Plan for Today's Lecture(s)

- Introduction to Relationships and Structures
- Semantic Perspective
- Lexical Perspective

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

3 October 2013
Lecture 11.1 – Introduction to Relationships and Structures

Defining "Relationship"

- “An association among several things, with that association having a particular significance”
- “Relationships are the stuff out of which information is made”
- The reason is an important part of the relationship
- Multiple relationships can exist among the same objects, so the order of the objects matters

Relationship != Relation

- In mathematics and computer science a “relation” is a set of ordered elements (“tuples”) of equal degree
- A binary relation is a set of element pairs, a ternary relation is a set of 3-tuples, and so on.
- The elements in each tuple are “related” but they do not need to have any “significant association” or “relationship” among them

Five Perspectives on Relationships (1)

- SEMANTIC: the meaning of the association
- LEXICAL: how the conceptual description of a relationship is expressed using words in a specific language
- STRUCTURAL: analyzes the patterns of association, arrangement, proximity, or connection between resources

Five Perspectives on Relationships (1)

- ARCHITECTURAL: emphasizes the number and abstraction level of the components of a relationship, which together characterize its complexity
- IMPLEMENTATION: how the relationship is implemented in a particular notation and syntax and the manner in which relationships are arranged and stored in some technology environment.

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

3 October 2013
Lecture 11.2 – The Semantic Perspective on Relationships

The Semantic Perspective on Relationships

- The essence of relationships; why the resources are related
- Relies on or expresses information that is not directly available from perception
- “Homer is married to Marge" is a semantic assertion; "Homer is standing next to Marge" is not
- It is difficult to define what a relationship means in a non-circular way that doesn't rely on lots of undefined relationships...

Making Sense (1)

I saw :	with :
a Man	a Telescope
a Star	a Microscope
a Molecule	Binoculars

How many of these 9 combinations make sense?

Making Sense (2)

- Draw a diagram that represents these assertions:
  - "Mary and Sue are sisters“
  - "Sally and Martha are mothers”
- Now try:
  - “Bob saw the plane flying over Denver”
  - “Bob saw the mountains flying over Denver”

Language and Meaning

- Words and sentence structure only hint at meaning
- Meaning is constructed not just from the literal language used, but from all the clues or cues in the context of use -- common knowledge, assumptions, previous discourse, the present situation, and inferences from all of these

Specifying Semantic Relationships

- "A sequence of categories, that includes one thing from each category“ (Kent)
- Predicate/Argument Syntax:

  is-married-to (Homer Simpson, Marge Simpson)

