Lecture:
Lexical Semantics

CS 585, Fall 2017
Introduction to Natural Language Processing
http://people.cs.umass.edu/~brenocon/inlp2017

Brendan O'Connor
College of Information and Computer Sciences
University of Massachusetts Amherst

[most slides borrowed from J&M 3rd ed. website]

- Word sparsity is a problem!
  - Show me tweets about voting irregularities
  - Train a classifier on 50 documents
- Idea: external database of word meaning information, for word types.
  - Today: word senses and taxonomies
  - Thursday: sentiment lexicons and lexicon expansion
  - Post-midterm: vectors, word embeddings, distributional semantics

Terminology: lemma and wordform

- A **lemma** or **citation form**
  - Same stem, part of speech, rough semantics
- A **wordform**
  - The inflected word as it appears in text

Wordform	Lemma
banks	bank
sung	sing

Lemmas have senses

Lemmas have senses

- One lemma “bank” can have many meanings:
  - `…a` **`bank`** `can hold the investments in a custodial account…`
  - `“…as agriculture burgeons on the east` **`bank`** `the river will shrink even more”`
- **Sense** (or **word sense**)
  - A discrete representation

    of an aspect of a word’s meaning.
- The lemma **bank** here has two senses

Lemmas have senses

- One lemma "bank" can have many meanings:
  - Sense 1: …a **bank** can hold the investments in a custodial account…
    1
  - "…as agriculture burgeons on the east **bank** the river will shrink even more"
- **Sense** (or **word sense**)
  - A discrete representation

    of an aspect of a word's meaning.
- The lemma **bank** here has two senses

Lemmas have senses

- One lemma "bank" can have many meanings:
  - Sense 1: …a **bank**$_1$ can hold the investments in a custodial account…
  - Sense 2: "…as agriculture burgeons on the east **bank**$_2$ the river will shrink even more"
- **Sense** (or **word sense**)
  - A discrete representation

    of an aspect of a word's meaning.
- The lemma **bank** here has two senses

Homonymy

Homonyms: words that share a form but have unrelated, distinct meanings:

- bank$_1$: financial institution, bank$_2$: sloping land
- bat$_1$: club for hitting a ball, bat$_2$: nocturnal flying mammal

1. Homographs (bank/bank, bat/bat)
2. Homophones:
    1. Write and right
    2. Piece and peace

Homonymy causes problems for NLP applications

- Information retrieval
  - "bat care"
- Machine Translation
  - bat: murciélago (animal) or bate (for baseball)
- Text-to-Speech
  - bass (stringed instrument) vs. bass (fish)

Polysemy

- 1. The **bank** was constructed in 1875 out of local red brick.
- 2. I withdrew the money from the **bank**
- Are those the same sense?
  - Sense 2: “A financial institution”
  - Sense 1: “The building belonging to a financial institution”
- A **polysemous** word has **related** meanings
  - Most non-rare words have multiple meanings

Metonymy or Systematic Polysemy: A systematic relationship between senses

- Lots of types of polysemy are systematic
  - School, university, hospital
  - All can mean the institution or the building.
- A systematic relationship:
  - Building ⟺ Organization
- Other such kinds of systematic polysemy:

Author (Jane Austen wrote Emma)

Works of Author (I love Jane Austen)

Metonymy or Systematic Polysemy: A systematic relationship between senses

- Lots of types of polysemy are systematic
  - `School, university, hospital`
  - All can mean the institution or the building.
- A systematic relationship:
  - Building ↔ Organization
- Other such kinds of systematic polysemy:

Author (`Jane Austen wrote Emma`)

Works of Author (`I love Jane Austen`)

Tree (`Plums have beautiful blossoms`)

Metonymy or Systematic Polysemy: A systematic relationship between senses

- Lots of types of polysemy are systematic
  - `School, university, hospital`
  - All can mean the institution or the building.
- A systematic relationship:
  - Building ↔ Organization
- Other such kinds of systematic polysemy:

Author (`Jane Austen wrote Emma`)

↔ Works of Author (`I love Jane Austen`)

Tree (`Plums have beautiful blossoms`)

↔ Fruit (`I ate a preserved plum`)

How do we know when a word has more than one sense?

How do we know when a word has more than one sense?

- The "zeugma" test: Two senses of serve?
  - Which flights **serve** breakfast?
  - Does Lufthansa **serve** Philadelphia?
  - ?Does Lufthansa serve breakfast and San Jose?

How do we know when a word has more than one sense?

- The "zeugma" test: Two senses of `serve`?
  - `Which flights` **`serve`** `breakfast?`
  - `Does Lufthansa` **`serve`** `Philadelphia?`
  - ?Does Lufthansa serve breakfast and San Jose?
- Since this conjunction sounds weird,
  - we say that these are **two different senses of "serve"**

Word meaning relations

- Relationships between pairs of word meanings
- Synonymy: same meaning
- Antonymy: opposite meanings
- Hypernymy/hyponymy: more general/specific meanings
- Meronymy: part-whole relations
- etc.

Synonyms

- Word that have the same meaning in some or all contexts.
    - filbert / hazelnut
    - couch / sofa
    - big / large
    - automobile / car
    - vomit / throw up
    - Water / H_20
- Two lexemes are synonyms
    - if they can be substituted for each other in all situations
    - If so they have the same **propositional meaning**

Synonyms

Synonyms

- But there are few (or no) examples of perfect synonymy.
  - Even if many aspects of meaning are identical
  - Still may not preserve the acceptability based on notions of politeness, slang, register, genre, etc.

Synonyms

- But there are few (or no) examples of perfect synonymy.
  - Even if many aspects of meaning are identical
  - Still may not preserve the acceptability based on notions of politeness, slang, register, genre, etc.
- Example:
  - Water/H_20
  - Big/large
  - Brave/courageous

Synonymy is a relation between senses rather than words

Synonymy is a relation between senses rather than words

- Consider the words *big* and *large*

Synonymy is a relation between senses rather than words

- Consider the words *big* and *large*
- Are they synonyms?
  - How **big** is that plane?
  - Would I be flying on a **large** or small plane?

