

WHAT DO WORDS MEAN?

- ▶ Referents in the physical or social world
 - ▶ But not usually useful in text analysis
- ▶ Their dictionary definition
 - ▶ But dictionary definitions are necessarily circular
 - ▶ Only useful if meaning is already understood
- ▶ Their relationships with other words
 - ▶ Also circular, but more practical

THREE KINDS OF SEMANTICS

- ▶ Lexical semantics
 - ▶ How the meanings of words relate to one another
- ▶ Distributional semantics
 - ▶ How words that appear together reflect their meaning
- ▶ Logical semantics
 - ▶ How facts about the world are represented

WORDS AND SENSES

- ▶ Orthographic word type is wrong unit for semantic analysis

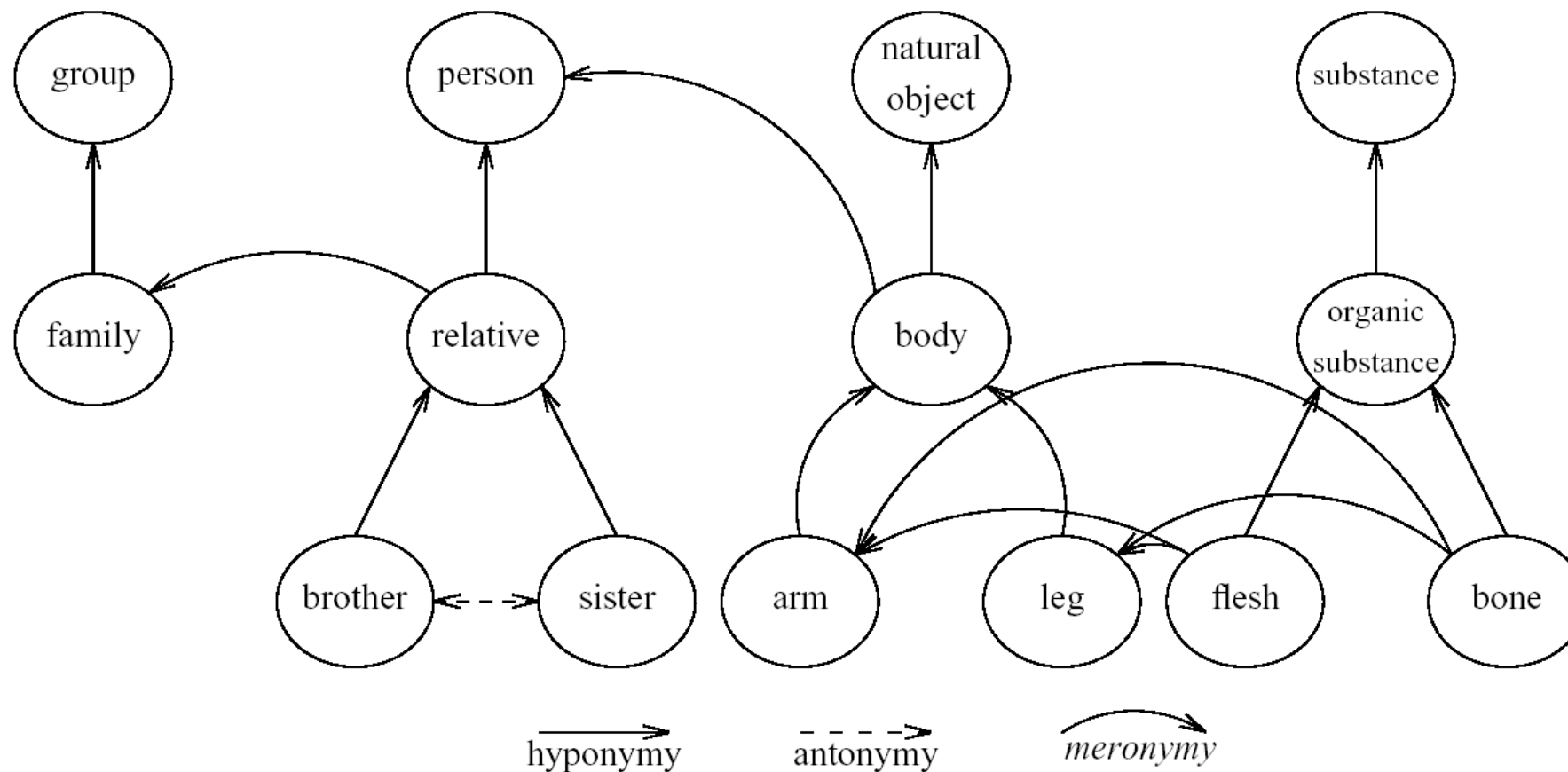
Bank (noun):

