WELCOME!
Bookkeeping Matters

- **Time:** Fridays 2:15-3:30 (or 11:00-12:30?)
- **Place:** Ventura 17
- **http://lingo.stanford.edu/courses/03/wl/**
- **Evaluation:** written report on some aspect of word learning
- **Please send us email if you are interested in taking the course:** {tbaldwin,dwiddows}@csli.stanford.edu
Class Format

• 1.5 hour informal lecture including lots of time for questions, comments, etc.

• (Hopefully) modular course structure

• Recommended readings for each week announced a week in advance on the web page

• Slides posted on the web page after class each week

• Web demos, etc. for hands-on familiarisation
Caveats

- Language orientation (English-centricity)
- Coverage of topics (instructor-centricity)
- Selection of readings (accessibility-centricity)
- Weird accents, spelling and examples (dialect-eccentricity)
BASIC INTRODUCTION
What is Computational Word Learning?

• Working definition: the process of discovering word commonalities, embellishing information contained in existing lexical resources and/or expanding existing lexical resources through computational means

• Computational word learning is also known as lexical acquisition
Brief History of Word Learning

- Word learning first came to prominence in the late 1980’s as electronic lexical resources (LRs) became available
- Early efforts focused on extracting “knowledge” out of LRs
- Shift in the late-1990’s towards corpus-based approaches
Recurring Themes in Word Learning

- Corpora
- **Supervised** vs. **unsupervised** methods
- Tokens and types
- Ambiguity and disambiguation
- Words and multiword expressions (MWEs)
Corpora

- A **corpus** (plural **corpora**) is a body of written or spoken language, generally either from a homogeneous source or balanced across multiple sources in an attempt to be representative of a given language type.

- Examples: British National Corpus (BNC), Penn Treebank (Brown, WSJ, Switchboard)
Supervision

• **Supervised** methods have prior knowledge of a closed set of word classes and set out to discover and categorise new words according to those classes.

• **Unsupervised** methods dynamically discover the word classes in the process of categorising the words.

• **Supervised or unsupervised?** James Cook, George de Mestral (inventor of Velcro), David Livingstone
Types and tokens

- The number of **types** in a corpus is the number of unique word forms, and the number of **tokens** is the total word count.

  > Pease porridge hot
  > Pease porridge cold
  > Pease porridge in the pot
  > Nine days old

- **Types**: 10 (Pease, porridge, hot, cold, ...)

- **Tokens**: 14 (Pease, porridge, hot, Pease, ...)
Ambiguity and Disambiguation

- **Ambiguity:** observation that a given word occurs in multiple configurations

- **Disambiguation:** determination of which of a fixed set of classes a given word conforms to

The gang held up the bank
The boat pulled up at the bank
We stopped by the bank
Words and Multiword Expressions

• **(Escapist definition)** A word is what we would expect to occur as an atomic, independent entry in a dictionary (e.g. reconsider, *shōgakkō* “primary school”)

• **(Narrow definition)** A multiword expression (MWE) is made up of multiple words and is syntactically and/or semantically idiosyncratic (e.g. look up, phone book, off screen)
Basic Topics

Word Discovery

Morphology

Subcategorisation Frames

Selectional Preferences

Diathesis Alternations

Lexico-syntactic clues to similarity

Semantic Compositionality

Compound Nominals

Conceptual Ambiguity & Disambiguation

Noun Countability

Vector Space Methods

26 September, 2003
Word and MWE Discovery

- Segmentation ("word splitting") of non-segmenting languages (e.g. *spot the breaks in this string*)
- Extraction of collocations and multiword expressions (MWEs) (e.g. *pick out the MWE*)
- Morpheme discovery (e.g. *antidisestablishmentarianism*)
Morphology

• When are words with the same stem also related semantically?
Subcat Frame Acquisition

• What complements does a given word (verb/noun/adjective) take?

  \[
  \text{give } [Kim]_{NP} \quad [\text{a present}]_{NP} \\
  \text{refer } [\text{to the article}]_{PP} \\
  \text{fondness } [\text{for chocolate}]_{PP}
  \]

• How to distinguish between arguments and modifiers computationally
Selectional Preferences

• What set of things can normally fly?
  ✴ Eg. birds, aircraft, missiles

• How do we learn sets like this from corpora?
Diathesis Alternations

- **Diathesis alternation**: systematic valence-level correspondence between subcat frames of a given verb (or verb paradigm)

  *Kim opened the door* ↔ *The door opened*

- Corpus- and dictionary-driven methods for extracting verbs which participate in given alternations

- Diathesis alternations in verb clustering
Noun Countability

- Noun countability: lexical property that determines which uses a noun can occur in
  
  a/one dog, two/many/some dogs
  φ/much/some information

- Corpus-based methods for determining the countabilities of English nouns

- Cross-linguistic countability prediction: English ↔ Dutch
Lexico-syntactic patterns

- **Lexico-syntactic pattern**: a recurring pattern which can often signify a regular semantic relationship between some of its constituents

  *France and Spain* and *France it and other European countries*

- How reliable are such patterns?

- How can the relationships be checked / validated?
Vector Space Models

- **Vectors**: Lists of numbers

- How do we map words to lists of numbers so that ‘similar’ words have ‘similar’ numbers?

- What can we do with this? (apart from information retrieval!)
Semantic Compositionality

- The degree of semantic compositionality/idiomaticity varies across MWEs:
  - red army, red dwarf, red herring
  - chicken out, bow out, make out

- Different methods for estimating the relative compositionality of MWEs
Conceptual Ambiguity and Disambiguation

• Nearly all the words we might learn are potentially ambiguous

• Even when we don’t think of them as ambiguous

• How do we learn

  ★ which words are ambiguous
  ★ when a particular meaning is being used
  ★ when does it matter?
Compound Nominals

- Compound nominals are highly productive and semantically varied:
  
  diesel truck/tanker, phone book, apple juice seat, cloud bus

- How is it possible to constrain the range of interpretations and determine the default interpretation for a nominal compound in isolation and in context?
NEXT WEEK
Word and MWE Discovery

- Readings listed web site with links to PDF copies of the papers
  - total of 4 recommended papers + 1 extra reading
  - focus on nature of problem and the basic techniques employed in each paper
  - look out for online demo of morpheme discovery

- http://lingo.stanford.edu/courses/03/wl/