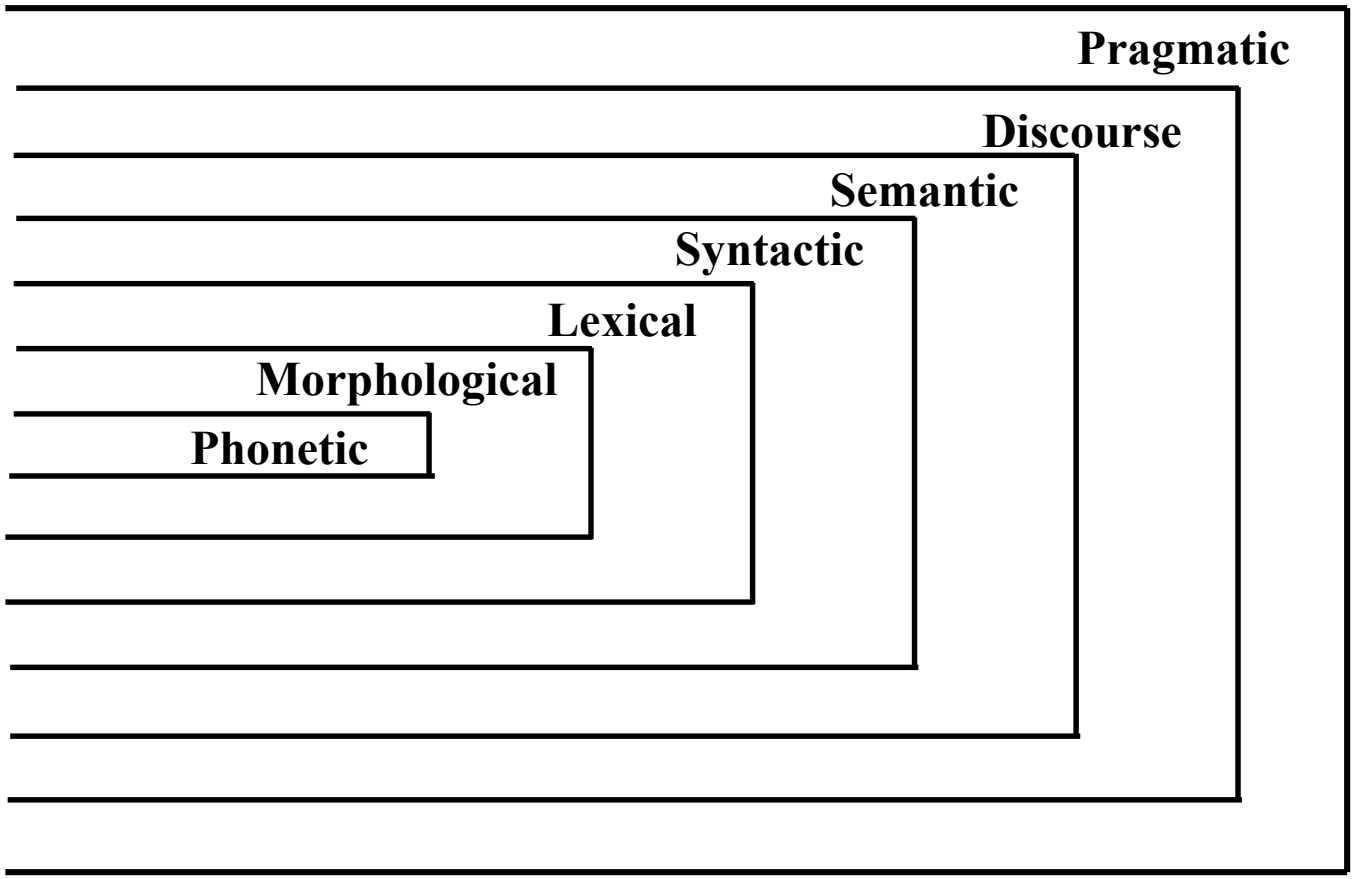
LEVELS OF LANGUAGE USED BY NATURAL LANGUAGE PROCESSING

adopted some materials developed in previous courses by Nancy McCracken, Liz Liddy and others; and some instructor resources for the book “Speech and Language Processing” by Daniel Jurafsky and James H. Martin

