mining, question answering, machine translation, AI/reasoning,...)

What's special about WordNet?

- Traditional paper dictionaries are organized alphabetically
- As a result, words that are found together (on the same page) are not related by *meaning*
- WordNet is organized by meaning: words in close proximity are semantically similar
- Human users and computers can browse WordNet and find words that are meaningfully related to their queries (somewhat like in a hyperdimensional thesaurus)
- Meaning similiarity can be measured and quantified to support Natural Language Understanding, in particular Word Sense Disambiguation

Language is a bit random

WordNet allows one to investigate to what extent the language systematically encodes/lexicalizes (labels with a word) a concept

Global and local systematicity

Where are “holes” (lexical gaps?) And are these indicative of concepts that happen not to be lexicalized?

Lexical gaps

- Simple example: kinship terms
- English encodes both vertical and horizontal relations
- But arguably not (as) systematically (as other languages)

English kinship terms

English does not lexically distinguish

-younger and older siblings (cf. Japanese)

-male and female cousins (cf. French, German, Arabic)

-maternal and paternal aunts and uncles (Arabic)

WordNet: A bit of history

Late 1960s, 70s: Artificial Intelligence (AI), cognitive science attempt to understand and model the human mind

Language is one of the most complex ways in which the human mind manifests itself

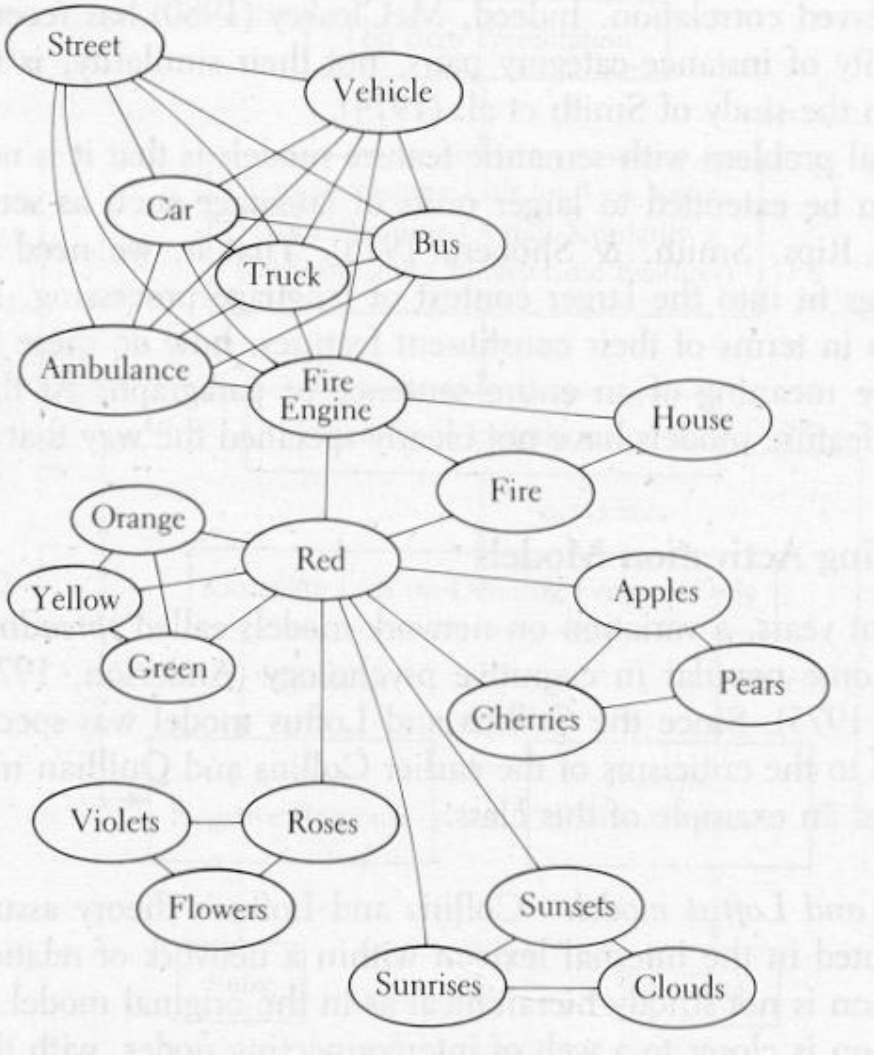
Language and Mind

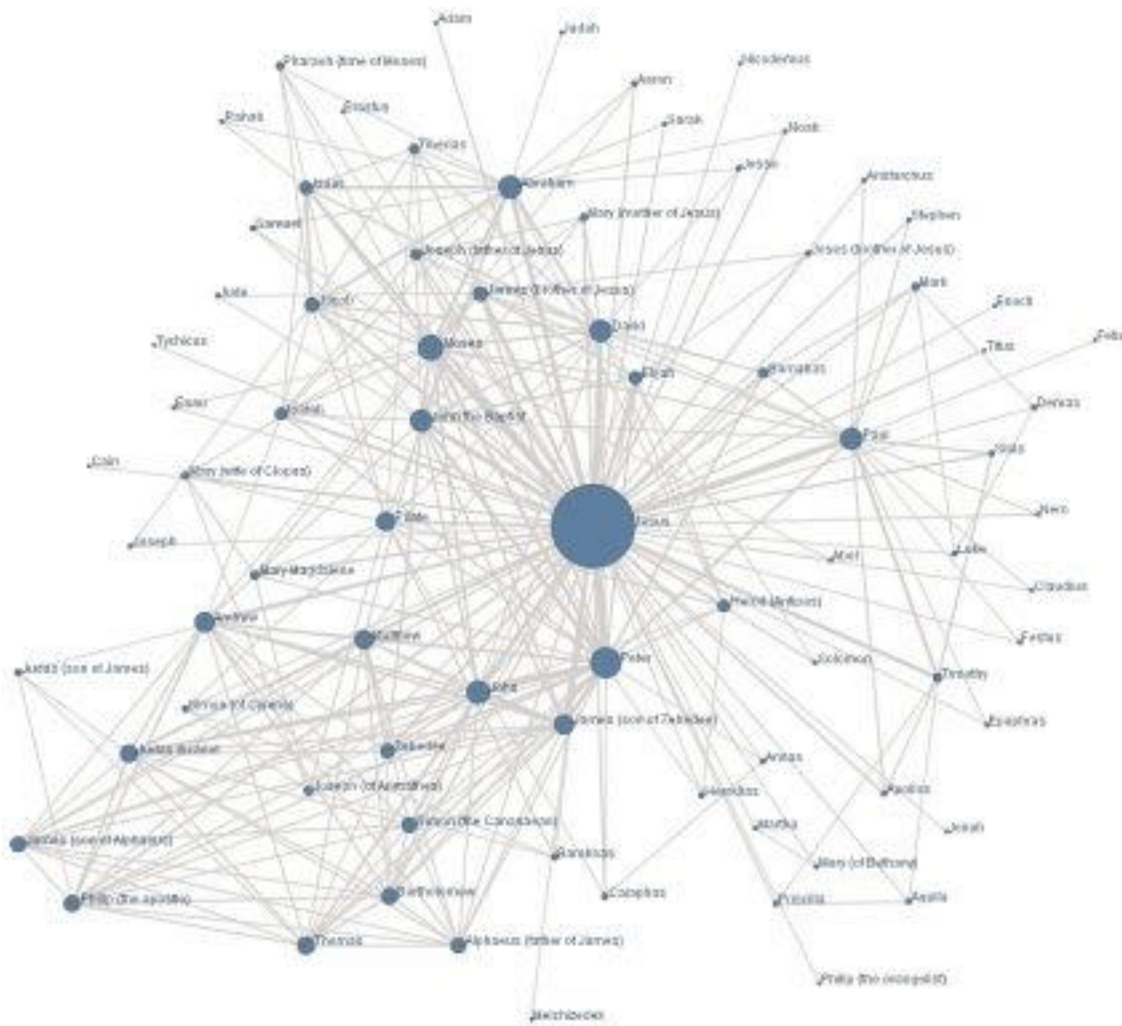
How do humans store and access knowledge about concept?

Hypothesis: concepts are interconnected via meaningful relations

Semantic network representation

(Collins and Quillian 1969, 1970, 1972)


Street
Vehicle
Car
Bus
Truck
Ambulance
Fire Engine
House
Fire
Orange
Red
Yellow
Apples
Green
Pears
Cherries
Violets
Roses
Flowers
Sunsets
Sunrises
Clouds


Adam
Judah
Pharaoh (time of Moses)
Nicodemus
Aaron
Erastus
Rahab
Sarah
Noah
Tiberius
Jesse
Isaiah
Abraham
Aristarchus
Mary (mother of Jesus)
Stephen
Samuel
Joseph (father of Jesus)
Joses (brother of Jesus)
Jude
Jacob
James (brother of Jesus)
Mark
Enoch
David
Tychicus
Moses
Titus
Felix
Elijah
Barnabas
Esther
Joshua
John the Baptist
Demas
Paul
Silas
Cain
Mary (wife of Clopas)
Nero
Pilate
Jesus
Joseph
Joel
Luke
Mary Magdalene
Claudius
Herod (Antipas)
Andrew
Matthew
Festus
Peter
Solomon
Judas (son of James)
Timothy
John
Simon (of Cyrene)
James (son of Zebedee)
Epaphras
Judas Iscariot
Zebedee
Ananias
Joseph (of Arimathea)
Herodias
Apollos
Simon (the Canaanite)
Martha
Jonah
James (son of Alphaeus)
Barabbas
Mary (of Bethany)
Bartholomew
Caiaphas
Philip (the apostle)
Priscilla
Aquila
Thomas
Alphaeus (father of James)
Philip (the evangelist)
Melchizedek