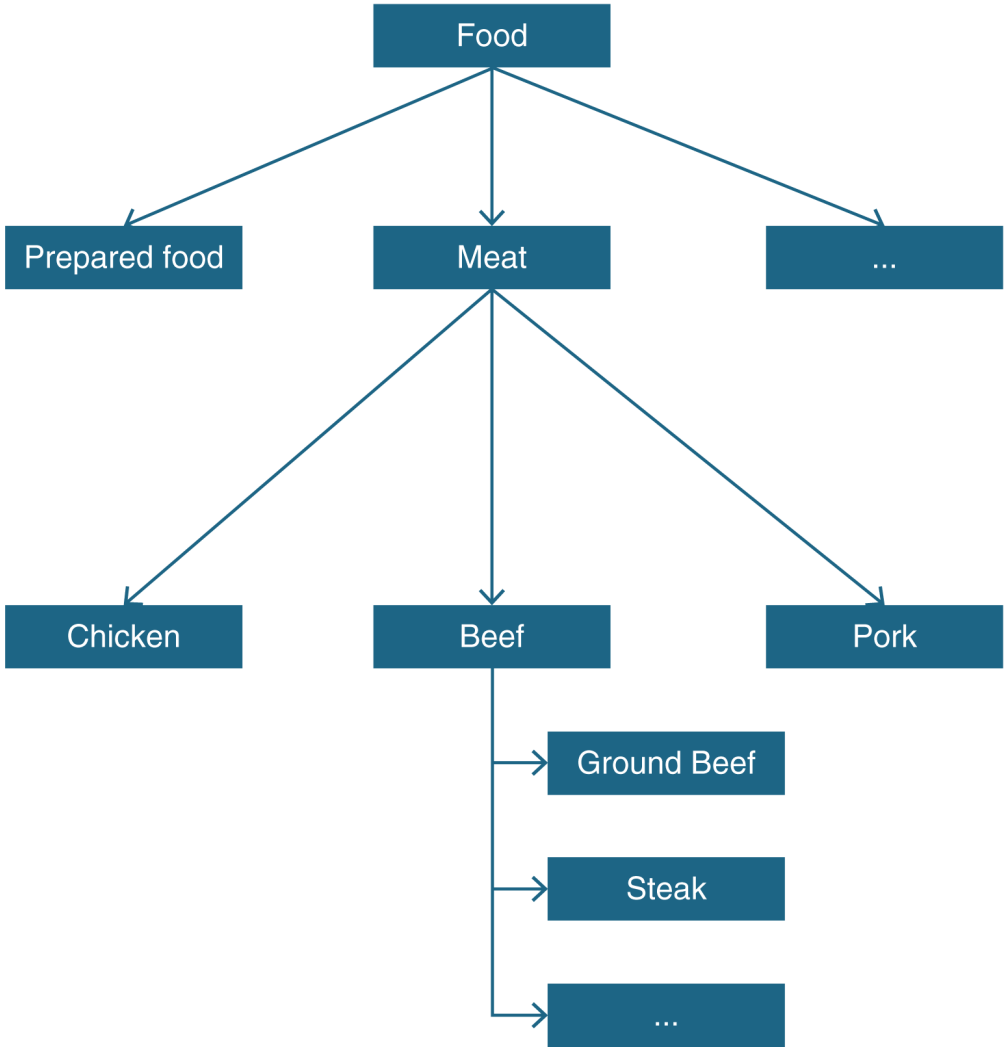
- Subject / Predicate / Argument Syntax:

  Homer Simpson => is-married-to => Marge Simpson

The Semantics of "Inclusion"

- One class or type of resources logically contains another; predicate is usually *is-a* or "*is-a-type-of*"
- A set of interconnected class inclusion relationships defines a HIERARCHY or TAXONOMY
- An inclusion relationship between an instance and a class -- assigning an instance to a class - is CLASSIFICATION

