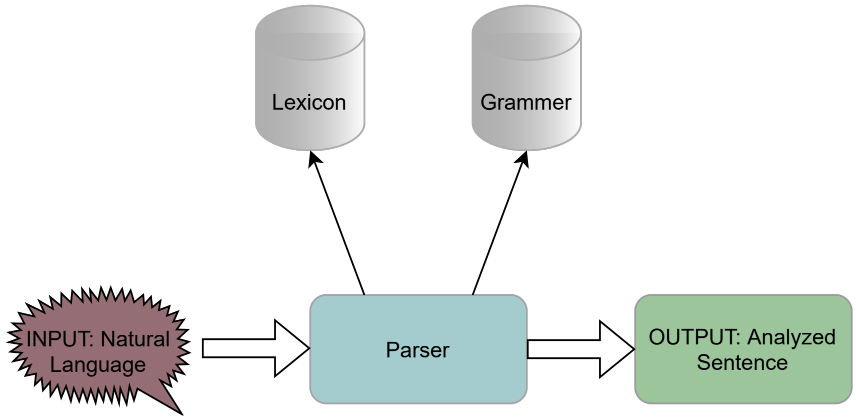
Chapter NLP:II

II. Corpus Linguistics

- Empirical Research
- Text Corpora
- Text Statistics
- Data Acquisition

Text Statistics

Classic processing model for language:



Statistical aspects of language

- The lexical entries are not used equally often
- The grammatical rules are not used equally often
- The expected value of certain word forms or word form combinations depends on the technical language used (Sub Language)

Text Statistics

Questions:

- How many words are there?
- How do we count?

  bank$^{(1)}$ (the financial institution),
  bank$^{(2)}$ (land along the side of a river or lake),
  banks$^{(1)}$, banks$^{(2)}$, ...

- How often does each word occur?

Text Statistics

Questions:

- How many words are there?
- How do we count?

  bank$^{(1)}$ (the financial institution),
  bank$^{(2)}$ (land along the side of a river or lake),
  banks$^{(1)}$, banks$^{(2)}$, ...

- How often does each word occur?

Experiment:

- Read a text left to right (beginning to end); make a tally of every new word seen.
- n words seen in total, $v(n)$ different words so far.
- How does the vocabulary V (set of distinct words) grow? $\rightarrow$ Plot $v(n)$.

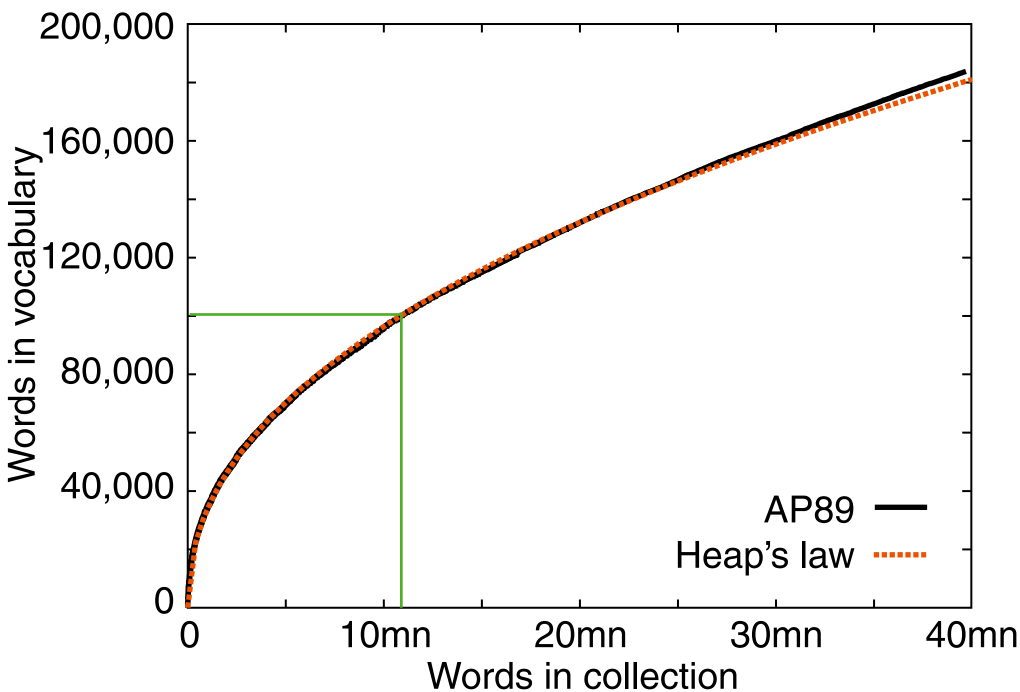
Text Statistics

Vocabulary Growth: Heaps' Law

The vocabulary V of a collection of documents grows with the collection.
Vocabulary growth can be modeled with Heaps' Law:

$$|V| = k \cdot n^{\beta},$$

where n is the number of non-unique words, and k and β are collection parameters.