Theory of semantic processing

Spreading Activation (Collins and Loftus, 1975)

A node in the network (a concept/word) gets activated and activates other, nearby nodes

Activation level diminishes with distance from entry point

Links among nodes are weighted

Assumptions

Knowledge of concepts

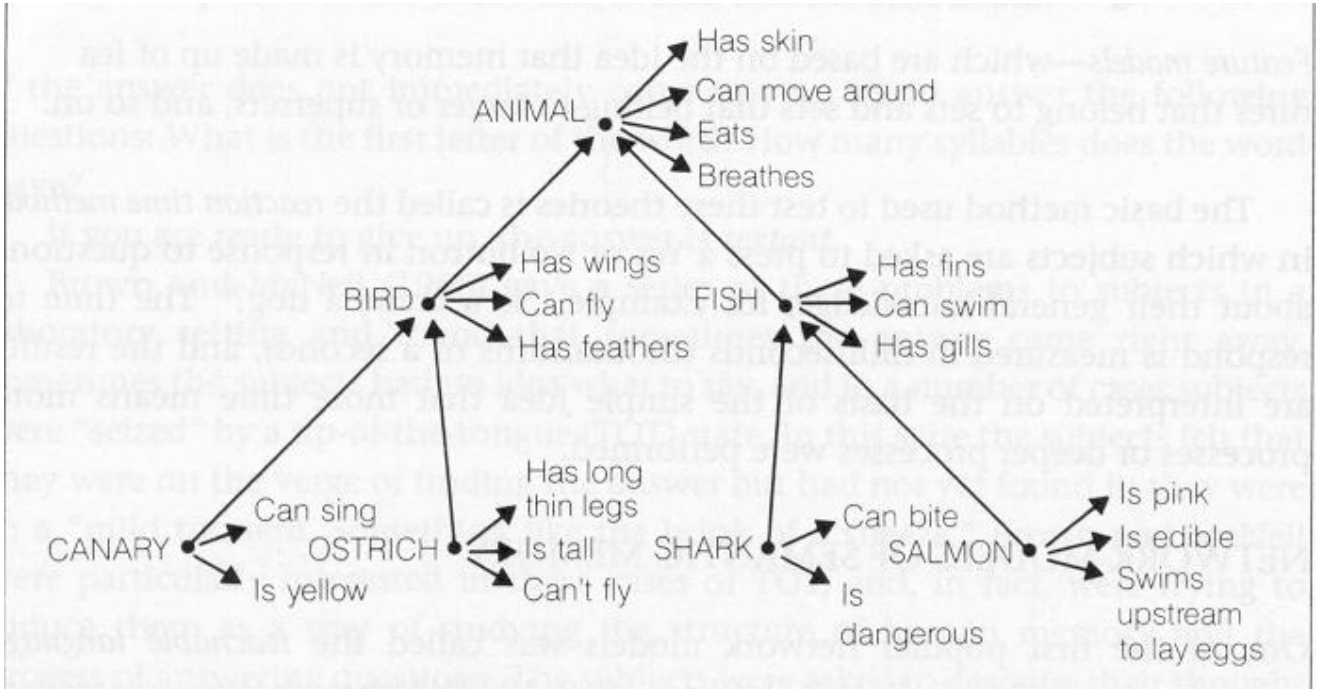
--stored economically in our minds/brains

--computed “on the fly”

--via access to general concepts

Claim: we know that “canaries fly” because “birds fly” and “canaries are a kind of bird”

Collins & Quillian Semantic network



A model of semantic organization

Knowledge is stored **only once** at the highest possible node and inherited downward (not re-stored)

animals move, birds fly, canaries sing

no redundant storage: birds move, canaries fly

unidirectional inheritance: *animals fly and sing

Collins & Quillian (1969) measured reaction times to statements involving knowledge distributed across different “levels”

Collins & Quillian experiment

Responses to statements like

Do birds move?

Do canaries move?

Do canaries have feathers?

Are canaries yellow?

