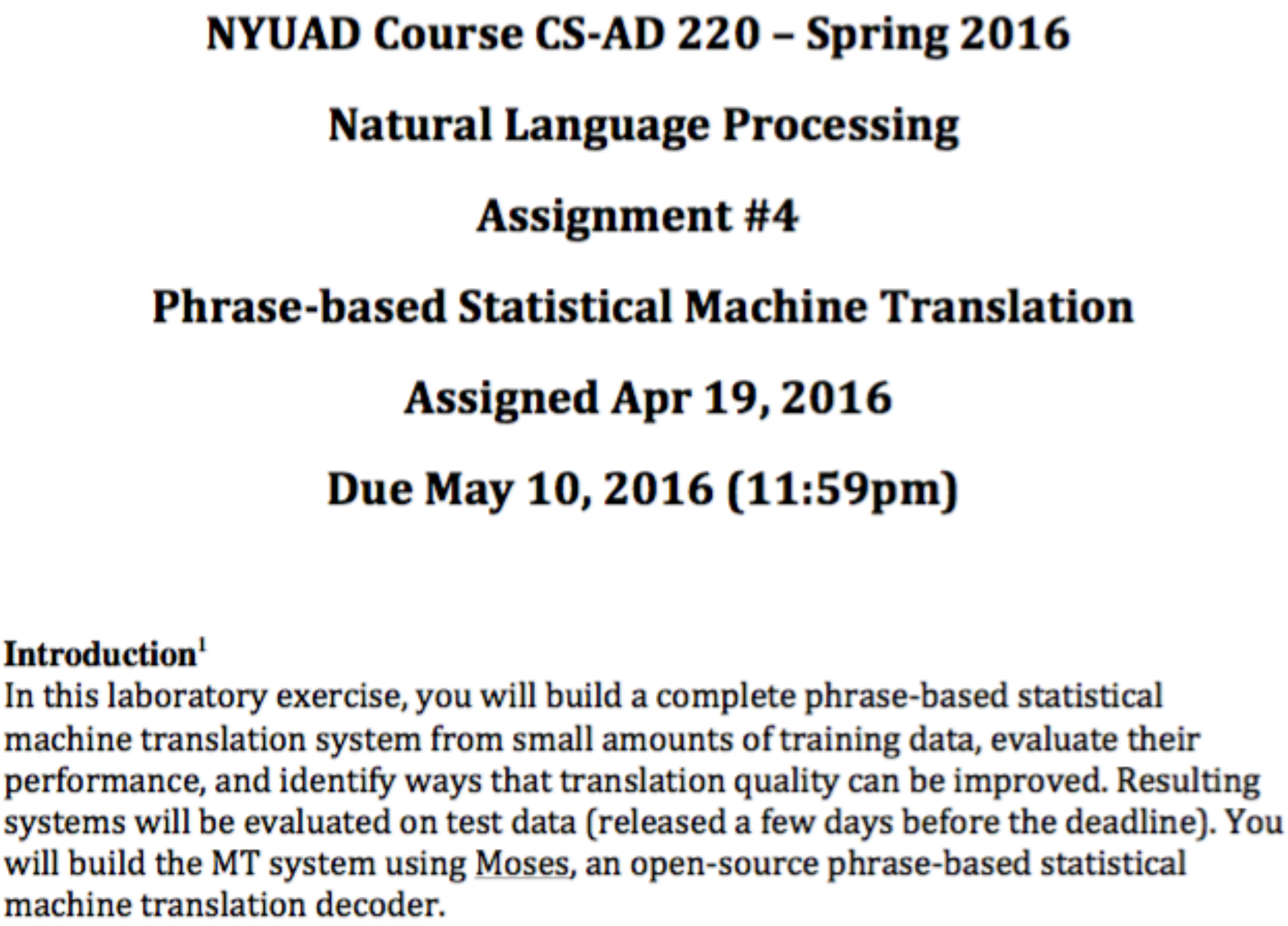
CS-AD 220 – Spring 2016

Natural Language Processing

Session 25: 28-Apr-16

Prof. Nizar Habash

Slides from Jurafsky and Martin with modifications



NYUAD Course CS-AD 220 – Spring 2016

Natural Language Processing

Assignment #4

Phrase-based Statistical Machine Translation

Assigned Apr 19, 2016

Due May 10, 2016 (11:59pm)

Introduction[1]

In this laboratory exercise, you will build a complete phrase-based statistical machine translation system from small amounts of training data, evaluate their performance, and identify ways that translation quality can be improved. Resulting systems will be evaluated on test data (released a few days before the deadline). You will build the MT system using Moses, an open-source phrase-based statistical machine translation decoder.