- 1. A financial institution; a building where a financial institution offers services; a repository; a container for holding money*
- 2. Land sloping down to a body of water*

- ▶ *Bank* has many senses (more than just these)
- ▶ 1 and 2 are *homonyms*
 - ▶ Considered different lexical items by lexicographers
- ▶ 1 shows *polysemy*
 - ▶ Related senses of the same lexical item

BASIC LEXICAL RELATIONS

- ▶ Synonyms (same) and antonyms (opposite/complementary)
- ▶ Hypernyms (generic), hyponyms (specific)
- ▶ Meronyms (part) and holonyms (whole)



WORDNET

- ▶ A database of lexical relations
- ▶ English WordNet includes ~120,000 nouns, ~12,000 verbs, ~12,000 adverbs, ~4,000 adverbs
- ▶ WordNets available in most major languages (www.globalwordnet.org)
- ▶ English version freely available (accessible via NLTK)

SYNSETS

- ▶ The nodes of WordNet are not words, but meanings
- ▶ They are represented by sets of synonyms, or *synsets*

```
>>> nltk.corpus.wordnet.synsets('bank')
```

```
[Synset('bank.n.01'), Synset('depository_financial_institution.n.01'), Synset('bank.n.03'),  
Synset('bank.n.04'), Synset('bank.n.05'), Synset('bank.n.06'), Synset('bank.n.07'),  
Synset('savings_bank.n.02'), Synset('bank.n.09'), Synset('bank.n.10'), Synset('bank.v.01'),  
Synset('bank.v.02'), Synset('bank.v.03'), Synset('bank.v.04'), Synset('bank.v.05'), Synset('deposit.v.02'),  
Synset('bank.v.07'), Synset('trust.v.01')]
```

```
>>> nltk.corpus.wordnet.synsets('bank')[0].definition()
```

```
u'sloping land (especially the slope beside a body of water)'
```

```
>>> nltk.corpus.wordnet.synsets('bank')[1].lemma_names()
```

```
[u'depository_financial_institution', u'bank', u'banking_concern', u'banking_company']
```

LEXICAL RELATIONS IN WORDNET

- ▶ Connections between nodes are lexical relations
- ▶ Including all the major ones mentioned earlier

```
>>> nltk.corpus.wordnet.synsets('relative')[0].hypernyms()
```

```
[Synset('person.n.01')]
```

```
>>> nltk.corpus.wordnet.synsets('body')[0].part_meronyms()
```

```
[Synset('arm.n.01'), Synset('articulatory_system.n.01'), Synset('body_substance.n.01'),  
Synset('cavity.n.04'), Synset('circulatory_system.n.01'), Synset('crotch.n.02'),  
Synset('digestive_system.n.01'), Synset('endocrine_system.n.01'), Synset('head.n.01'), Synset('leg.n.01'),  
Synset('lymphatic_system.n.01'), Synset('musculoskeletal_system.n.01'), Synset('neck.n.01'),  
Synset('nervous_system.n.01'), Synset('pressure_point.n.01'), Synset('respiratory_system.n.01'),  
Synset('sensory_system.n.02'), Synset('torso.n.01'), Synset('vascular_system.n.01')]
```

```
>>> print nltk.corpus.wordnet.lemmas('sister')[0].antonyms()
```

```
[Lemma('brother.n.01.brother')]
```