A PARTIAL TAXONOMY OF FOOD

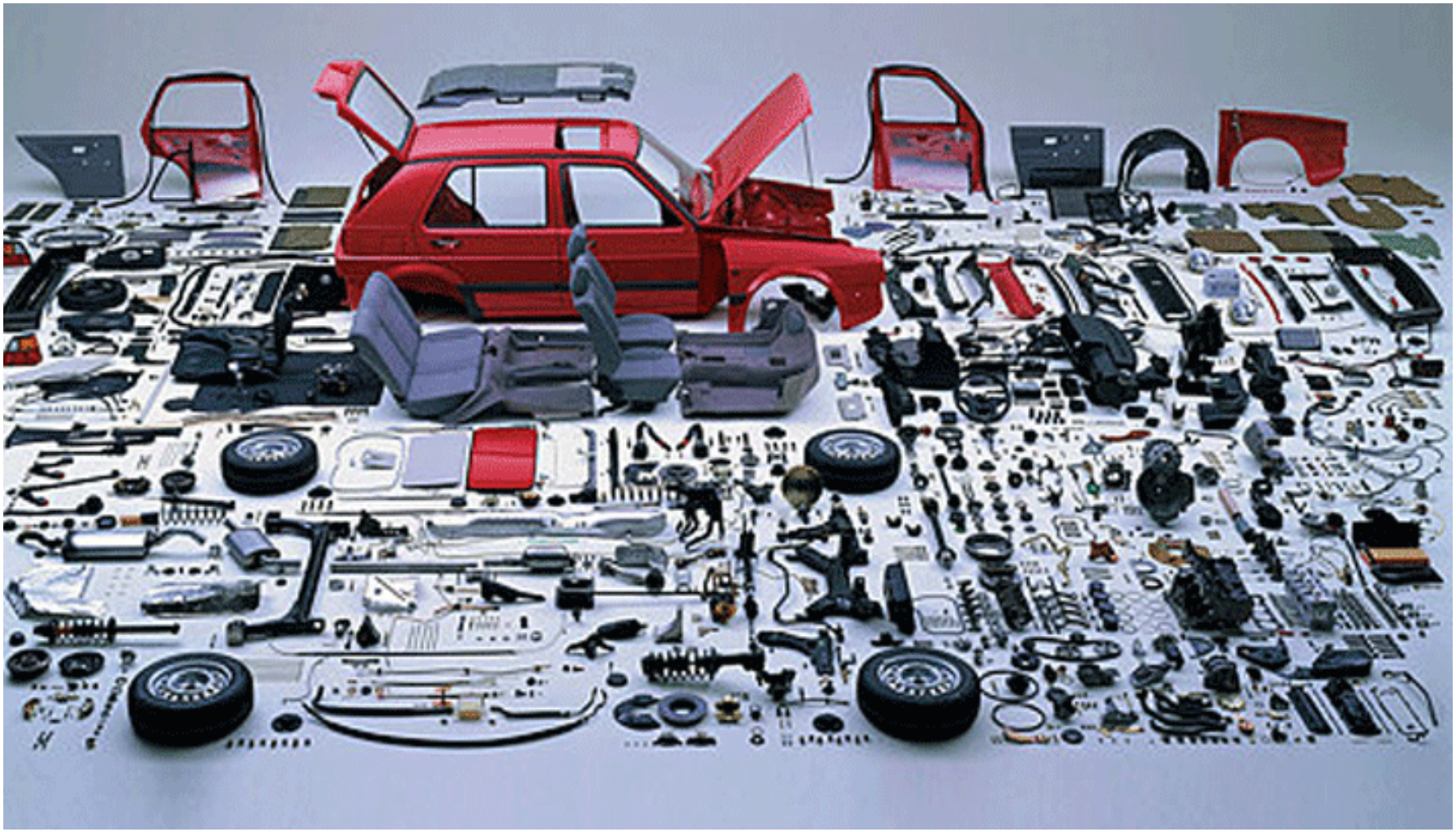


TDO
Figure 5.1

Part-Whole or Meronymic Inclusion

- A resource class or instance physically, topologically, or temporally contains another
  - The engine is part of the car
  - The Napa Valley is part of Northern California
  - Extra time is part of a football match
- expressed using “is-part-of,” “is-partly,” or with other similar predicate expressions

Part-Whole Relationship (Component-Object)



The Semantics of "Attribution"

- In contrast to inclusion expressions that state relationships between resources, attribution relationships assert or assign values to properties for a particular resource
- Attribution can be thought of as a binary relationship whose predicate has a single argument:

  Martin the Gecko => is-small

  "Title" : "Moby Dick"
- Or with two arguments:

  Martin => has-size => small

Properties of Semantic Relationships: Symmetry (1)

- SYMMETRY: the order in which the arguments of a binary relationship doesn't matter

  Homer Simpson => is-married-to => Marge Simpson

  Marge Simpson => is-married-to => Homer Simpson

Properties of Semantic Relationships: Symmetry (2)

- Relationships that are symmetric are also called BI-DIRECTIONAL
- ASYMMETRIC relationships are those where the order of the arguments matters

Homer Simpson => is-parent-of => Bart Simpson

Properties of Semantic Relationships: Transitivity (1)

- TRANSITIVITY: if X and Y have a relationship, and Y and Z have the same relationship, then X also has the relationship with Z

Homer Simpson => is-taller-than => Bart Simpson

Bart Simpson => is-taller-than => Maggie Simpson

IMPLIES THAT

Homer Simpson => is-taller-than => Maggie Simpson

Properties of Semantic Relationships: Transitivity (2)

- Any relationship based on inclusion or ordering is transitive
- Ordering includes numerical, alphabetic, and chronological relationships as well as those that imply qualitative or quantitative measurement

Properties of Semantic Relationships: Equivalence

- Any relationship that is both symmetric and transitive is an EQUIVALENCE relationship
- We often need to assert that a particular class or property has the same meaning as another class or property or that it is generally substitutable for it