Assignment #4 posted on NYU Classes

START EARLY!

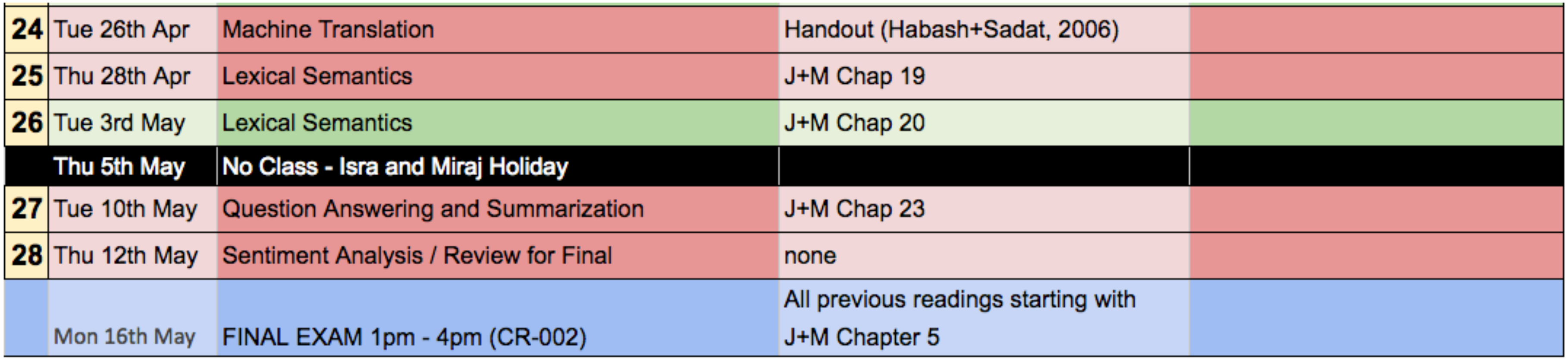
DEADLINE IS May 10 (11:59pm)

MT Assignment Update

- For Train and LM, please use the data in the directory unescaped.
- For Test and Tune data have not been escaped.
- The file Blind_Test_data.src.ar is the blind test you need to use for the Open Challenge.

Update to Syllabus

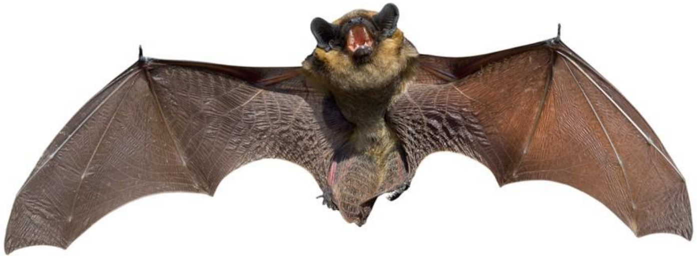
- May 10 and 12 changed
- May 12 will include review for final
- **Final is May 16, 1pm to 4pm in CR-002**

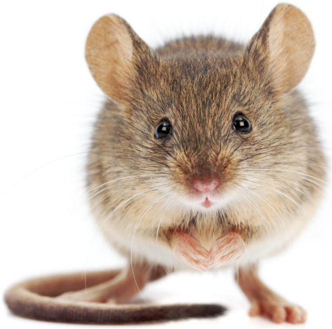


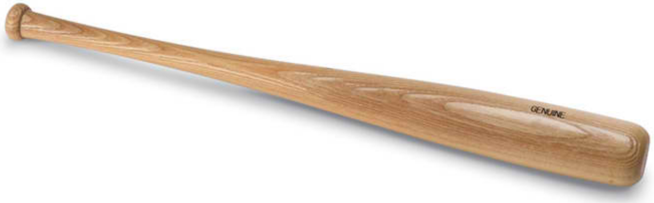
24	Tue 26th Apr	Machine Translation	Handout (Habash+Sadat, 2006)	
25	Thu 28th Apr	Lexical Semantics	J+M Chap 19	
26	Tue 3rd May	Lexical Semantics	J+M Chap 20	
	Thu 5th May	No Class - Isra and Miraj Holiday		
27	Tue 10th May	Question Answering and Summarization	J+M Chap 23	
28	Thu 12th May	Sentiment Analysis / Review for Final	none	
	Mon 16th May	FINAL EXAM 1pm - 4pm (CR-002)	All previous readings starting with J+M Chapter 5	

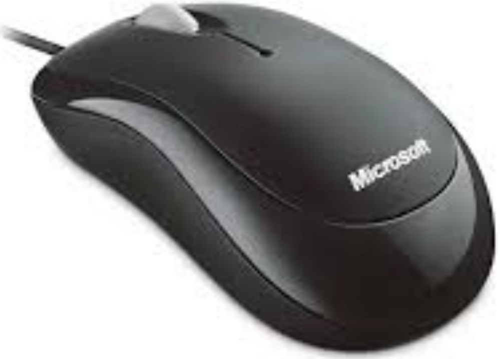
Lexical Semantics

- how similar is a bat to a mouse?