Synonymy is a relation between senses rather than words

- Consider the words *big* and *large*
- Are they synonyms?
  - How **big** is that plane?
  - Would I be flying on a **large** or small plane?
- How about here:
  - Miss Nelson became a kind of **big** sister to Benjamin.
  - ?Miss Nelson became a kind of **large** sister to Benjamin.

Synonymy is a relation between senses rather than words

- Consider the words *big* and *large*
- Are they synonyms?
  - How **big** is that plane?
  - Would I be flying on a **large** or small plane?
- How about here:
  - Miss Nelson became a kind of **big** sister to Benjamin.
  - ?Miss Nelson became a kind of **large** sister to Benjamin.
- Why?
  - *big* has a sense that means being older, or grown up
  - *large* lacks this sense

Antonyms

- Senses that are opposites with respect to one feature of meaning
- Otherwise, they are very similar!

  dark/light   short/long   fast/slow   rise/fall

  hot/cold   up/down   in/out

- More formally: antonyms can
  - define a binary opposition
    or be at opposite ends of a scale
    - long/short, fast/slow
  - be **reversives**:
    - rise/fall, up/down

Hyponymy and Hypernymy

- One sense is a **hyponym** of another if the first sense is more specific, denoting a subclass of the other
  - car is a hyponym of *vehicle*
  - *mango* is a hyponym of *fruit*
- Conversely **hypernym/superordinate** ("hyper is super")
  - *vehicle* is a **hypernym** of *car*
  - *fruit* is a hypernym of *mango*

Superordinate/hyper	vehicle	fruit	furniture
Subordinate/hyponym	car	mango	chair

Hyponymy more formally

- Extensional:
  - The class denoted by the superordinate extensionally includes the class denoted by the hyponym
- Entailment:
  - A sense A is a hyponym of sense B if *being an A* entails *being a B*
- Hyponymy is usually transitive
  - (A hypo B and B hypo C entails A hypo C)
- Another name: the **IS-A hierarchy**
  - A IS-A B (or A ISA B)
  - B **subsumes** A

Hyponyms and Instances

- WordNet has both **classes** and **instances**.
- An **instance** is an individual, a proper noun that is a unique entity
    - `San Francisco` is an **instance** of `city`
  - But `city` is a class
    - `city` is a **hyponym** of `municipality...location...`

Meronymy

- The part-whole relation
  - A *leg* is part of a *chair*; a *wheel* is part of a *car*.
- *Wheel* is a **meronym** of *car*, and *car* is a **holonym** of *wheel*.

Computing with a Thesaurus

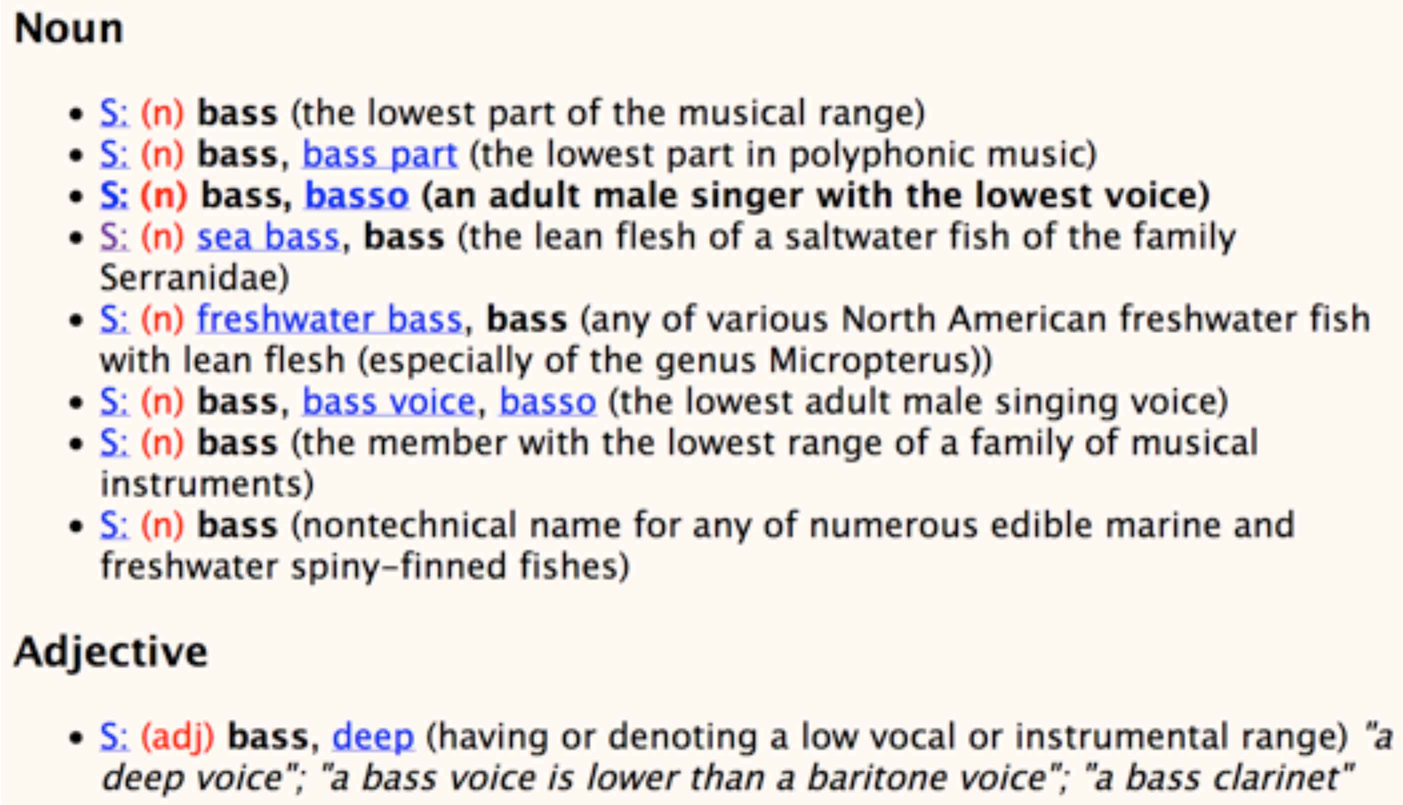
WordNet

WordNet 3.0

- A hierarchically organized lexical database
- On-line thesaurus + aspects of a dictionary
    - Some other languages available or under development
        - (Arabic, Finnish, German, Portuguese…)

