Corpus Statistics
using word and bigram frequencies, and bigram mutual information scores

What is Corpus Statistics/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
- Statistical analysis
  - Word frequencies
  - Collocations of words: bigrams, trigrams, etc.
- Often used in “Digital Humanities” as ways to characterize properties of corpora
  - Where the “properties” of interest may govern choices of words to highlight

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

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

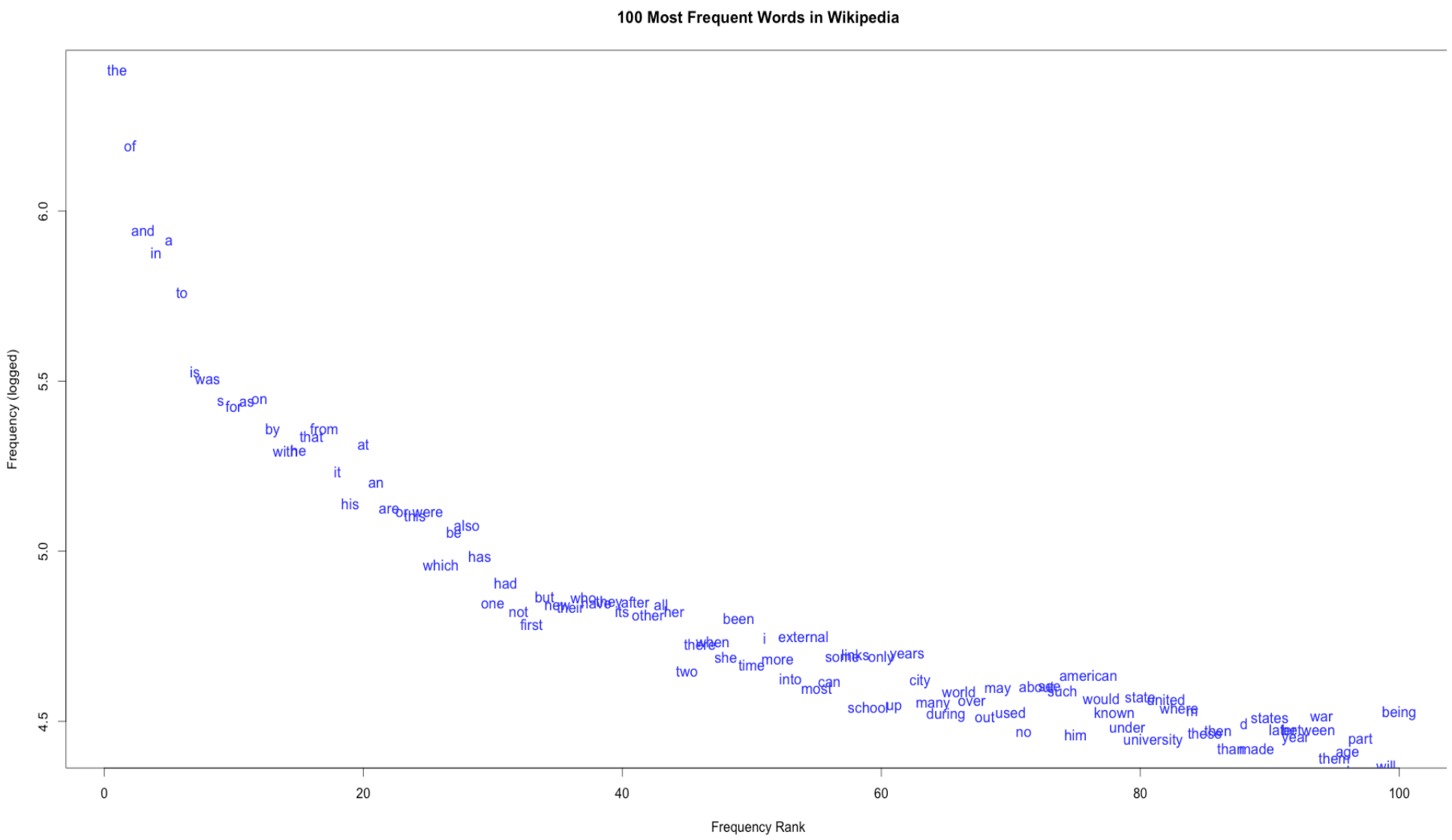
from Dan Jurafsky

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

Zipf's Law

- 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.