Reaction times varied depending on how many nodes had to be traversed to access the information

Critique

Results are not compelling:

reaction times are influenced (at least) by

--prototypicality (how typical an exemplar of the category bird is canary?)

--word frequency (statement with robin are processed faster than with canary)

--category size (how many birds and associated information has to be searched/discarded?)

--uneven semantic distance across levels (big jump from animal to bird; smaller jump from canary to bird)

Semantic network

inspired WordNet (1986), which asked:

What would such a network look like exactly nd on a large scale?

Can most/all of the lexicon (of any language?) be represented as a semantic network?

Would some words be unconnected and left hanging in space? (If so, which ones?)

Later: crosslingual perspective

WordNet

If the (English) lexicon can be represented as a semantic network (a graph), what are the links that connect the nodes?

WN distinguishes two kinds of links

Links among nodes (concepts) are conceptual-semantic (e.g., *bird-feather*)

Links among specific words are lexical (e.g., *feather-feathery*)

Lexical links subsume conceptual-semantic links (links based on word form are also always semantic in WN)

Whence the relations? Psycholinguistic evidence

Inspection of association norms:

stimulus: *hand* reponse: *finger, arm*

stimulus: *help* response: *aid*

stimulus: *thin* response: *fat*

stimulus: *rodent* response: *rat*

Speech errors: substitution of, e.g., *week* for *day*

Data show systematic relations among words

(Also: syntagmatic and idiosyncratic relations)

Whence the relations? Distributional evidence

- Semantically related words co-occur in a given context
- Cf. Chomsky's famous example of a semantically ill-formed sentence: *colorless green ideas sleep furiously

Principle of semantic coherence within a context aids word sense disambiguation

*Knowing how to mix **drinks** at the bar is very important*

*In the U.S., admission to the bar is the granting of permission by a particular **court** system to a **lawyer** to practice **law**…*

*Police have arrested four teenagers over an attack of a 15-year-old boy involving metal bars and wooden **stakes**…*

Organizing by meaning

Lexicon-as-library metaphor

WordNet as a large-scale model of human lexical-semantic organization

Basic relation: synonymy

Each node in the semantic network is a “concept”

“Concept” is expressed by several different word forms

Synonym sets (“synsets”) are the building blocks of WordNet

{beat, hit, strike}

{car, motorcar, auto, automobile}

{ big, large}

{queue, line}

Synset members are unordered

All express/denote/refer to the same concept

WN disregards differences in frequency, connotation, register, genre…

“cognitive synonymy” (Cruse 1986)

Polysemy

WordNet gives information about two fundamental, universal properties of human language:

synonymy and **polysemy**

Synonymy = one:many mapping of meaning and form

Polysemy = one:many mapping of form and meaning

Polysemy

One word form expresses multiple meanings

{*table*, tabular_array}
{*table*, piece_of_furniture}
{*table*, mesa}
{*table*, postpone}

Polysemy in WordNet

A word form that appears in *n* synsets
is *n*-fold polysemous

{*table*, tabular_array}
{*table*, piece_of_furniture}
{*table*, mesa}
{*table*, postpone}

table is fourfold polysemous/has four senses

Some current WordNet stats

Part of speech	Word forms	Synsets
noun	117,798	82,115
verb	11,529	13,767
adjective	21,479	18,156
adverb	4,481	3,621
total	155,287	117,659

The “Net” part of WordNet

Synsets are interconnected

Bi-directional arcs express semantic relations

Result: large semantic network

(directed acyclic graph/DAG)

Relations among synsets

Based on psycholinguistic evidence, distributional properties of words

Two principal relations among concepts expressed by nouns

Already present in classical ontology (Aristotle's *Metaphysics*):

IS-A (kind/type of), hyponymy/hyperonymy:

poodle-dog-animal

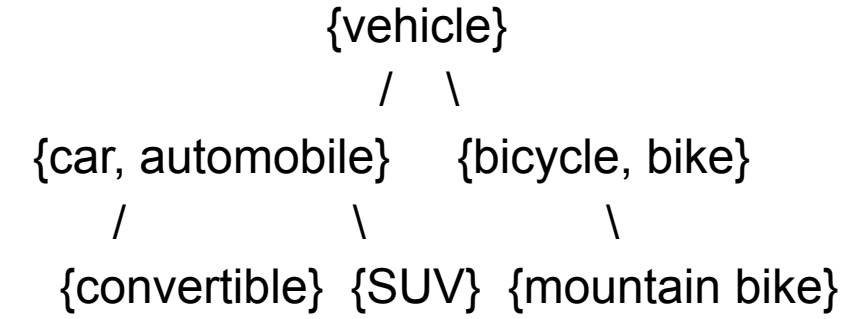
HAS-A (part-of), meronymy-holonymy:

dog-tail

Hypo-/hypernymy relates noun synsets

Relates more/less general concepts

Creates hierarchies, or “trees”