Category	Unique Strings
Noun	117,798
Verb	11,529
Adjective	22,479
Adverb	4,481

Senses of “bass” in Wordnet



Noun

- S: (n) **bass** (the lowest part of the musical range)
- S: (n) **bass**, bass part (the lowest part in polyphonic music)
- **S: (n) bass, basso (an adult male singer with the lowest voice)**
- S: (n) sea bass, **bass** (the lean flesh of a saltwater fish of the family Serranidae)
- S: (n) freshwater bass, **bass** (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
- S: (n) **bass**, bass voice, basso (the lowest adult male singing voice)
- S: (n) **bass** (the member with the lowest range of a family of musical instruments)
- S: (n) **bass** (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

Adjective

- S: (adj) **bass**, deep (having or denoting a low vocal or instrumental range) *"a deep voice"; "a bass voice is lower than a baritone voice"; "a bass clarinet"*

How is “sense” defined in WordNet?

- **The synset (synonym set),** the set of near-synonyms, instantiates a sense or concept, with a gloss
- Example: chump as a noun with the gloss:

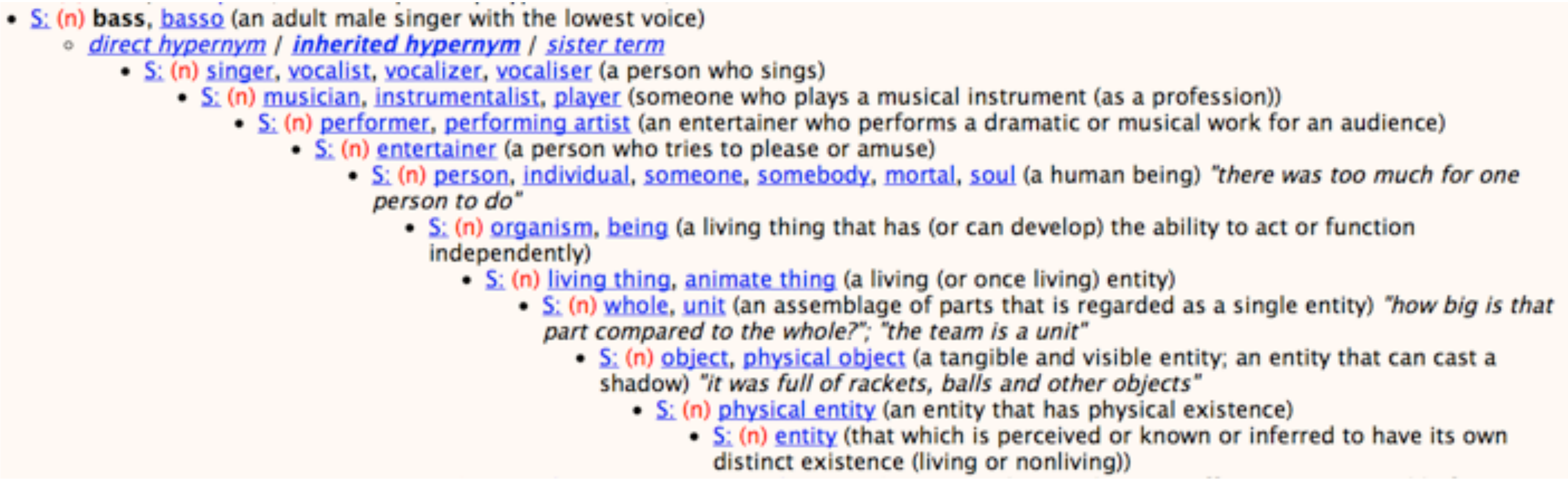
  “a person who is gullible and easy to take advantage of”

- This sense of “chump” is shared by 9 words:

  `chump`[1], `fool`[2], `gull`[1], `mark`[9], `patsy`[1], `fall guy`[1], `sucker`[1], `soft touch`[1], `mug`[2]

