Lexical Semantics

- So far, we have listed **words** in our **lexicon** or **vocabulary** assuming a single meaning per word.
- However, the same word *bank* means two different things but we cannot distinguish between them using the traditional definition of word.
- To deal with this issue, we combine the **spelling** or **pronunciation** of a word and the **meaning**. In the **lexicon** we now store **lexemes** instead of words. A lexeme pairs a particular spelling or pronunciation with a particular meaning.
The meaning part of a lexeme is called a **sense**. For CL, our interest is in relations between lexemes or disambiguating different senses of a word. 

word: bank → lexeme: **bank**\(^1\)  OR  word: bank → lexeme: **bank**\(^2\)

Note that meanings are often not definitions, but often are simple listings of compatible lexemes.

cf. dictionary defns: **red**, *n.* the color of blood or ruby; **blood**, *n.* red liquid circulating in animals
Homonyms

- **Homonyms: words that have the same form but different meanings**
  1. *Instead, the chemical plant was found in violation of several environmental laws*
  2. *Stanley formed an expedition to find a rare plant found along the Amazon river*

- **Same orthographic form: plant but two senses:** plant$^1$ and plant$^2$
Homonyms

- Text vs. speech: fly-casting for *bass* vs. rhythmic *bass* chords
  These cases are homonyms in text, but not in speech. Referred to as **homographs**

- Speech vs. text: *would* vs. *wood*
  These cases are not homonyms in text, but easily confused in speech. Referred to as **homophones**

- Note that this problem in some cases can be solved using *part of speech tagging*
  Can you think of a case which cannot be solved using POS tagging?
Applications

- Spelling correction: homophones: *weather* vs. *whether*
- Speech recognition: homophones: *to*, *two*, *too*. Also homonyms (see $n$-gram e.g.)
- Text to speech: homographs: *bass* vs. *bass*
- Information retrieval: homonyms: *latex*
Polysemy

- Consider the homonym: *bank* → commercial **bank**¹ vs. river **bank**²
- Now consider
  1. A PCFG can be trained using derivation trees from a tree bank annotated by human experts
- Is this a new sense of *bank*?
Polysemy

- Senses can be derived from a particular lexeme. This process is known as **polysemy**
  In previous case we would say that the use of *bank* is a sense derived from commercial *bank*¹
- In some cases, splitting into different lexemes has other supporting evidence: *bank*¹ has a Romance (Italian) origin vs. *bank*² has a Germanic (Scandinavian) origin
  1. A PCFG can be trained using a bank of derivation trees called a tree-bank annotated by human experts
- How can we tell between homonyms and polysemous uses of a word?
Consider the case for a verb like serve

1. *Does United serve breakfast?*
2. *Does United serve Philadelphia?*
3. *Does United serve breakfast and dinner?*
4. *#Does United serve breakfast and Philadelphia?*
Word Sense Disambiguation

▶ Consider a noun like *bank*
  1. *How many senses does it have?*
  2. *How are these senses related?*
  3. *How can they be reliably distinguished?*

▶ For NLP software, among these three questions, typically at runtime we need to automatically find the answer to the last question: given a word in context, map it to the correct lexeme: *word-sense disambiguation*
Synonyms

- Synonyms: Different lexemes with the same meaning
  1. How big/large is that plane?
  2. Would I be flying on a big/large or small plane?
- Synonyms clash with polysemous meanings
  1. Seema is my big sister
  2. #Seema is my large sister
WordNet

- WordNet is an electronic database of word relationships, handcrafted from scratch by researchers at Princeton University (George Miller, Christine Fellbaum, et al.)
- WordNet contains 3 databases: for verbs, nouns and one for adjectives and adverbs

<table>
<thead>
<tr>
<th>Category</th>
<th>Unique Forms</th>
<th>Number of Senses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>94474</td>
<td>116317</td>
</tr>
<tr>
<td>Verb</td>
<td>10319</td>
<td>22066</td>
</tr>
<tr>
<td>Adjective</td>
<td>20170</td>
<td>29881</td>
</tr>
<tr>
<td>Adverb</td>
<td>4546</td>
<td>5677</td>
</tr>
</tbody>
</table>
Ask the question: how many senses per noun or verb? The distribution of senses follows Zipf’s (2nd) Law.

- WordNet provides multiple lexeme entries for each word and for each part of speech, e.g. *plant* as noun has 3 senses; *plant* as verb has 2 senses
- WordNet also provides *domain-independent* lexical relations such as IS-A, HasMember, MemberOf, ...
WordNet: noun relations

<table>
<thead>
<tr>
<th>Relation</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypernym</td>
<td>this is a kind of</td>
<td>breakfast → meal</td>
</tr>
<tr>
<td>Hyponym</td>
<td>this has a specific instance</td>
<td>meal → lunch</td>
</tr>
<tr>
<td>Has-Member</td>
<td>this has a member</td>
<td>faculty → professor</td>
</tr>
<tr>
<td>Member-Of</td>
<td>this is member of a group</td>
<td>copilot → crew</td>
</tr>
<tr>
<td>Has-Part</td>
<td>this has a part</td>
<td>table → leg</td>
</tr>
<tr>
<td>Part-Of</td>
<td>this is part of</td>
<td>course → meal</td>
</tr>
<tr>
<td>Antonym</td>
<td>this is an opposite of</td>
<td>leader → follower</td>
</tr>
</tbody>
</table>
### WordNet: verb relations

<table>
<thead>
<tr>
<th>Relation</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypernym</td>
<td>this event is a kind of</td>
<td>fly → travel</td>
</tr>
<tr>
<td>Tropynym</td>
<td>this event has a subtype</td>
<td>walk → stroll</td>
</tr>
<tr>
<td>Entails</td>
<td>this event entails</td>
<td>snore → sleep</td>
</tr>
<tr>
<td>Antonym</td>
<td>this event is opposite of</td>
<td>increase → decrease</td>
</tr>
</tbody>
</table>
Sense 1: Canada
  ⇒ North American country, North American nation
    ⇒ country, state, land
      ⇒ administrative district, administrative division, territorial division
        ⇒ district, territory
          ⇒ region
            ⇒ location
              ⇒ entity, physical thing
Sense 3: Vancouver
⇒ city, metropolis, urban center
  ⇒ municipality
  ⇒ urban area
  ⇒ geographical area
  ⇒ region
  ⇒ location
  ⇒ entity, physical thing
⇒ administrative district, territorial division
  ⇒ district, territory
  ⇒ region
  ⇒ location
  ⇒ entity, physical thing
⇒ port
  ⇒ geographic point
  ⇒ point
  ⇒ location
  ⇒ entity, physical thing
A **synset** in WordNet is a list of synonyms (interchangeable words)

{ chump, fish, fool, gull, mark, patsy, fall guy, sucker, schlemiel, shlemiel, soft touch, mug }

How can we use this information like synsets, hypernyms, etc. from WordNet to benefit NLP applications?

Consider one example: PP attachment in “parsing”, words plus word classes extracted from the hypernym hierarchy increase accuracy from 84% to 88% (Stetina and Nagao, 1998)
Another example of WordNet used in NLP applications: **selectional restrictions**

We have considered subcategorization:

\[ VP\text{-}with\text{-}NP\text{-}complement \rightarrow V(eat)\ NP \] “eat six bowls of rice ”

But not selectional restrictions of the verb itself: “eat tomorrow ”

Consider what do you want to eat tomorrow

We can use the **synset** \{ food, nutrient \} to describe the NP argument of eat – then the 60K lexemes under these nodes in the WordNet hierarchy will be acceptable. (however, what about “eat my shorts ”)

→ several other applications have been explored