“A car is is a kind of vehicle” <=>“The class of vehicles includes cars, bikes”

Noun hierarchies can have up to 16 levels

Tree(s)

About a dozen high-level concepts:

person, animal, artifact, location, motion, communication,…

All link to a single root, *entity*

Trees can have as many as 16 level*s*

Hyponymy

Transitivity:

A car is a kind of vehicle

An SUV is a kind of car

=> *An SUV is a kind of vehicle*

Hyponymy

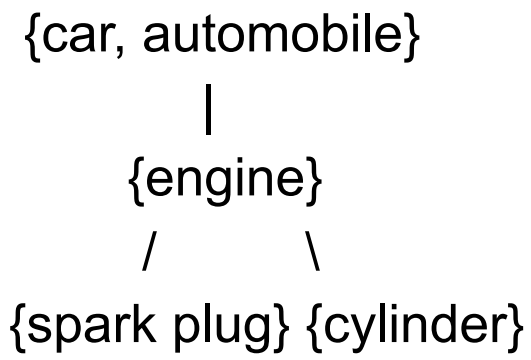
For natural species: folk taxonomy or scientific taxonomy?

Folk terms: *shrub, bush*,..

Linneus's taxonomy based on shared features is likely to be replaced by DNA-based similarity

Domain experts structure their terms differently from naïve speakers

Meronymy/holonymy (part-whole relation)



“An engine has spark plugs”

“Spark plus and cylinders are parts of an engine”

Meronymy/Holonymy

Inheritance:

A finger is part of a hand

A hand is part of an arm

An arm is part of a body

=>a finger is part of a body

(Note that statements like “a fingernail is a part of an arm” seem odd--though they are true--while others like “a fingernail is a part of the body” seem natural. Why is that?)

Meronymy

WordNet distinguishes three kinds of meronymy

Proper parts (count nouns):

arm-body, page-book, *branch-tree*

Substance/Stuff (mass nouns):