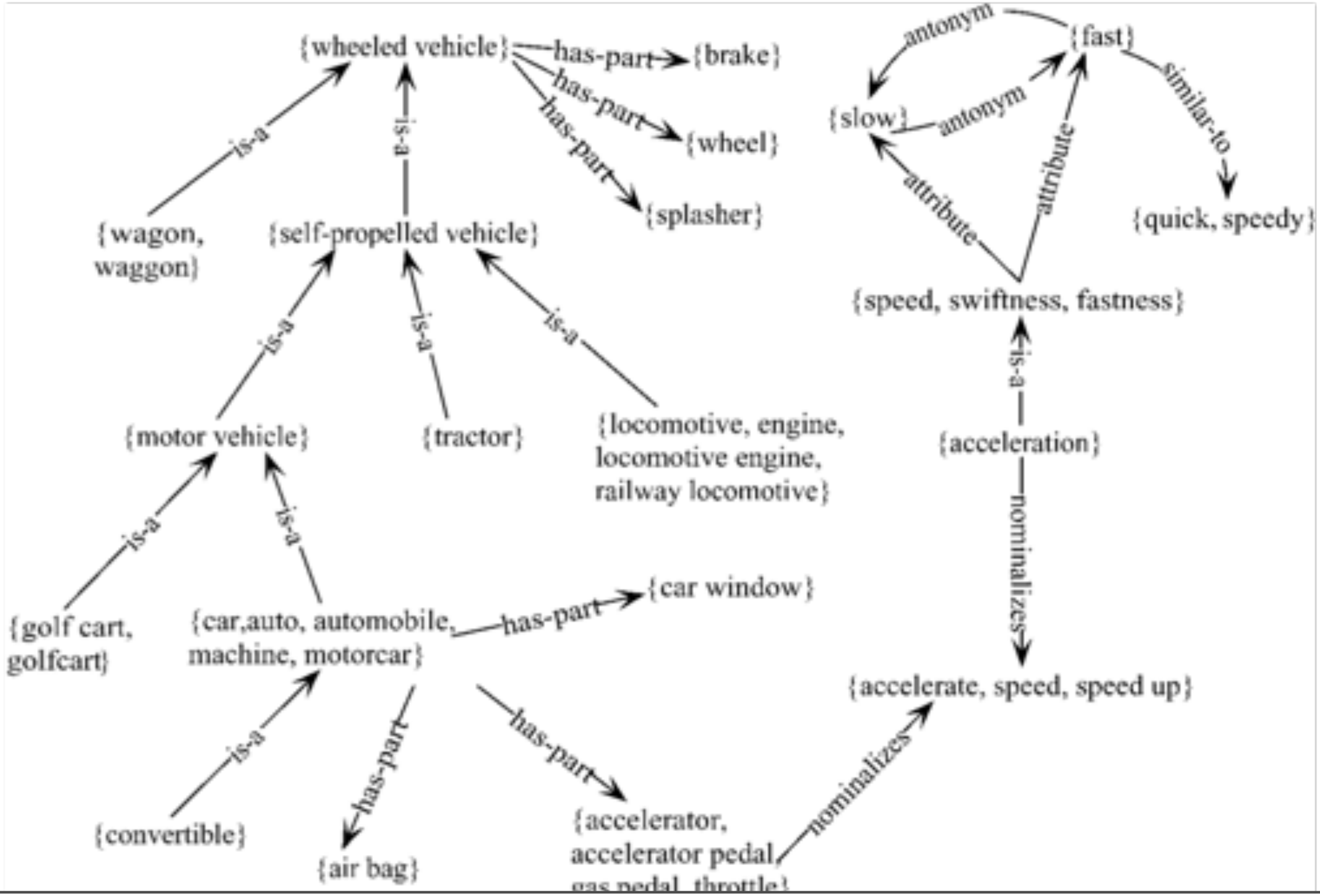
- Each of **these** senses have this same gloss
  - (Not **every** sense; sense 2 of gull is the aquatic bird)

WordNet Hypernym Hierarchy for “bass”



- S: (n) **bass**, basso (an adult male singer with the lowest voice)
    - *direct hypernym* / ***inherited hypernym*** / *sister term*
        - S: (n) singer, vocalist, vocalizer, vocaliser (a person who sings)
            - S: (n) musician, instrumentalist, player (someone who plays a musical instrument (as a profession))
                - S: (n) performer, performing artist (an entertainer who performs a dramatic or musical work for an audience)
                    - S: (n) entertainer (a person who tries to please or amuse)
                        - S: (n) person, individual, someone, somebody, mortal, soul (a human being) *"there was too much for one person to do"*
                            - S: (n) organism, being (a living thing that has (or can develop) the ability to act or function independently)
                                - S: (n) living thing, animate thing (a living (or once living) entity)
                                    - S: (n) whole, unit (an assemblage of parts that is regarded as a single entity) *"how big is that part compared to the whole?"; "the team is a unit"*
                                        - S: (n) object, physical object (a tangible and visible entity; an entity that can cast a shadow) *"it was full of rackets, balls and other objects"*
                                            - S: (n) physical entity (an entity that has physical existence)
                                                - S: (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