WORD SIMILARITY WITH PATHS

- ▶ Want to go beyond specific lexical relations
 - ▶ E.g. *money* and *nickel* are related, despite no direct lexical relationship
- ▶ Given WordNet, find similarity based on path length in hypernym/hyponym tree

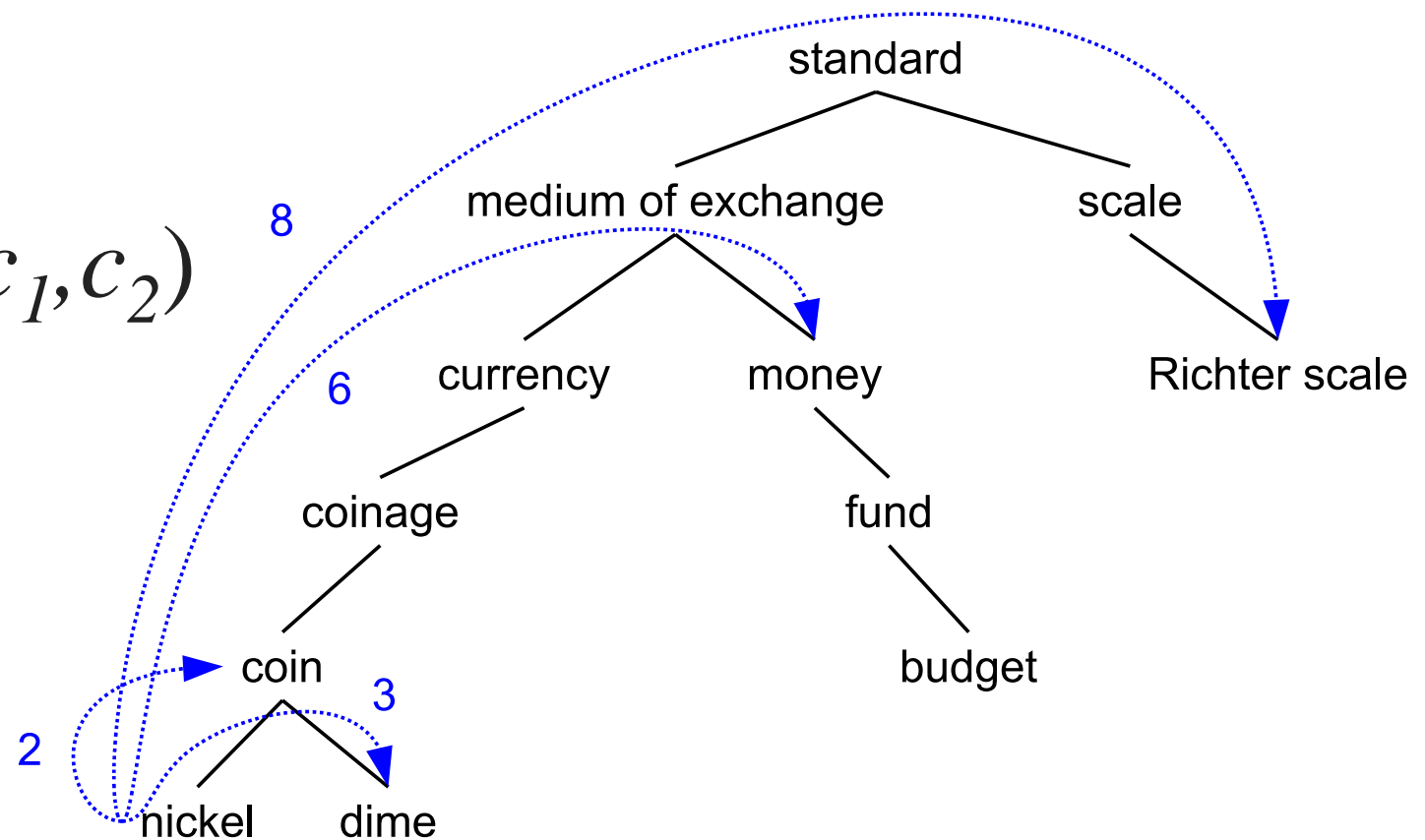
$$\text{simpath}(c_1, c_2) = 1/\text{pathlen}(c_1, c_2)$$