Bigrams

- Examples of bigrams are any two words that occur together
  - In the text: “two great and powerful groups of nations”, the bigrams are “two great”, “great and”, “and powerful”, etc.
- **The *frequency* of an n-gram** is the percentage of times the n-gram occurs in all the n-grams of the corpus and could be useful in corpus statistics
  - For bigram xy:
    - Count of bigram xy / Count of all bigrams in corpus
  - Examples are in the Google N-gram corpus

Google N-Gram Release

All Our N-gram are Belong to You

By Peter Norvig - 8/03/2006 11:26:00 AM

Posted by Alex Franz and Thorsten Brants, Google Machine Translation Team

Here at Google Research we have been using word n-gram models for a variety of R&D projects, such as statistical machine translation, speech recognition, spelling correction, entity detection, information extraction, and others. While such models have usually been estimated from training

to share this enormous dataset with everyone. We processed 1,024,908,267,229 words of running text and are publishing the counts for all 1,176,470,663 five-word sequences that appear at least 40 times. There are 13,588,391 unique words, after discarding words that appear less than 200 times.

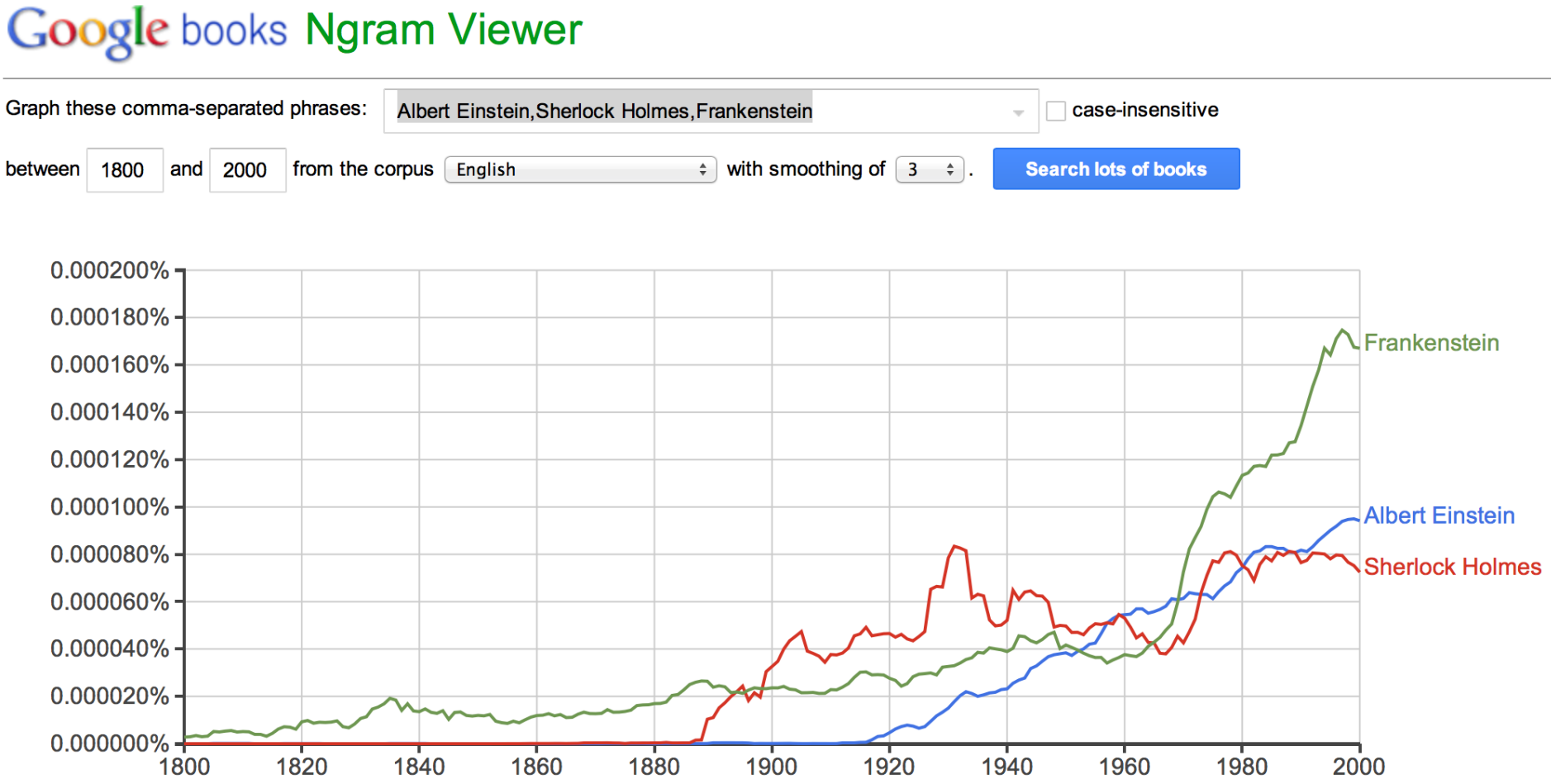
Example Data

- Examples of 4-gram frequencies from the Google N-gram release
  - `serve as the incoming 92`
  - `serve as the incubator 99`
  - `serve as the independent 794`
  - `serve as the index 223`
  - `serve as the indication 72`
  - `serve as the indicator 120`
  - `serve as the indicators 45`
  - `serve as the indispensable 111`
  - `serve as the indispensible 40`
  - `serve as the individual 234`

Google n-gram viewer

- In 2010, Google placed on on-line n-gram viewer that would display graphs of n-gram frequencies of one or more n-grams, based on a corpus defined from Google Books
  - https://books.google.com/ngrams
  - And see also the “About NGram Viewer” link

Google n-gram viewer



Additional corpus measures

- Recall that so far, we have looked at one measure involving bigrams (and these definitions can be extended to n-grams):
  - Bigram frequency – percentage occurrence of the bigram in the corpus