WordNet: Viewed as a graph



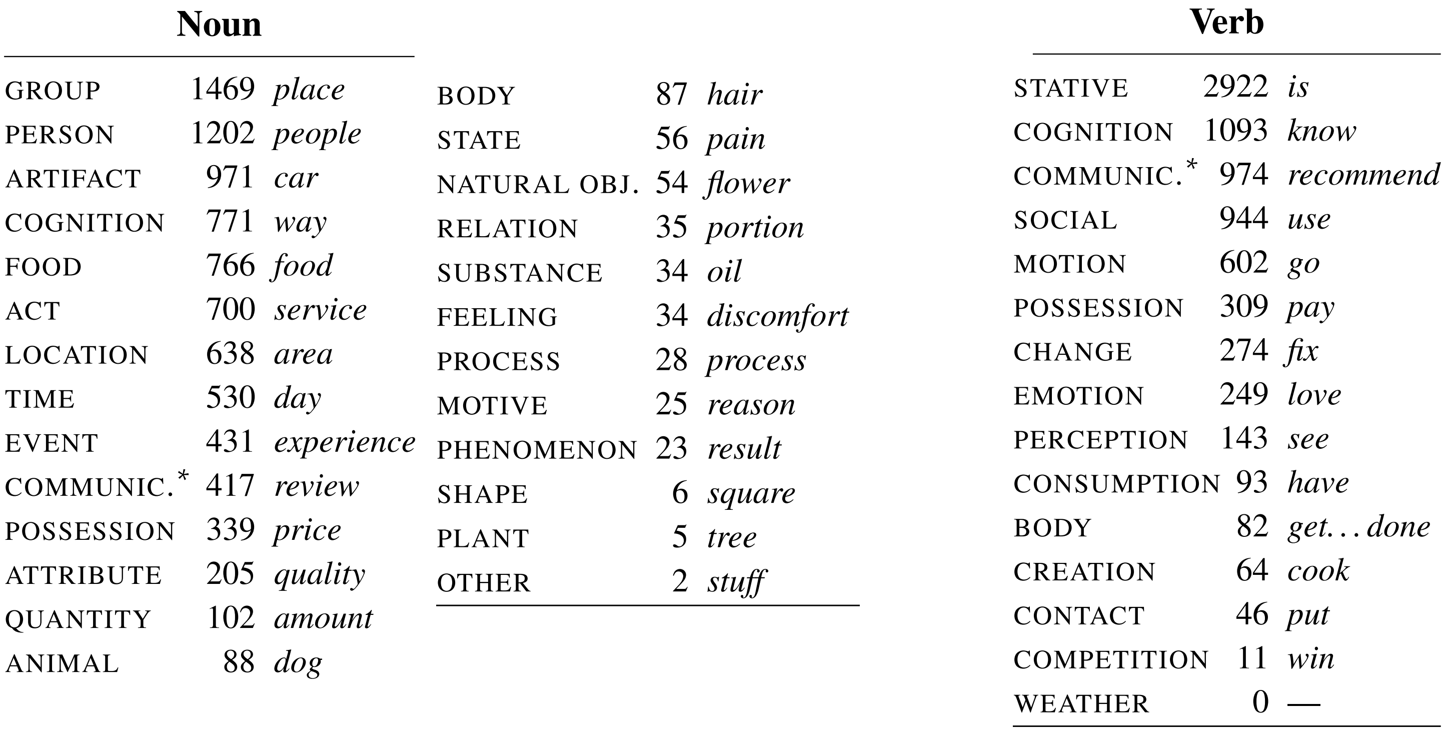
Hyponyms of “person” in WN

7588 total -- with most freq. sense restriction. *[from Michael Heilman]*

- vintager
- matrisib
- horseback rider
- ceo
- seeker
- fieldhand
- radiologist
- captain
- moujik
- research director
- damsel
- nibbler
- nailer
- nude person
- seismologist
- oddball
- prankster
- radiotherapist
- nebraskan
- cupbearer
- psychic
- accompanist
- plagiariser
- timberman
- photographer's model
- lombard
- debaser
- courtier
- dutch uncle
- schlemiel
- dizygotic twin
- mental case
- matriarch
- vocalist
- internist
- transplanter
- techie
- sniffler
- marrano
- first baseman
- government man
- child prodigy
- athenian
- hospital chaplain
- dominatrix
- bibliopole
- hombre
- east indian
- ballet master
- bad person
- rock 'n' roll musician
- flack catcher
- telephoner
- dominus
- cheater
- groveler
- accomplice
- herb doctor
- schoolfriend
- preteen
- gastronome
- concierge
- shogun
- flutist
- bottom dog
- imperialist
- emir
- libeler
- manichaean
- abnegator
- cousin-german
- masorite
- trouble maker
- villainess
- rajpoot
- calapooya
- overlord
- bank guard
- tumbler
- polycarp
- radiographer
- slave owner
- stick-in-the-mud
- audile
- deadbeat
- maltman
- jeweler
- pasha
- screwballer
- prioress
- crosspatch
- persecutor
- movie maker
- capo
- class act
- navvy
- golden boy
- sweet talker
- junior
- feminist
- villager
- specialiser
- scotsman

“Supersenses”

(counts from Schneider and Smith 2013’s Streusel corpus)



Noun

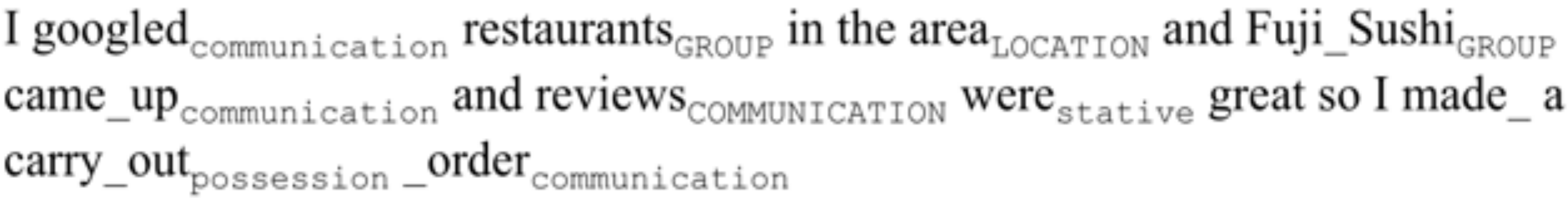
GROUP	1469	*place*
PERSON	1202	*people*
ARTIFACT	971	*car*
COGNITION	771	*way*
FOOD	766	*food*
ACT	700	*service*
LOCATION	638	*area*
TIME	530	*day*
EVENT	431	*experience*
COMMUNIC.*	417	*review*
POSSESSION	339	*price*
ATTRIBUTE	205	*quality*
QUANTITY	102	*amount*
ANIMAL	88	*dog*
BODY	87	*hair*
STATE	56	*pain*
NATURAL OBJ.	54	*flower*
RELATION	35	*portion*
SUBSTANCE	34	*oil*
FEELING	34	*discomfort*
PROCESS	28	*process*
MOTIVE	25	*reason*
PHENOMENON	23	*result*
SHAPE	6	*square*
PLANT	5	*tree*
OTHER	2	*stuff*

Verb

STATIVE	2922	*is*
COGNITION	1093	*know*
COMMUNIC.*	974	*recommend*
SOCIAL	944	*use*
MOTION	602	*go*
POSSESSION	309	*pay*
CHANGE	274	*fix*
EMOTION	249	*love*
PERCEPTION	143	*see*
CONSUMPTION	93	*have*
BODY	82	*get. . . done*
CREATION	64	*cook*
CONTACT	46	*put*
COMPETITION	11	*win*
WEATHER	0	—

Supersenses

- A word's supersense can be a useful coarse-grained representation of word meaning for NLP tasks



I googled$_{\text{communication}}$ restaurants$_{\text{GROUP}}$ in the area$_{\text{LOCATION}}$ and Fuji_Sushi$_{\text{GROUP}}$ came_up$_{\text{communication}}$ and reviews$_{\text{COMMUNICATION}}$ were$_{\text{stative}}$ great so I made_ a carry_out$_{\text{possession}}$ _order$_{\text{communication}}$

- See "STREUSEL" system http://www.cs.cmu.edu/~ark/LexSem/

- To use WordNet, or any lexical database, for NLP:
  - 1. Word Sense Disambiguation (WSD) a.k.a. Entity Linking
    - The [bank]$_3$ was open early.
  - 2. Use lexical entry information for features or inferences
    - When was that business open?
    - [bank]$_3$ <-hypo- [commercial institution]

WordNet 3.0

- Where it is:
  - http://wordnetweb.princeton.edu/perl/webwn
- Libraries
  - Python: WordNet from NLTK
  - Java: JWNL, extJWNL

MeSH: Medical Subject Headings thesaurus from the National Library of Medicine

- **MeSH (Medical Subject Headings)**
  - 177,000 entry terms that correspond to 26,142 biomedical “headings”