- Corpus: AP89
- $k = 62.95$, $\beta = 0.455$
- At $10,879,522$ words: $100,151$ predicted, $100,024$ actual.
- At $< 1,000$ words: poor predictions

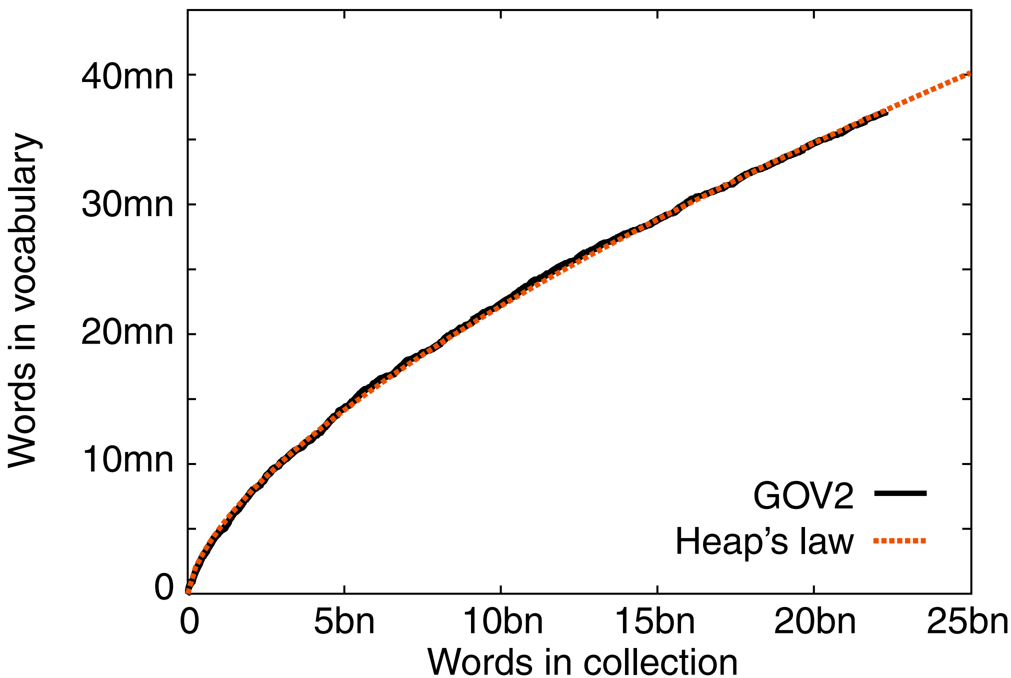
Text Statistics

Vocabulary Growth: Heaps' Law

The vocabulary V of a collection of documents grows with the collection.
Vocabulary growth can be modeled with Heaps' Law:

$$|V| = k \cdot n^{\beta},$$

where n is the number of non-unique words, and k and β are collection parameters.

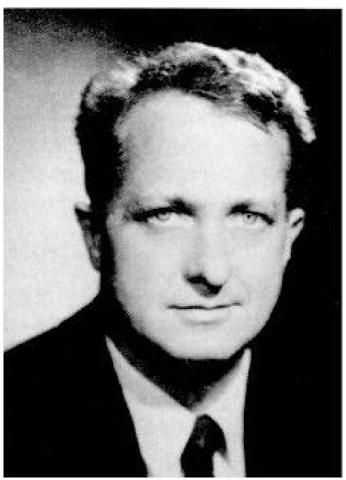


- Corpus: GOV2
- $k = 7.34$, $\beta = 0.648$
- Vocabulary continuously grows in large collections
- New words include spelling errors, invented words, code, other languages, email addresses, etc.

Text Statistics

Term Frequency: Zipf's Law

- The distribution of word frequencies is very *skewed*: Few words occur very frequently, many words hardly ever.
- For example, the two most common English words (the, of) make up about 10% of all word occurrences in text documents. In large text samples, about 50% of the unique words occur only once.



George Kingsley Zipf, an American linguist, was among the first to study the underlying statistical relationship between the frequency of a word and its rank in terms of its frequency, formulating what is known today as Zipf's law.

For natural language, the "Principle of Least Effort" applies.