$$\text{simpath}(\textit{nickel}, \textit{coin}) = 1/2 = .5$$

$$\text{simpath}(\textit{nickel}, \textit{currency}) = 1/4 = .25$$

$$\text{simpath}(\textit{nickel}, \textit{money}) = 1/6 = .17$$

$$\text{simpath}(\textit{nickel}, \textit{Richter scale}) = 1/8 = .13$$



BEYOND PATH LENGTH

- ▶ Problem: edges vary widely in actual semantic distance
 - ▶ Much bigger jumps near top of hierarchy
- ▶ Solution 1: include depth information (Wu & Palmer)
 - ▶ Use path to find lowest common subsumer (LCS)
 - ▶ Compare using depths

$$\text{simwup}(c_1, c_2) = \frac{2 * \text{depth}(\text{LCS}(c_1, c_2))}{\text{depth}(c_1) + \text{depth}(c_2)}$$

$$\text{simwup}(\textit{nickel}, \textit{money}) = 2 * 2 / (3 + 6) = .44$$

$$\text{simwup}(\textit{nickel}, \textit{Richter scale}) = 2 * 1 / (3 + 6) = .22$$

INFORMATION CONTENT

- ▶ But count of edges is still poor semantic distance metric
- ▶ Solution 2: include statistics from corpus (Resnik; Lin)
 - ▶ $P(c)$: probability that word in corpus is instance of concept c

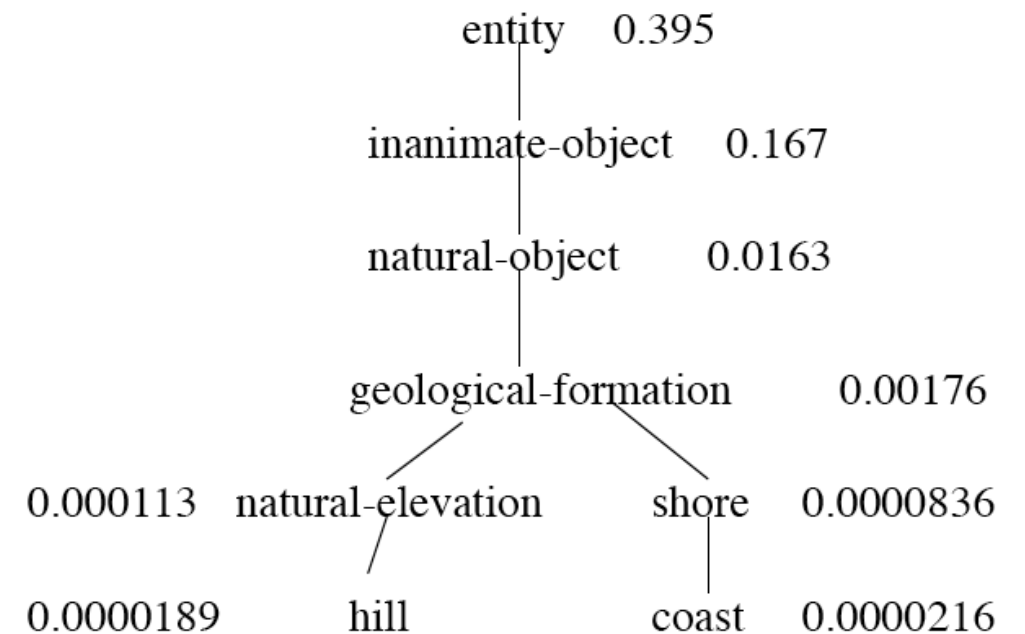
$$P(c) = \frac{\sum_{w \in \text{words}(c)} \text{count}(w)}{N}$$

- ▶ information content (IC)

$$IC(c) = -\log P(c)$$

- ▶ Lin distance

$$\text{simlin}(c_1, c_2) = \frac{2 * IC(\text{LCS}(c_1, c_2))}{IC(c_1) + IC(c_2)}$$



