Corpus Linguistics

What is Corpus Linguistics?

- A methodology to process text and provide information about the text, usually at the one or two word level.
- 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
- Statistical analysis
  - Word frequencies
  - Collocations
  - Concordances
- Often used in “Digital Humanities” as ways to characterize properties of corpora

Preliminary Text Processing Required :

- Find the words:
  - Filter out ‘junk data’
    - Formatting / extraneous material
    - First be sure it doesn’t reveal important information
  - Deal with upper / lower case issues
  - Tokenize
    - Decide how you define a ‘word’
    - 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)

- Word segmentation
  - No white space in Japanese language
  - Compound words – *"Lebensversicherungsgesellschaftsangestellter"*
- Additional issues if OCR'd data or speech transcripts
- Morphology (To stem or not to stem?)
  - Depends on the application

Word Counting in Corpora

- After corpus preparation, additional decisions
  - Ignore capitalization at beginning of sentence? Is "They" the same word as "they"?
  - Ignore other capitalization? is "Company" the same word as "company"
  - Stemming? Is "cat" the same word as "cats"
- Terminology for word occurrences:
  - Tokens – the total number of words
  - Distinct Tokens (sometimes called word types) – the number of distinct words, not counting repetitions
    - 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.*

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
- Used for comparison and characterization of text
  - See the article on the State of the Union (SOTU)Speeches by Nate Silver

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

Also see xkcd.com/1133/
How to describe rocket only using words from most common 1,000

Zipf's Law

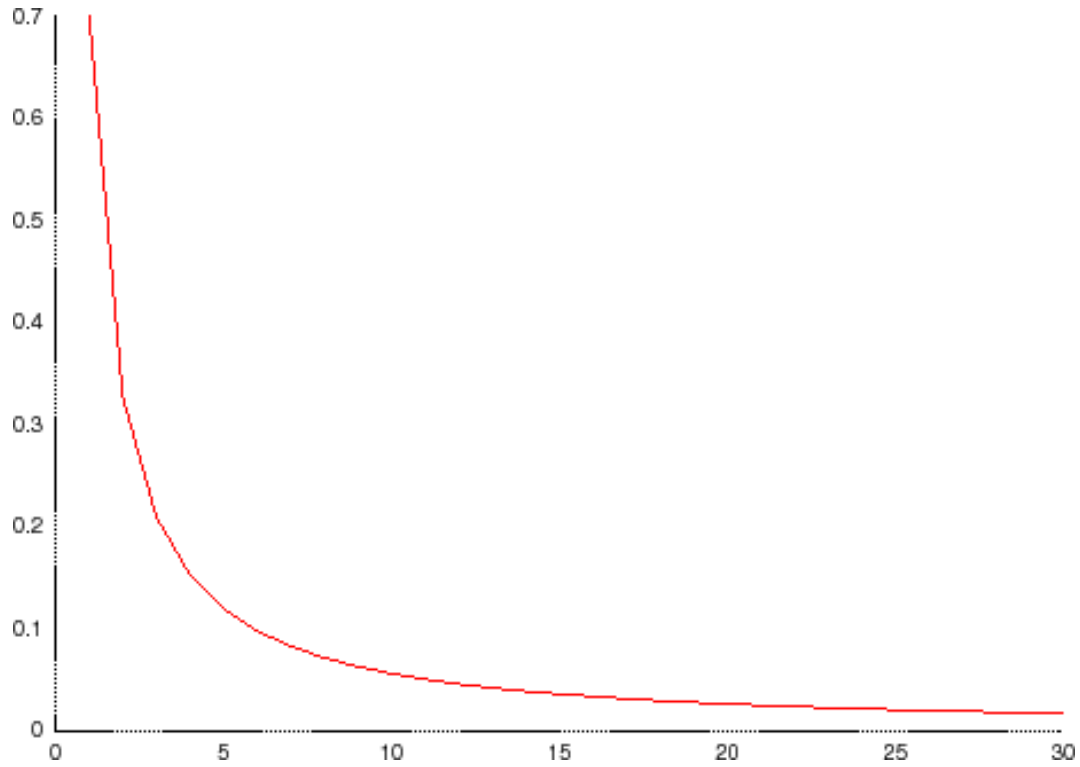
- Rank (r): The numerical position of a word in a list sorted by decreasing frequency (f).
- Zipf (1949) "discovered" that:

$$f \cdot r = k \quad (\text{for constant } k)$$

- If probability of word of rank r is p_r and N is the total number of word occurrences:

$$p_r = \frac{f}{N} = \frac{A}{r} \quad \text{for corpus indp. const. } A \approx 0.1$$

Zipf curve



A typical Zipf-law rank distribution. The y-axis represents word occurrence frequency, and the x-axis represents rank (highest at the left).

* Diagram from planetmath.org.

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 size of vocabulary in a text

- Bad News: For most words, gathering sufficient data for meaningful statistical analysis (e.g. for correlation analysis for query expansion) is difficult since they are extremely rare.