Properties of Semantic Relationships: Inverse

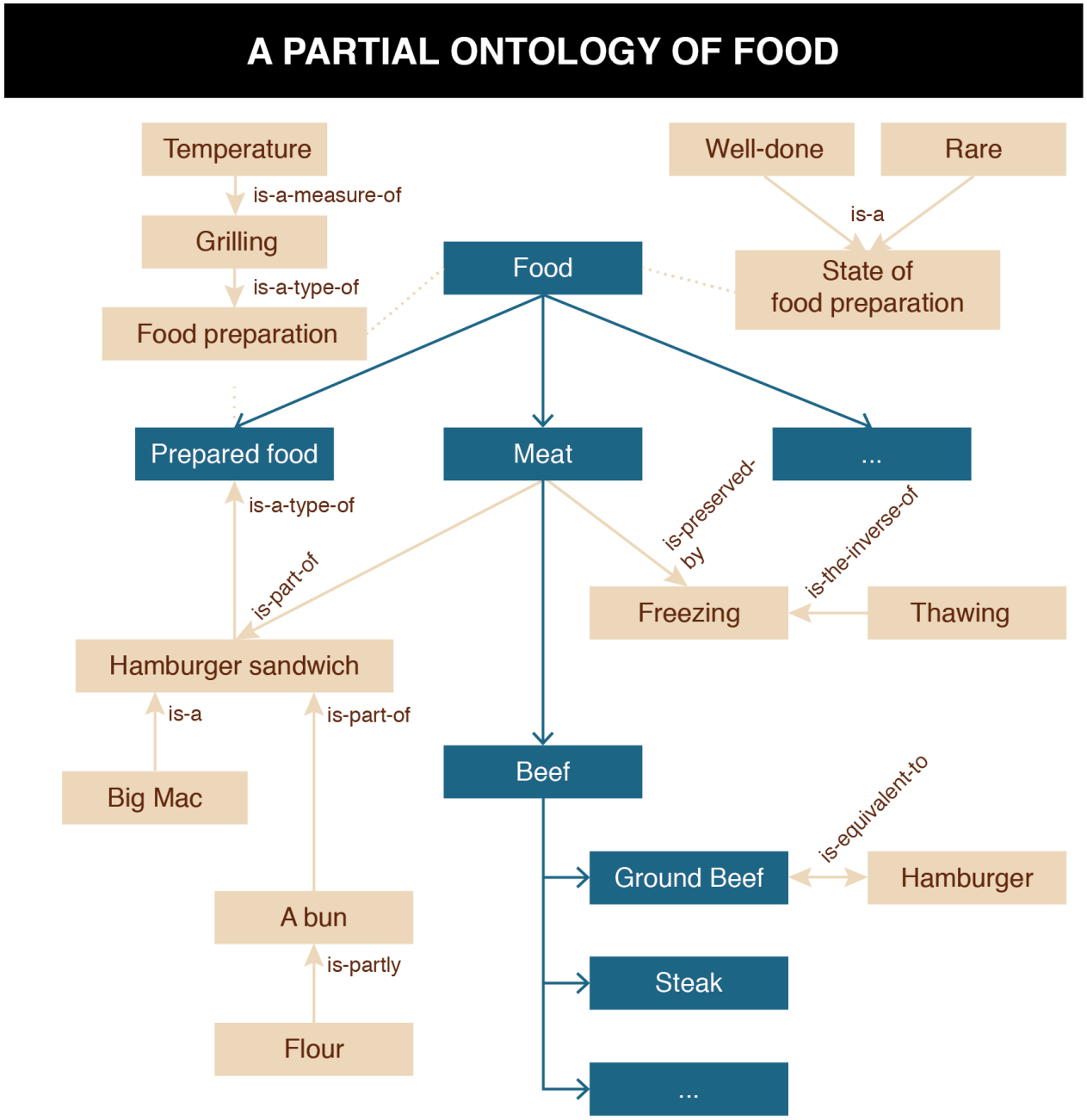
- For asymmetric relationships, it is often useful to be explicit about the meaning of the relationship when the order of the arguments in the relationship is reversed
- The resulting relationship is called the INVERSE or the converse of the first relationship