- **Hemoglobins**

  Entry Terms: Eryhem, Ferrous Hemoglobin, Hemoglobin

  Definition: The oxygen-carrying proteins of ERYTHROCYTES. They are found in all vertebrates and some invertebrates. The number of globin subunits in the hemoglobin quaternary structure differs between species. Structures range from monomeric to a variety of multimeric arrangements

MeSH: Medical Subject Headings thesaurus from the National Library of Medicine

- **MeSH (Medical Subject Headings)**
  - 177,000 entry terms that correspond to 26,142 biomedical "headings"

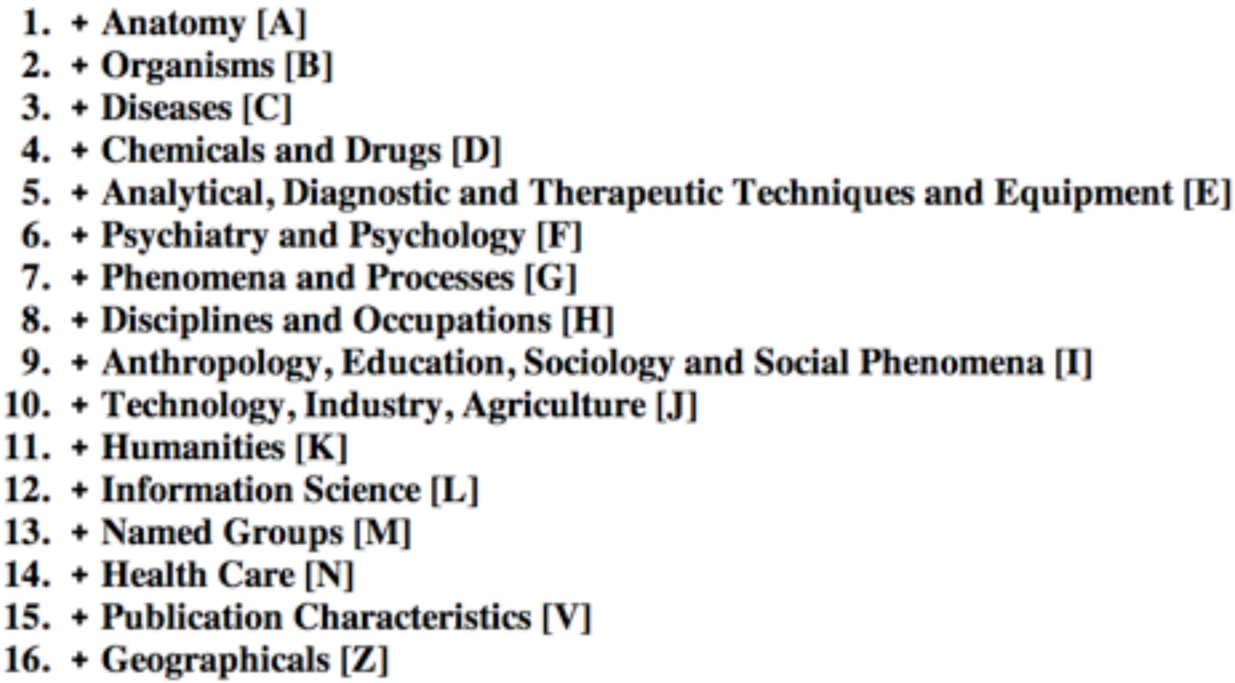
- **Hemoglobins**

  Entry Terms: Eryhem, Ferrous Hemoglobin, Hemoglobin

  Synset

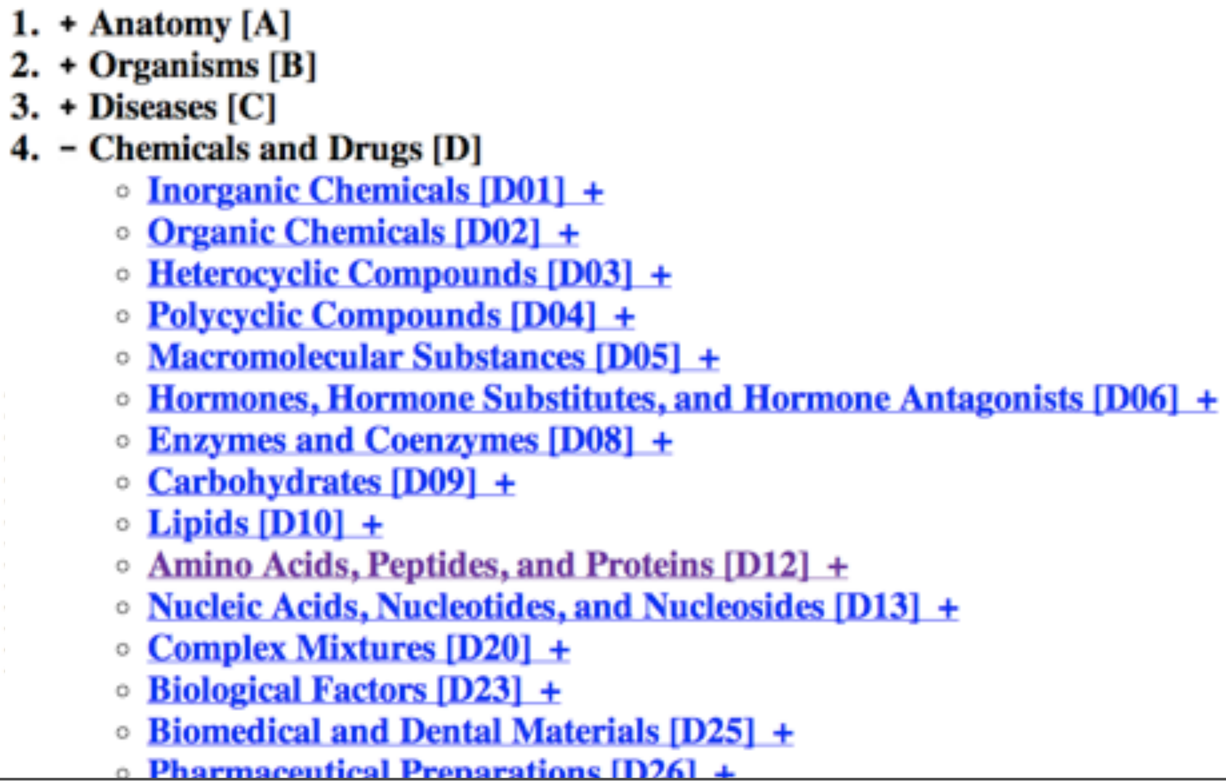
  Definition: The oxygen-carrying proteins of ERYTHROCYTES. They are found in all vertebrates and some invertebrates. The number of globin subunits in the hemoglobin quaternary structure differs between species. Structures range from monomeric to a variety of multimeric arrangements