Reminder: 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
duermes	dormir
wakatabnAhA وكتبناها	katab كتب

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

Homonyms are both homographs and homophones

1. Homographs
   a. Bass and bass
   b. Lead (v) and lead (n)
2. Homophones:
   a. Write and right
   b. 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`)

Tree (`Plums have beautiful blossoms`)

↔ Fruit (`I ate a preserved plum`)

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"**

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

- 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

- 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 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...`

Applications of Thesauri and Ontologies

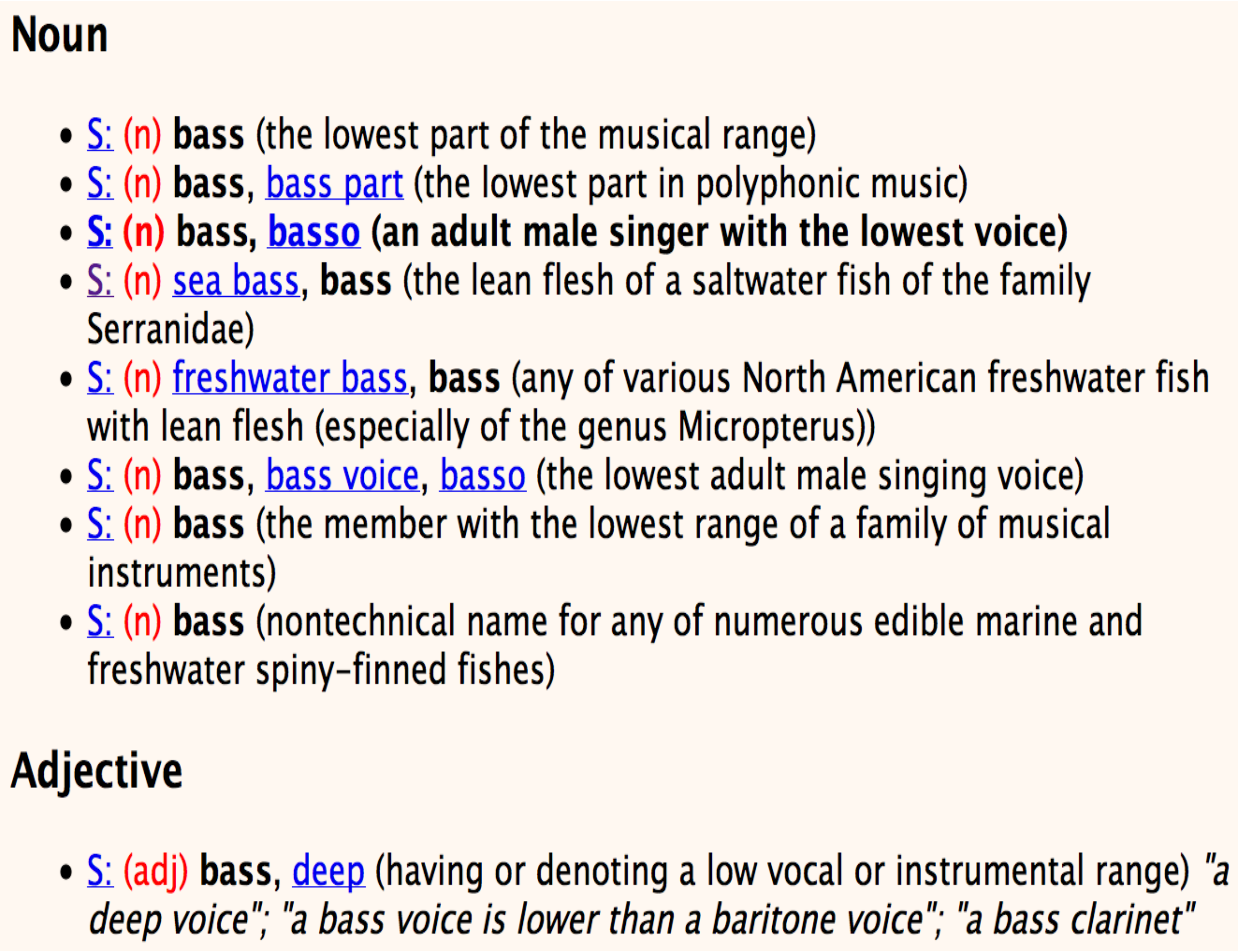
- Information Extraction
- Information Retrieval
- Question Answering
- Bioinformatics and Medical Informatics
- Machine Translation