oxygen-water, flour-pizza

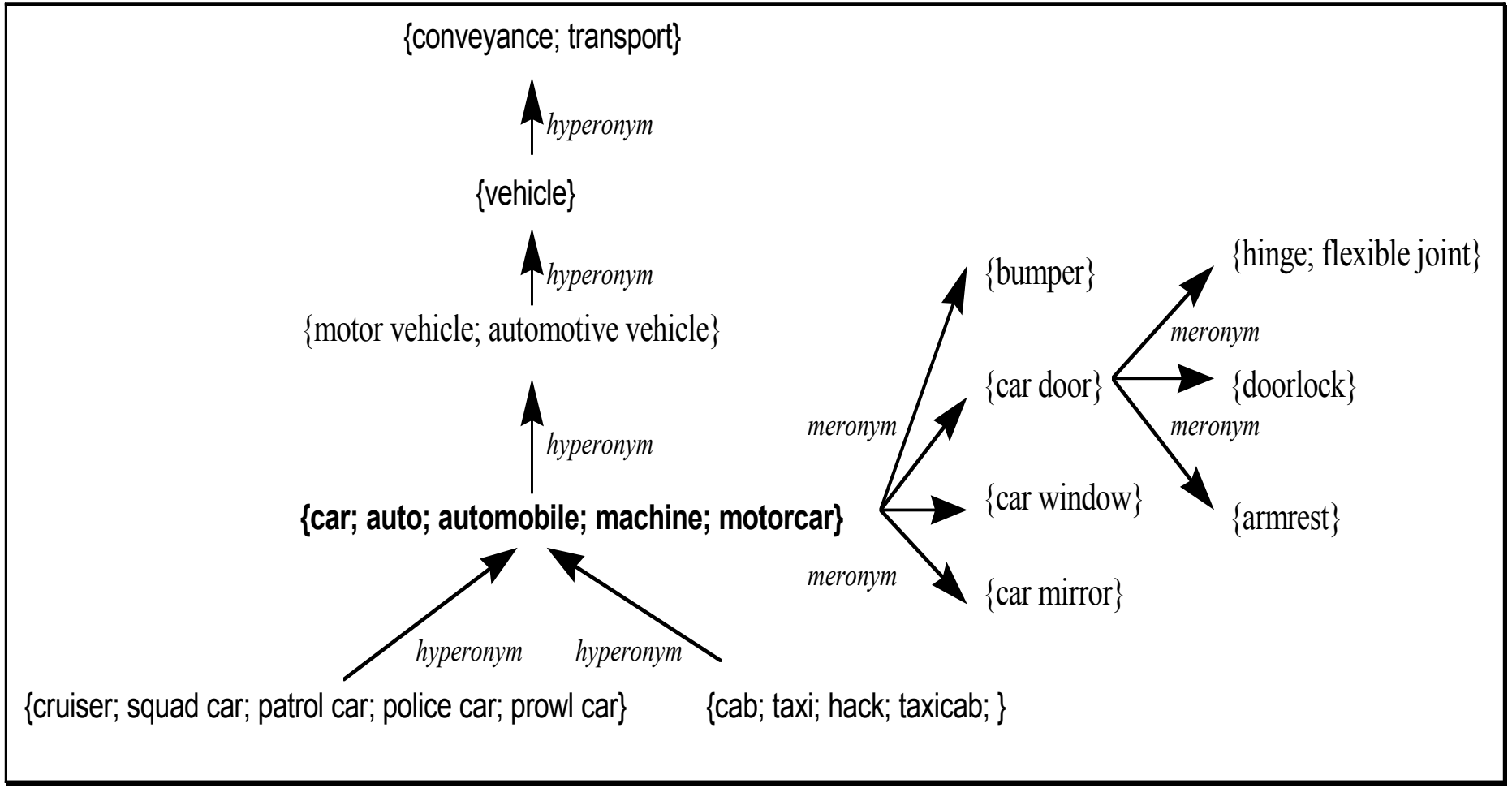
Member-group*:*

student-class, tree-forest, bird-flock

(the whole would not exist but for the members)

There are arguably more kinds of meronymy (Chaffin et al.)

Structure of WordNet (Nouns)



Classes are real

- In some cases of aphasia following a stroke, patients lose entire categories such as tools or animals
- Same with early dementia/Alzheimer's
- Neuroimaging indicates close proximity of class members in a given brain regions

Types vs. Instances

Instances are leaf nodes

Proper names

Automatically retrieved all persons, place names from WN

Manually checked whether these are instances (two people)

Some cases are hard, result in disagreement: *book ->Bible->vulgate*

Adjective relations: antonymy

Strong mutual association between members of antonymous adjective pairs:

hot-cold, old-new, high-low, big-small,...

Distributional overlap (shared selectional restrictions): what can be cold can also be hot

Highly frequent, polysemous:

High/low building/stock market/opinion/ income…

Adjective relations: antonymy

Statistically high co-occurrence in the same sentence (Justeson and Katz 1991)