The MeSH Hierarchy



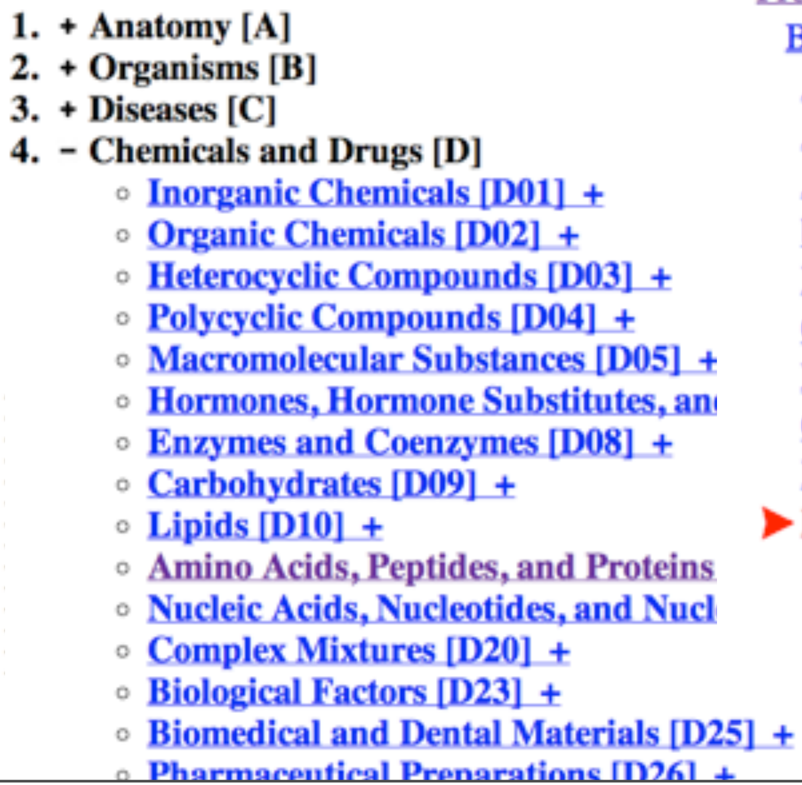
1. + Anatomy [A]
2. + Organisms [B]
3. + Diseases [C]
4. + Chemicals and Drugs [D]
5. + Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]
6. + Psychiatry and Psychology [F]
7. + Phenomena and Processes [G]
8. + Disciplines and Occupations [H]
9. + Anthropology, Education, Sociology and Social Phenomena [I]
10. + Technology, Industry, Agriculture [J]
11. + Humanities [K]
12. + Information Science [L]
13. + Named Groups [M]
14. + Health Care [N]
15. + Publication Characteristics [V]
16. + Geographicals [Z]

The MeSH Hierarchy

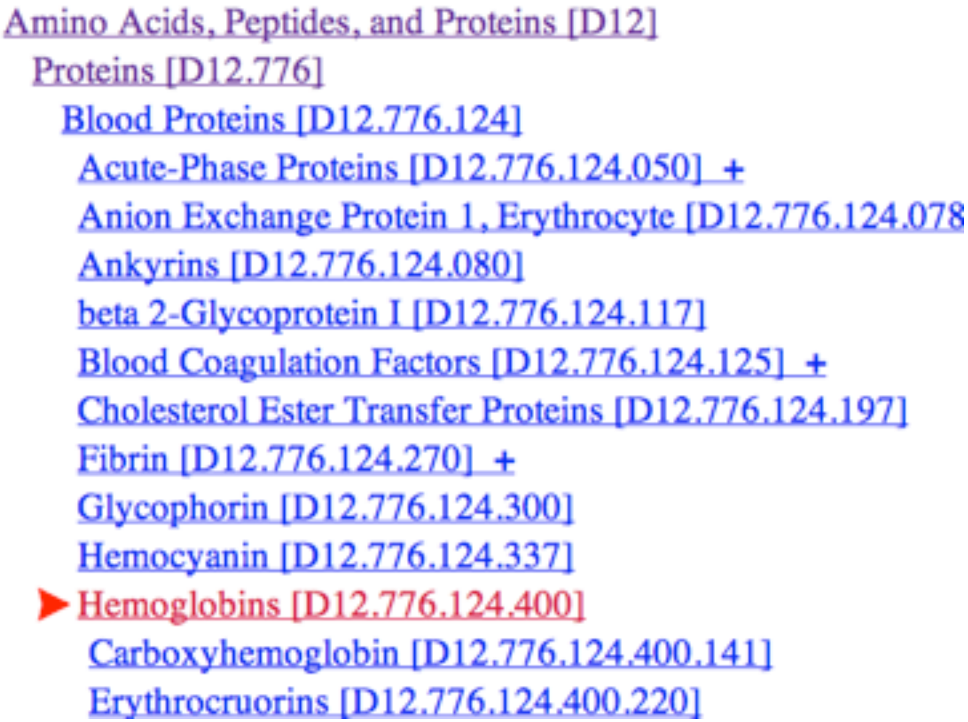


1. + Anatomy [A]
2. + Organisms [B]
3. + Diseases [C]
4. − Chemicals and Drugs [D]
    - Inorganic Chemicals [D01] +
    - Organic Chemicals [D02] +
    - Heterocyclic Compounds [D03] +
    - Polycyclic Compounds [D04] +
    - Macromolecular Substances [D05] +
    - Hormones, Hormone Substitutes, and Hormone Antagonists [D06] +
    - Enzymes and Coenzymes [D08] +
    - Carbohydrates [D09] +
    - Lipids [D10] +
    - Amino Acids, Peptides, and Proteins [D12] +
    - Nucleic Acids, Nucleotides, and Nucleosides [D13] +
    - Complex Mixtures [D20] +
    - Biological Factors [D23] +
    - Biomedical and Dental Materials [D25] +
    - Pharmaceutical Preparations [D26] +