Levels Of Language Analysis

- Use the synchronic model to guide computational techniques to analyze text (as much as possible)



Synchronic Model Of Language

- The more exterior the level of language processing:
  - The larger the unit of analysis
    - phoneme-> morpheme -> word -> sentence -> text -> world
    - The less precise the language phenomena
  - The more free choice & variability
    - less rule-oriented, more exceptions to regularities
  - The more levels it presumes a knowledge of or reliance on
  - Theories used to explain the data move more into the areas of cognitive psychology and AI

- Lower levels of the model have been more thoroughly investigated and incorporated into NLP systems

The "Non-level" NLP Analysis

- Corpus Statistics
  - Frequencies of words
  - Frequencies of word pairs, using co-occurrence or semantic measures

- Classification or other Machine Learning
  - Use NLP to produce features, also known as attributes, of the text
  - Classify the text according to a set of labels
    - Classify customer reviews as positive or negative
    - Classify news articles according to topic

CORPUS LINGUISTICS USING WORD FREQUENCIES

WHAT IS CORPUS LINGUISTICS?

- A methodology to process text and provide information about the text
- The Corpus is a collection of text
  - Utilizes a representative sample of machine-readable text of a language or a particular variety of text or language
  - Many contain linguistic annotations, such as POS tags, named entities, syntactic structures, semantic roles, etc.
- Statistical analysis
  - Word frequencies
  - Collocations
  - Concordances
- Often used in "Digital Humanities" as ways to characterize properties of corpora
  - Where the "properties" of interest may govern choices of words to highlight

Text Corpus Structure

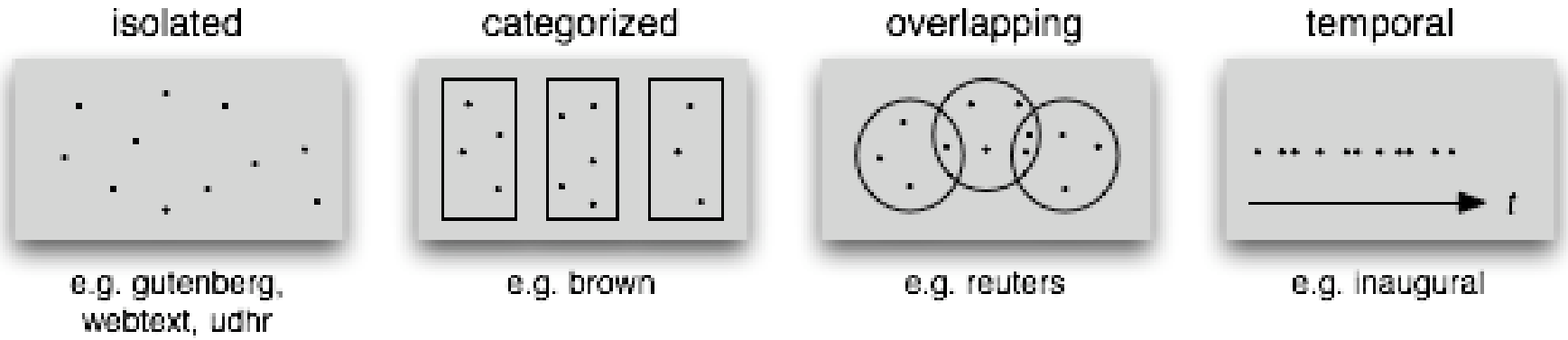


Image from: http://www.nltk.org/book/ch02.html

Preliminary Text Processing Required :