WORD SENSE DISAMBIGUATION

- ▶ But how to go from words in text to senses in WordNet?
 - ▶ Hacky (but popular) “solutions”:
 - ▶ Assume the most popular sense
 - ▶ For word similarity, take minimum across senses
 - ▶ The proper (but difficult) solution: Word Sense Disambiguation
- ▶ WordNet as sense inventory a great resource for WSD
 - ▶ Also problematic, because too comprehensive
- ▶ Beyond WordNet, good WSD useful for many tasks in NLP
 - ▶ In practice, often ignored because good WSD too hard

SUPERVISED WSD

- ▶ Apply standard machine classifiers such as SVMs
- ▶ Requires sense tagged corpora
 - ▶ E.g. SENSEVEL, SEMCOR (available in NLTK)
 - ▶ Very time consuming to create!
- ▶ Feature vectors typically words and syntax around target
 - ▶ But context is ambiguous too!
 - ▶ How big should context window be? (typically very small)

LESS SUPERVISED APPROACHES

- ▶ Lesk: Use overlap of context with dictionary gloss from WN
- ▶ Yarowsky: Bootstrap method
 - ▶ Create a small seed training set
 - ▶ E.g for *plant* (factory vs. vegetation): *manufacturing plant*, *plant life*
 - ▶ Iteratively expand training set with untagged examples by
 - ▶ Assuming *one sense per collocation*, *one sense per document*
 - ▶ Training a statistical classifier on current training set, add confidently predicted examples to training set
- ▶ Graph methods in WordNet
- ▶ Clustering for Word Sense Induction

FRAMENET

- ▶ A lexical data base of *frames*, typically prototypical situations
 - ▶ E.g. “apply_heat” frame
- ▶ Includes lists of *lexical units* that evoke the frame
 - ▶ E.g. *cook, fry, bake, boil*, etc.
- ▶ Lists of *semantic roles* or *frame elements*
 - ▶ E.g. “the cook”, “the food”, “the container”, “the instrument”
 - ▶ Distinguish between core and non-core roles
- ▶ Semantic relationships among frames
 - ▶ “apply_heat” is Causative of “absorb_heat”, is Used by “cooking_creation”

OTHER USEFUL LEXICONS IN NLTK

- ▶ Names: List of male and female names
- ▶ Gazetteer List: lists of cities and countries
 - ▶ Comprehensive lists of locations at www.geonames.org
- ▶ WordList: lists of words for various languages
- ▶ Stopwords: list of stopwords for various languages
- ▶ CmuDict: a pronunciation dictionary

LEXICONS FOR TEXT ANALYSIS

- ▶ General Inquirer lexicon
 - ▶ Large words tagged for 150+ categories
 - ▶ Tags for psychological, social, and topic distinctions
 - ▶ Best known in NLP for positive/negative tags
- ▶ Linguistic Inquiry and Word Count (LIWC) lexicons
 - ▶ Largest and most well known text analysis tool
 - ▶ Major lexical categories: affect; social; cognitive processes; perpetual processes; biological processes; core drives and needs; time orientation; relativity; personal concerns; informal speech

MULTIWORD LEXICONS

- ▶ Many lexical items involve multiple words
 - ▶ Semantically *non-compositional* (United States \neq United + States)
 - ▶ Sometimes non-contiguous (*take him/her/them for a ride*)
- ▶ Both WordNet and FrameNet contain *multiword expressions* (MWEs)
 - ▶ But far from comprehensive
- ▶ In fact, no comprehensive collection of MWEs exists
 - ▶ MWE/collocation identification is a classic NLP task
- ▶ See <http://www.cs.cmu.edu/~ark/LexSem/> for a good collection taken from various sources

MOVING ON TO THE CORPUS

- ▶ Manually-tagged lexical resources an important starting point for text analysis
- ▶ But much modern work attempts to derive semantic information directly from corpora, without human intervention
- ▶ Let's add some distributional information

FURTHER READING

- ▶ JM2 19.1-19.4 (lexical semantics and Word/FrameNet)
- ▶ JM2 20.1-20.6 (WSD and word similarity)