The MeSH Hierarchy



1. + Anatomy [A]
2. + Organisms [B]
3. + Diseases [C]
4. − Chemicals and Drugs [D]
   - Inorganic Chemicals [D01] +
   - Organic Chemicals [D02] +
   - Heterocyclic Compounds [D03] +
   - Polycyclic Compounds [D04] +
   - Macromolecular Substances [D05] +
   - Hormones, Hormone Substitutes, an
   - Enzymes and Coenzymes [D08] +
   - Carbohydrates [D09] +
   - Lipids [D10] +
   - Amino Acids, Peptides, and Proteins
   - Nucleic Acids, Nucleotides, and Nucl
   - Complex Mixtures [D20] +
   - Biological Factors [D23] +
   - Biomedical and Dental Materials [D25] +
   - Pharmaceutical Preparations [D26] +



- Amino Acids, Peptides, and Proteins [D12]
  - Proteins [D12.776]
    - Blood Proteins [D12.776.124]
      - Acute-Phase Proteins [D12.776.124.050] +
      - Anion Exchange Protein 1, Erythrocyte [D12.776.124.078
      - Ankyrins [D12.776.124.080]
      - beta 2-Glycoprotein I [D12.776.124.117]
      - Blood Coagulation Factors [D12.776.124.125] +
      - Cholesterol Ester Transfer Proteins [D12.776.124.197]
      - Fibrin [D12.776.124.270] +
      - Glycophorin [D12.776.124.300]
      - Hemocyanin [D12.776.124.337]
      - ➤ Hemoglobins [D12.776.124.400]
        - Carboxyhemoglobin [D12.776.124.400.141]
        - Erythrocruorins [D12.776.124.400.220]

Uses of the MeSH Ontology

- Provide synonyms ("entry terms")
  - E.g., glucose and dextrose
- Provide hypernyms (from the hierarchy)
  - E.g., glucose ISA monosaccharide
- Indexing in MEDLINE/PubMED database
  - NLM's bibliographic database:
    - 20 million journal articles
    - Each article hand-assigned 10-20 MeSH terms

Entity-focused lexical databases

- General-domain, derived from Wikipedia
    - DBpedia http://wiki.dbpedia.org/
    - Freebase: now discontinued
- Google Knowledge Graph and other proprietary databases (Bing, Facebook, etc.)
    - Lots of relations/attributes, aimed for consumer internet use
    - Can be used directly to answer queries
    - Internally, they entity-link document texts against them

Word Similarity

- **Synonymy**: a binary relation
  - Two words are either synonymous or not
- **Similarity** (or **distance**): a looser metric
  - Two words are more similar if they share more features of meaning
- Similarity is properly a relation between **senses**
  - The word “`bank`” is not similar to the word “`slope`”
  - Bank^1 is similar to fund^3
  - Bank^2 is similar to slope^5
- But we’ll compute similarity over both words and senses

Why word similarity

- A practical component in lots of NLP tasks
  - Question answering
  - Natural language generation
  - Automatic essay grading
  - Plagiarism detection
- A theoretical component in many linguistic and cognitive tasks
  - Historical semantics
  - Models of human word learning
  - Morphology and grammar induction

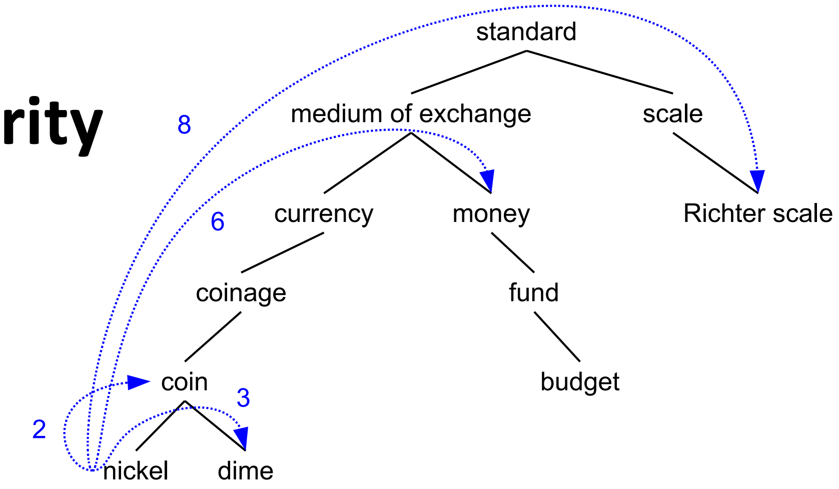
Word similarity and word relatedness

- We often distinguish **word similarity** from **word relatedness**
  - **Similar words**: near-synonyms
  - **Related words**: can be related any way
    - `car, bicycle:` **similar**
    - `car, gasoline:` **related**, not similar

Two classes of similarity algorithms

- Thesaurus-based algorithms
  - Are words “nearby” in hypernym hierarchy?
  - Do words have similar glosses (definitions)?
- Distributional algorithms
  - Do words have similar distributional contexts?
  - Distributional (Vector) semantics on Thursday!

Path based similarity



- Two concepts (senses/synsets) are similar if they are near each other in the thesaurus hierarchy
  - =have a short path between them
  - concepts have path 1 to themselves

Evaluating similarity

- Extrinsic (task-based, end-to-end) Evaluation:
  - Question Answering
  - Spell Checking
  - Essay grading
- Intrinsic Evaluation:
  - Correlation between algorithm and human word similarity ratings
    - Wordsim353: 353 noun pairs rated 0-10. *sim(plane,car)=5.77*
  - Taking TOEFL multiple-choice vocabulary tests
    - `Levied is closest in meaning to:`

      `imposed, believed, requested, correlated`