Ontology

- An ontology defines the concepts and terms used to describe and represent some domain or area of knowledge
- To fully understand a domain you need to know class and part-whole inclusion, attribution, equivalence, and many other relationships about and between the resources it contains
- Class inclusion relationships can form a kind of backbone or framework to which other kinds of relationships attach, creating a network of relationships

Analyzing the Definition of Ontology

- The word ontology has been used to describe artifacts with different degrees of structure that differ:
  - ... according to how precisely the terms are defined
  - ... according to how precisely the relationships among them are expressed
- So the simplest ontology is a dictionary
- A thesaurus is a somewhat more complex ontology
- More complete ontologies are expressed using formal logic-based language



TDO Figure 5.2

Ontological Assertions That Annotate the Food Taxonomy

Hamburger → is-equivalent-to → Ground Beef
Hamburger → is-prepared-by → Grilling
Grilling → is-a-type-of → Food Preparation
Hamburger Sandwich → is-a-type-of → Prepared Food
BigMac → is-a → Hamburger Sandwich
A bun → is-part-of → Hamburger Sandwich
A bun → is-partly → flour
Temperature → is-a-measure-of → Grilling
Rare → is-a → State of Food Preparation
Well-done → is-a → State of Food Preparation
Meat → is-preserved-by → Freezing
Thawing → is-the-inverse-of → Freezing

Ontologies and the "Vocabulary Problem"

- The Artificial Intelligence solution (for computers) to the vocabulary problem is to give an information system all the knowledge -- including "commonsense" -- that is needed to interpret every user's expressions in every context
- A great deal of work in AI has been dedicated to building knowledge bases to support language understanding, reasoning, problem solving applications
- The most famous / infamous effort is the Cyc project (http://www.cyc.com)

Cyc – Formalized Commonsense Knowledge

- Cyc knows a few hundred thousand basic concepts and a few million human-entered assertions about the world -- "facts, rules of thumb, and heuristics for reasoning about the objects and events of everyday life“
- The Cyc knowledge base consists of terms and assertions that relate those terms
- This kind of common sense is a pre-requisite for computers to achieve anything approaching human competence on natural language processing tasks (once you get outside of narrow, constrained domains)
- How will Cyc change the world in the next 10 years?

SNOMED

- Systematized Nomenclature of Medical-Clinical Terms, http://www.ihtsdo.org/snomed-ct/
- Important aid in accurately recording "healthcare encounters" to improve communication and interoperability among information systems, especially electronic health records
- Contains over 300,000 concepts and formal definitions, organized in several different hierarchies to enable terms to be located and interrelated from different perspectives

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

3 October 2013
Lecture 11.3 – Lexical Perspective on Relationships

The Lexical Perspective on Relationships

- A prototypical word is the minimal "meaning bearing" element of language
- Words express concepts and relationships, but not all concepts and relationships have words for them (are "lexicalized")
- These "lexical gaps" differ from language to language
- Whereas "conceptual gaps" -- the things we can't think of -- may be innate and universal

Linguistic Relativity in Iceland

Relationships Among Words / Relationships Among Concepts

- Representing the relationships among words is a task for linguistics and cognitive science
- In contrast, representing the relationships among concepts is a task for computer science (e.g., ontologies like Cyc)
- Understanding both kinds of relationships is essential for designers of organizing systems

Relations Among Words

- Hyponymy/Hyperonymy
- Metonymy
- Synonymy
- Polysemy
- Antonymy

Hyponymy/Hyperonymy

- When words encode IS-A or inclusion relationships, the word for the more specific class is the HYPONYM and the other is the HYPERNYM
- So there can be a "lexical hierarchy" that represents the "semantic hierarchy"
- Often used to situate "basic categories" with respect to superordinate and subordinate categories

Hyponymy/Hyperonymy

- A is a hyponym of B if A is a type of B
- A robin is a hyponym of bird
- A is a hypernym of B if B is a type of A
- An animal is a hypernym of bird