- Define the words so that you can count them:
  - Filter out 'junk data'
    - Formatting / extraneous material
    - First be sure it doesn't reveal important information
  - Deal with upper / lower case issues
    - Ignore capitalization at beginning of sentence? Is "They" the same word as "they"?
    - Ignore other capitalization? In a name such as "Unilever Corporation" is "Corporation" the same word as "corporation"

Preliminary Text Processing Required (Cont' d):

- Tokenization (or word segmentation):
  - Decide how to separate the characters in the sentence into individual words
    - Words are separated by "white space" or by special characters in English
    - No white space in Japanese language
    - In some languages, there are complex compound words – *"Lebensversicherungsgesellschaftsangestellter"*
  - Requires decisions on how to recognize and deal with punctuation
    - Apostrophes (one word *it's* vs. two words *it 's*
    - Hyphens (*snow-laden* vs. *New York-New Jersey*)
    - Periods (kept with abbreviations vs. separated as sentence markers)

Preliminary Processing Required: (Cont'd)

- Morphology (To stem or not to stem?)
  - Depends on the application
  - With stemming
    - "cat" is the same word as "cats"
    - "computing" is the same word as "compute"

- Additional issues if OCR'd data or speech transcripts in order to correct transcription errors

OCR (Optical Character Recognition): the recognition of printed or written text characters by a computer

Word Counting In Corpora

- Terminology for word occurrences:
  - Tokens – the total number of words
  - Distinct Tokens (sometimes called word types) – the number of distinct words, not counting repetitions – sometimes called vocabulary

The following sentence from the Brown corpus has 16 tokens and 14 distinct tokens:

They picnicked by the pool, then lay back on the grass and looked at the stars.

Note: we did not consider punctuations here. In NLTK, word_tokenize(text) function considers punctuations as well.

Word Frequencies

- Count the number of each token appearing in the corpus (or sometimes single document)
- A frequency distribution is a list of all tokens with their frequency, usually sorted in the order of decreasing frequency
- Used to make "word clouds"
  - For example, http://www.tumblr.com/tagged/word+cloud, http://stateoftheunion.onetwothree.net/#
- Used for comparison and characterization of text
  - See the State of the Union (SOTU) Speeches by Nate Silver http://fivethirtyeight.com/features/obamas-sotu-clintonian-in-good-way/
  - Methodology: choose topic words of interest and plot frequencies of these words vs. different speeches

How Many Words In A Corpus?

- Let N be the number of tokens
- Let V be the size of the vocabulary (the number of distinct tokens)