    - Seen in the Google N-gram data
- Other measures can be defined about the occurrences of bigrams in a corpus
  - Mutual information, …
  - More of these can be found in the NLTK

Corpus Statistics: Mutual Information (MI)

- N-Gram probabilities predict the next word – Mutual Information computes probability of two words occurring in sequence
- Given a pair of words, compares probability that the two occur together as a joint event to the probability they occur individually & that their co-occurrences are simply the result of chance
  - The more strongly connected 2 items are, the higher will be their MI value

Mutual Information

- Based on work of Church & Hanks (1990), generalizing MI from information theory to apply to words in sequence
  - They used terminology *Association Ratio*
- P(x) and P(y) are estimated by the number of observations of x and y in a corpus and normalized by N, the size of the corpus
- P(x,y) is the number of times that x is followed by y in a window of w words
- Mutual Information score (also sometimes called PMI, Pointwise Mutual Information):

$$\text{PMI}(x,y) = \log_2 \left(P(x,y) / P(x)\,P(y) \right)$$

MI values based on 145 WSJ articles

x	freq (x)	y	freq (y)	freq (x,y)	MI
Gaza	3	Strip	3	3	14.42
joint	8	venture	4	4	13.00
Chapter	3	11	14	3	12.20
credit	15	card	11	7	11.44
average	22	yield	7	5	11.06
appeals	4	court	47	4	10.45
…..					
said	444	it	346	76	5.02

Uses of Mutual Information

- Used in similar NLP applications as Language Models
  - Idiomatic phrases for MT
  - Sense disambiguation (both statistical and symbolic approaches)
  - Error detection & correction in speech analysis and spell-checking
- Used for distributional semantics in “deep learning”
- Used in non-text applications such as comparing features in machine learning

More on Corpus Statistics

- After collecting frequencies of words and bigrams or mutual information scores, use these to characterize some aspect of the text
  - Contents, topics
  - Style, informal vs. formal, differences in gender usage
- Example of word frequencies 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