A Formula for Definitions

hyponym = {adjective+} hypernym {distinguishing clause+}

- Robin = Migratory BIRD with clear melodious song, a reddish breast, gray or black upper plumage
- Doesn't mention every characteristic of hyponym, only those needed to distinguish from other hyponyms

Metonymy

- The lexical representation of or analogy to part-whole or meronymic relationships
- Using one aspect of something to stand for the whole
  - "Bank" as building stands for the institution of the "bank"
  - "The White House released new budget figures today"

Synonymy

- Synonyms are different word forms that can express the same concept
  - cat, feline, Siamese cat
- Absolute synonyms that can always substitute for each other probably don't exist
- Propositional synonyms are those where substituting one for another isn’t likely to change the truth value of the statement

Polysemy

- Many "word forms" (particular spelling patterns) are polysemous with multiple senses established in the language -- they are semantically ambiguous

  - That dog has floppy ears
  - She has a good ear for jazz.

- "bank" (financial) has related senses:

  - a building (the bank on Shattuck)
  - specific financial firm (Wells Fargo)
  - where money is kept (abstract notion)
  - where anything of value is kept (more abstract)

Polysemy vs Homonymy

- A HOMOGRAPH is a word with multiple senses that are not conceptually related

  - bank (financial sense)
  - bank (river sense)

- But what looks like homography may be polysemy from a historical perspective

  - mole (an animal that digs holes; an infiltrating spy)

- HOMOPHONES are words with the same pronunciation but different spellings and meanings

  - to, two, too; might, mite; clawed, claude

Antonymy

- Antonyms are lexical opposites
- Some are inherently binary “true antonyms”
  - dead / alive, true / false, on / off
- Others are "graded"
  - long / short, hot / cold
- Markedness: if one member of a pair is more restricted in its contexts it can stand out psychologically
  - long is unmarked, short is marked

Thesauri

- A THESAURUS is a tool for finding the "right" or "good" terms of a controlled vocabulary
- It is a collection of vocabulary terms annotated with lexical relationships to indicate terms that are:
  - Preferred (UF "used for")
  - Broader (BT "broader term")
  - Narrower (NT "narrower term")
  - Related (RT "related term" or "see also")
  - USE in a thesaurus refers the reader from a variant term to a preferred term; the inverse of UF

Examples from Food Domain (TDO 5.4.2)

Food BT Meat
Beef NT Meat
Food UF Sustenance, Nourishment
Food RT Cooking, Dining, Cuisine

Victuals USE Food

WordNet

- Another "ontological resource" -- a "semantic dictionary" - started in 1985 by George Miller and others at Princeton's Cognitive Science program
  - over 100,000 nouns in 80,000 “synonym sets”
  - 11000 verbs in 13000 synsets
- Instead of using spelling as the primary organizing principle for words, it uses their semantic properties and relationships
- Designed as a network to capture the idea that words and concepts are an interrelated system

Using WordNet

- The WordNet database is freely downloadable from http://wordnet.princeton.edu/
- http://www.synonym.com/ is easier to work with
- The Visual Thesaurus is a commercial product that uses WordNet (http://www.visualthesaurus.com)
- WordNet is very commonly used in natural language processing research and applications
- Most important automated use of WordNet is in sense disambiguation, identifying a sense based on contexts in which it occurs

Implications for Vocabulary Design

- Choosing vocabulary terms, and precisely defining their semantics, is essential but impossible to do perfectly
- Your vocabulary must express what YOU intend, so you "look inward" -- analyze how you think about a domain
- You want others to understand what you mean, so you need to "look outward" -- analyze the terms used by your users, competitors, or subject matter experts
- You should reuse other vocabularies or thesauri if they exist, especially for any "horizontal" components, to improve transformability and interoperability
- But these three approaches may suggest different terms

Readings for Next Lecture

- rest of TDO 5
- TDO 8 through 8.2.1.5
- Glushko, Robert J. – "Modeling Methods and Artifacts for Crossing the Data/Document Divide", people.ischool.berkeley.edu/~glushko/glushko_files/GlushkoXML2005.pdf
- www.oasis-open.org/org
- Getting started with schema.org schema.org/docs/gs.html