Members of antonymous pairs are acquired together by children

This likely accounts for the strong mental association

Language learners want to learn both members of a pair

Adjective relations

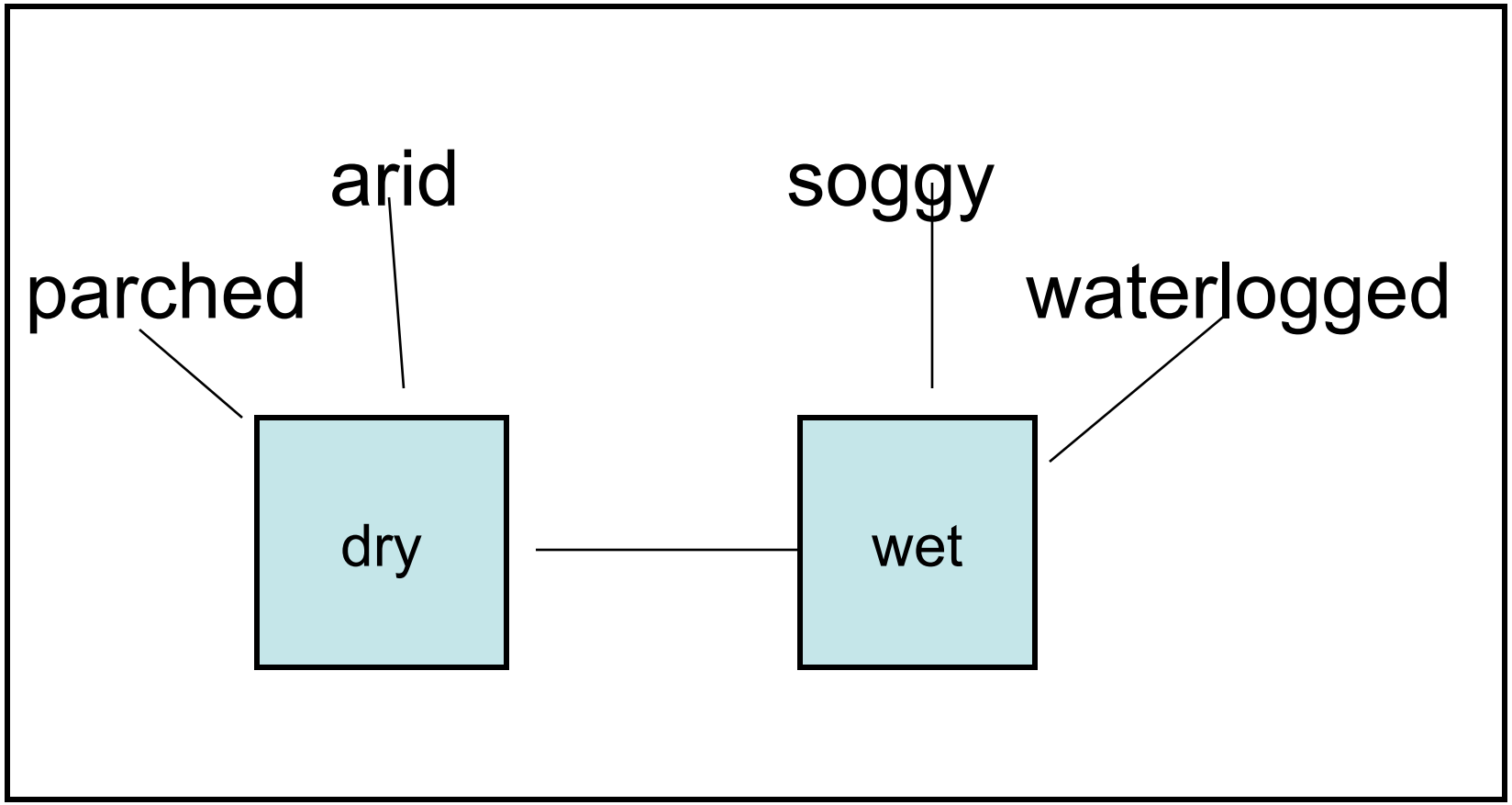
WordNet connects members of pairs like

hot-cold, long-short, new-old, wide-narrow,…

(“direct antonyms”)

For each adjective may there may be similar but less salient ones (e.g., *cool, lengthy, ancient,...)*

The “dumbbell” model



“Dumbbell” model

Direct antonyms: *dry-wet, long-short, old-new, high-low,* etc.

Indirect antonyms are “similar” to one member of the “dumbbell”

Experimental evidence

Reaction time for responses to questions like

"Is *dry* the opposite of *wet*?" (direct antonyms)

"Is *dry* the opposite of *waterlogged*?"

(direct-indirect)

"Is *arid* the opposite of *waterlogged*?"

(indirect-indirect)

Gross, Fischer, Miller (1989)

Experimental evidence

- Fastest response: direct-direct pairs
- Less fast: direct-indirect pairs
- Hesitation/slow response: indirect-indirect pairs

Problems

Some adjectives have no apparent direct or indirect antonyms (*angry, pregnant*)

Remainders

Not all adjectives fit into dumbbells

"Pertainyms" are derived from and linked in WordNet to nouns (*political-politics, nuclear-nucleus,* etc.)

Semantic relation is not specified

Current work

Explore encoding of scalar orderings for dimensional adjectives

cold<icy<arctic

{big, large}<huge<humongous

Relations among verbs

Manner relation (“troponymy”)

to x is to y in some manner

connects verbs like

move-walk, whisper-talk, smack-hit, gobble-eat

Can construct trees (not as deep as nouns):

move-run-jog-run

communicate-talk-whisper

Relations among verbs

Troponymy is polysemous: specific manner depends on verb category

Motion verbs:

Medium (air, land, water: *fly, walk, swim)*

speed: *run, jog*

Verb trees

No single top node: hundreds of flat "bushes" with no more than 5 levels

(what would a top node be and would it be useful?)

High-level nodes:

Verbs of motion, change of state, communication, cognition, contact, consumption, etc.