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

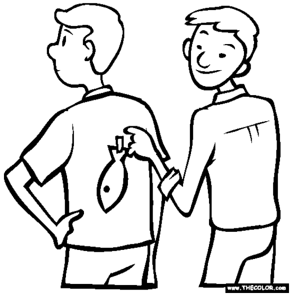
- S: (adj) **bass**, deep (having or denoting a low vocal or instrumental range) *"a deep voice"; "a bass voice is lower than a baritone voice"; "a bass clarinet"*

How is "sense" defined in WordNet?

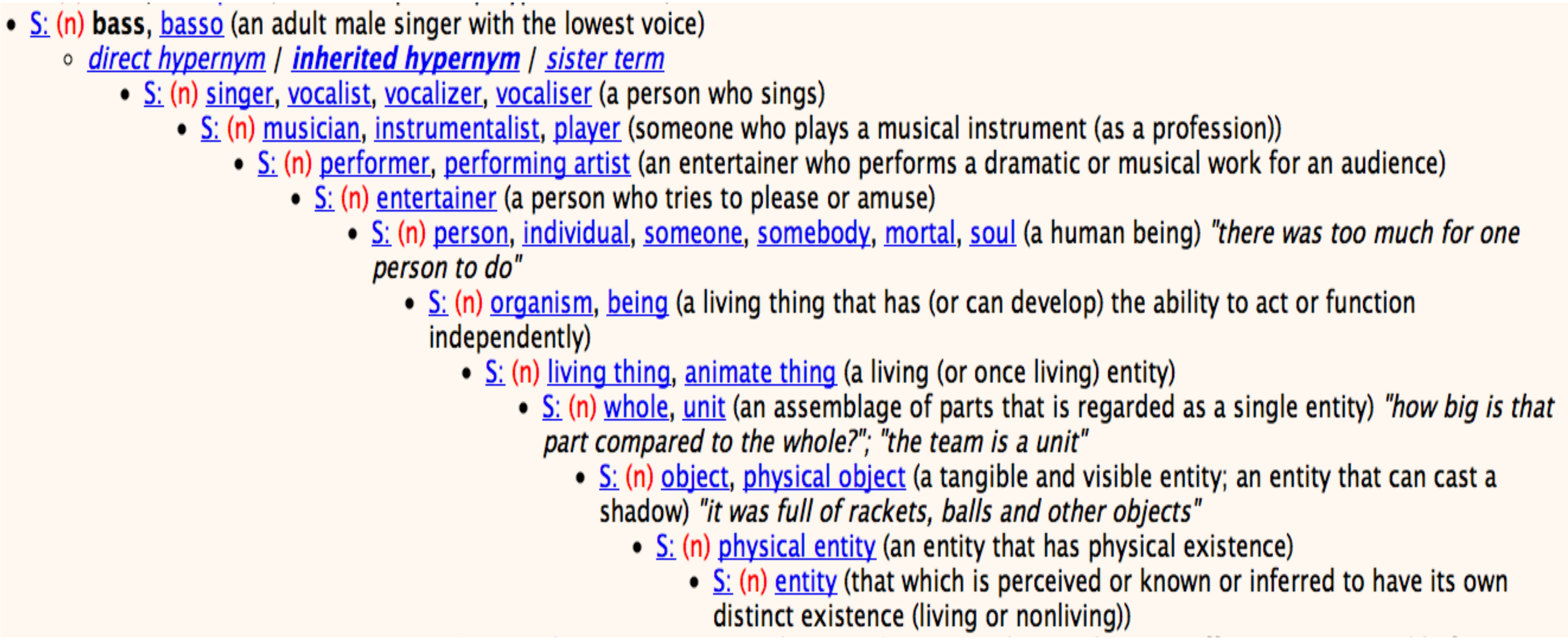
- **The synset (synonym set),** the set of near-synonyms, instantiates a sense or concept, with a gloss
- Example: chump as a noun with the gloss:

  "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 3.0

- Where it is:
  - http://wordnetweb.princeton.edu/perl/webwn
- Libraries
  - Python: WordNet from NLTK
    - http://www.nltk.org/Home
  - Java:
    - JWNL, extJWNL on sourceforge

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

- Information retrieval
- Question answering
- Machine translation
- Natural language generation
- Language modeling
- Automatic essay grading
- Plagiarism detection
- Document clustering

Next Time

- J+M Chap 20
- Come with questions about the MT assignment