Words and the Lexicon

Lecture #2.5 (short)

September 4th
2012

What is a word?

- A sequence of characters demarcated by white space
- Not exactly
  - Spoken language
  - Chinese
  - It's → it is
  - Other issues in tokenization (e.g., New York, rock 'n' roll)

Types versus Tokens

- A text will contain various words
- Some of these words may occur more than once
- "Chapter 2 introduced the regular expression, showing for example how a single search string could help a web search engine find both *woodchuck* and *woodchucks*."
- 25 words (tokens)

Types versus Tokens

- A text will contain various words
- Some of these words may occur more than once
- "Chapter 2 introduced the regular expression, showing for example how a single search string could help a web search engine find both *woodchuck* and *woodchucks*."
- 25 words (tokens)
- Some words (types) occur more than once (each a different token)
  - a – 2
  - search – 2

Types versus Tokens

- Word Token: an occurrence of a word at a particular spatio-temporal location (e.g., a sequential position in a text, an utterance event at a time and space).
- Word Type: a more abstract notion also termed lexeme – we speak of two tokens belonging to the same type.
- Also, *woodchuck* and *woodchucks* *are two grammatical forms of the same lexeme (woodchuck).*

5

Lexical Knowledge

- Phonology: sounds rhythm, variants, homophones
- Semantics: meanings of (parts of) parts of words, synonyms
- Morphology: related word forms (e.g., plural)
- Syntax: how to use the word in a sentence
- Pragmatics: appropriate situations for using the word
- Orthography: how the word is written variants
- Etymology: history of the word, obsolete meanings

6

Parts of Speech/Word Classes

Open Class Word Categories

- Nouns: person, place, or thing; proper vs. common, mass vs. count, number, gender, case
- Verbs: most referring to actions and processes; main verbs vs. auxiliaries; transitive (hit, keep) vs. intransitive (arrive, snore)
- Adjectives: terms that describe properties or qualities
- Adverbs: modify something; directional, locative, degree, manner, temporal

7

Parts of Speech/Word Classes

Closed Class Word Categories

- Determiners: definite (the), indefinite (a), demonstrative (this)
- Prepositions: occur before a noun phrase, semantically they are relational
- Conjunctions: coordinating (and), subordinating (if, that)
- Auxiliary verbs: can, may, should, are, have
- Pronouns: personal (she), possessive (her), interrogative (who), relative (who), reflexive (himself)
- Particles: combine with a verb to form a phrasal verb; up, down, on, off, in, out
- Numerals: one, two, three, first, second, third

8

Tests for Word Classes

Morphological (formal) tests:
- Look for closely related forms;
- E.g., only nouns can bear the plural affix

Syntactic tests:
- What words co-occur (i.e., immediately precede and follow) with the word?
- E.g., can you say the word directly after *the*?

Semantic (notational) tests:
- What kinds of things does the word denote?
- E.g., person, place, or thing?

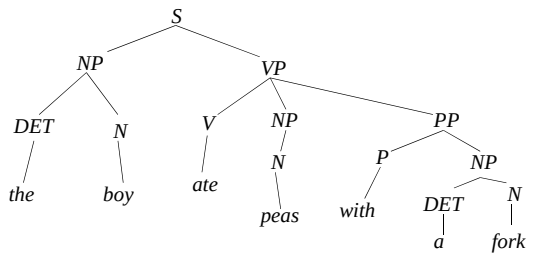
9

Major Syntactic Constituents

- Noun Phrase (NP): referring expressions (the blue shoe)
- Verb Phrase (VP): verbs plus complements (marks the sixth consecutive monthly decline)
- Prepositional Phrase (PP): direction, location, time, manner, etc. (in three minutes)
- Adjectival Phrase (AdjP): modified or complemented adjectives (much sharper, content to stay)
- Complementizers (COMP): (that, whether)

10

Constituent Structure (parse tree)



11

Lexicon

- Used in NLP systems to associate information with words (either for parsing or generation)
- Information about a word is called a lexical entry
- In parsing, each word in the input is scanned, and then lexical lookup retrieves one or more entries from the lexicon. Some of the information may be dynamically computed (e.g., by exploiting various lexical regularities).
- NLP-specific lexicons are similar to, but typically richer than, a printed dictionary
- Some NLP systems have used machine-readable versions of printed dictionaries
- Good source of links: http://www.clres.com/siglex.html

12

NLP Lexicon: abandon

- http://cs.nyu.edu/cs/faculty/grishman/comlex.html
- Lexical entry with syntactic frame:

  (verb :orth "abandon"
  :subc ((np-pp :p-val ("to")) (np)))

  – Words that take complements will have a subcategorization (:subc) feature. For example, the verb "abandon" can occur with a noun phrase followed by a prepositional phrase with the preposition "to" (e.g., "I abandoned him to the sea.") or with just a noun phrase complement ("I abandoned the ship").

13