Other relations among verbs reflect temporal or logical order between two events

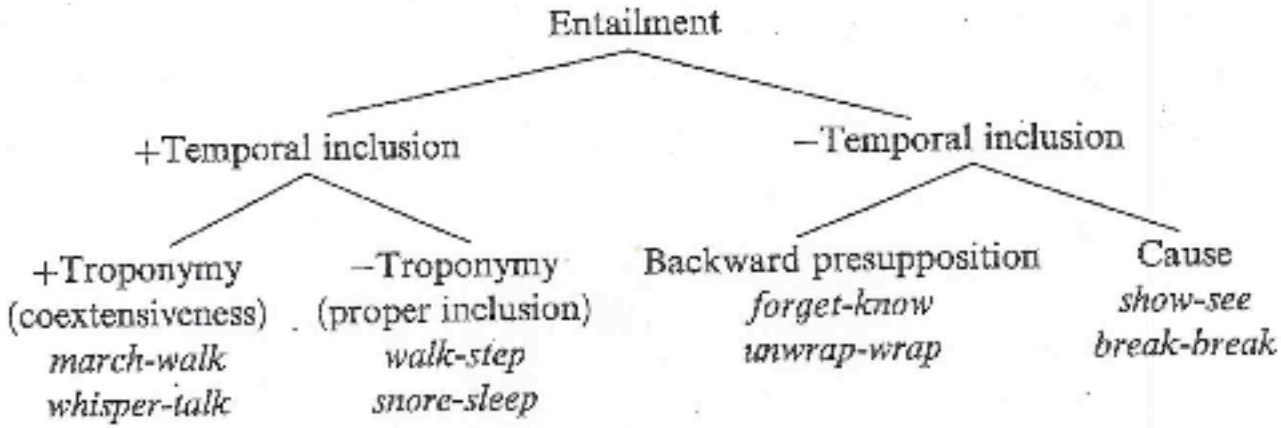
divorce-marry (backward presupposition)

snore-sleep, pay-buy (inclusion)

kill-die, fell-fall (cause)

One event unidirectionally entails the other

Entailment also holds among troponyms


Entailment
+Temporal inclusion
−Temporal inclusion
+Troponymy
(coextensiveness)
march-walk
whisper-talk
−Troponymy
(proper inclusion)
walk-step
snore-sleep
Backward presupposition
forget-know
unwrap-wrap
Cause
show-see
break-break

Is WordNet an ontology? A lexicon? A thesaurus?

Not quite either. But it's often referred to as a "lexical ontology"

Unlike a thesaurus, it makes the semantic relations explicit

Thesaurus gives you bags of words; WordNet has more structure

Basic Categories (Rosch 1976)

Categories are formed, learned by children via members

Some members are better examples than others

Categories include a prototypical member (e.g., for

American speakers, “carrot” is a prototypical member of the “vegetable” category, while “artichoke” is not)

Basic Levels (Rosch 1976)

The structure of many categories includes a “basic level”

Members at this level encode salient distinctions

E.g.,

dog, cat, horse

table, chair,bed

Basic categories

These BCs are very distinct from one another

Their hyponyms are not:

poodle, schnauzer, German shepherd,...

dining table, work table, coffee table,…

Their hypernyms are broad, underspecified (*furniture, mammal*)

Basic concepts

Universally lexicalized?

Words and concepts

Can the lexicon provide evidence for the existence of non-lexicalized concepts?

Intuitive subgroups suggest non-lexicalized superordinate category: *trams* and *trains* are different from *cars* and *motorbikes*

“vehicle on rails” vs. “wheeled vehicle”

Sorting

Words and concepts

Such “covert” categories are often lexicalized in some but not all languages

Lexical gaps?

WordNet's upper level has many "artificial" words like *unusual_person,* with hyponyms like *giant*

Is this just bad lexicography?

Are these classes, accidental gaps in English?

Ways to detect classes

Syntax can reveal categories

Two intuitive subclasses of verbs of creation:

- Verbs of creating something from a concrete material: *knit, mold*, *carve*…
- Verbs of creating something from abstract: *compose*, *formulate, concoct…*

Subclasses are real!
Revealed by syntax

All verbs have two arguments: Material and Product

All verbs can map these into the syntax

Somebody Vs Product out of/from Material

John carved a toy from the wood

Mary composed an aria from the folk song

But only the verbs with a “concrete material” allow an alternative syntax:

She carved the wood into a toy

He molded the clay into a figure

**She composed the folk melody into an aria*

**He formulated her words into a speech*

Words, concepts, categories

- Native speakers “know” this difference
- Have strong judgments about

(un)acceptable syntax

- Does this indicate the presence of two distinct unlexicalized (covert) categories, each with lexicalized members?

Another example

English allows only selected verbs to form “middle” constructions:

Chinese porcelain breaks easily

This door opens smoothly

**Birthday cards write easily*

**This door paints smoothly*

One claim: only verbs that “affect” the subject (i.e., cause it to change state) can form middles

But:

The car sold/*bought easily

The children photographed well

The book translated/read quickly

So what is the ontological status of the events denoted by the verbs that allow middle formation? (Or that of the subject?)

(Note that Romance, Germanic, Slavic languages do not have restrictions on middles.)

Words, concepts, categories

Should covert categories be represented in the lexicon? In an ontology?

Are such categories relevant for reasoning?

- So what has WN shown about the structure of the lexicon?
- Everything could be assigned a place in the network
- But relations are highly underspecified (or polysemous)