Church and Gale (1990): $|V| > O(N^{½})$

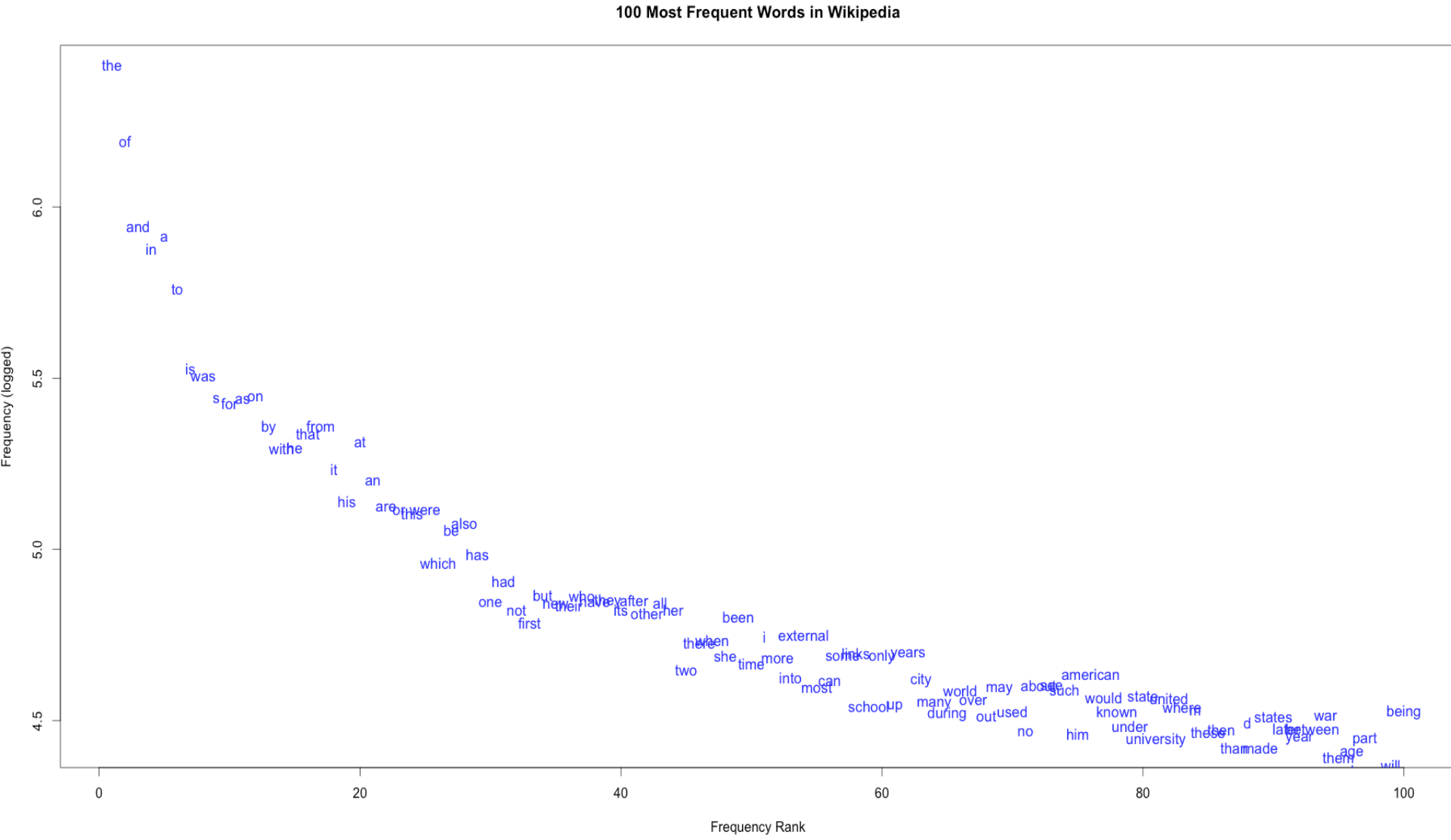
	Tokens = N	Types = |V|
Switchboard phone conversations	2.4 million	20 thousand
Shakespeare	884,000	31 thousand
Google N-grams	1 trillion	13 million

from Dan Jurafsky

Zipf's Law

- In a natural language corpus, the frequency of any word is inversely proportional to its rank in a frequency table
- Rank (*r*): The numerical position of a word in a list sorted by decreasing frequency (*f*).
- Zipf (1949) "discovered" that: $f \cdot r = k$ (for constant k)
  - Examples if k is 1:
    - Most frequent word (r = 1) is twice as frequent as 2nd most frequent
    - Most frequent (r = 1) is 3 times as frequent as 3rd most frequent, etc.

For example, in the Brown Corpus of American English text, the word "the" is the most frequently occurring word, and by itself accounts for nearly 7% of all word occurrences (69,971 out of slightly over 1 million). True to Zipf's Law, the second-place word "of" accounts for slightly over 3.5% of words (36,411 occurrences), followed by "and" (28,852). ----- from Wikipedia



a sample of 36.8 million words from Wikipedia, over 580,000 word types, nearly half (280,000) occur just once in the sample. --- image and this data from http://wugology.com/zipfs-law/

Zipf's Law Impact On Language Analysis

- Good News: Stopwords (commonly occurring words such as "the") will account for a large fraction of text so eliminating them greatly reduces the number of words in a text

- Bad News: For most words, gathering sufficient data for meaningful statistical analysis is difficult since they are extremely rare.