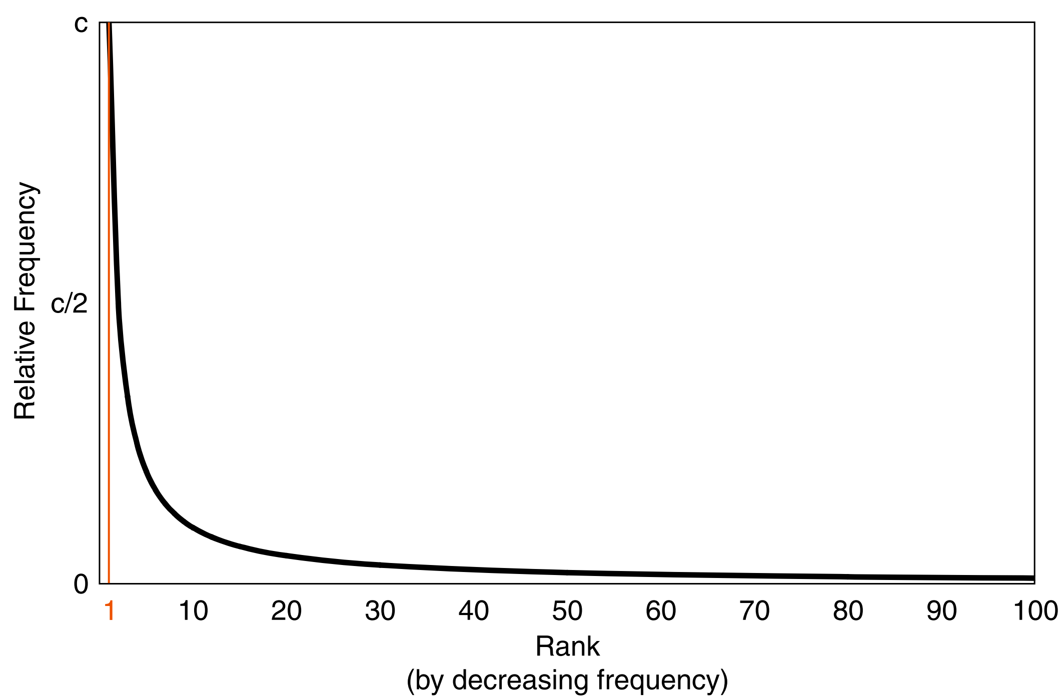
Text Statistics

Term Frequency: Zipf's Law (continued)

The relative frequency $P(w)$ of a word w in a sufficiently large text (collection) inversely correlates with its frequency rank $r(w)$ in a power law:

$$P(w) = \frac{c}{(r(w))^a} \qquad \Leftrightarrow \qquad P(w) \cdot r(w)^a = c,$$

where c is a constant and the exponent a is language-dependent; often $a \approx 1$.



Text Statistics

Term Frequency: Zipf's Law (continued)

Example: Top 50 most frequent words from AP89. Have a guess at c?

r	w	frequency	$P \cdot 100$	$P \cdot r$
1	the	2,420,778	6.09	0.061
2	of	1,045,733	2.63	0.053
3	to	968,882	2.44	0.073
4	a	892,429	2.25	0.090
5	and	865,644	2.18	0.109
6	in	847,825	2.13	0.128
7	said	504,593	1.27	0.089
8	for	363,865	0.92	0.073
9	that	347,072	0.87	0.079
10	was	293,027	0.74	0.074
11	on	291,947	0.73	0.081
12	he	250,919	0.63	0.076
13	is	245,843	0.62	0.080
14	with	223,846	0.56	0.079
15	at	210,064	0.53	0.079
16	by	209,586	0.53	0.084
17	it	195,621	0.49	0.084
18	from	189,451	0.48	0.086
19	as	181,714	0.46	0.087
20	be	157,300	0.40	0.079
21	were	153,913	0.39	0.081
22	an	152,576	0.38	0.084
23	have	149,749	0.38	0.087
24	his	142,285	0.36	0.086
25	but	140,880	0.35	0.089

r	w	frequency	$P \cdot 100$	$P \cdot r$
26	has	136,007	0.34	0.089
27	are	130,322	0.33	0.089
28	not	127,493	0.32	0.090
29	who	116,364	0.29	0.085
30	they	111,024	0.28	0.084
31	its	111,021	0.28	0.087
32	had	103,943	0.26	0.084
33	will	102,949	0.26	0.085
34	would	99,503	0.25	0.085
35	about	92,983	0.23	0.082
36	i	92,005	0.23	0.083
37	been	88,786	0.22	0.083
38	this	87,286	0.22	0.083
39	their	84,638	0.21	0.083
40	new	83,449	0.21	0.084
41	or	81,796	0.21	0.084
42	which	80,385	0.20	0.085
43	we	80,245	0.20	0.087
44	more	76,388	0.19	0.085
45	after	75,165	0.19	0.085
46	us	72,045	0.18	0.083
47	percent	71,956	0.18	0.085
48	up	71,082	0.18	0.086
49	one	70,266	0.18	0.087
50	people	68,988	0.17	0.087

Text Statistics

Term Frequency: Zipf's Law (continued)

Example: Top 50 most frequent words from AP89. For English: $c \approx 0.1$.

r	w	frequency	$P \cdot 100$	$P \cdot r$
1	`the`	2,420,778	6.09	0.061
2	`of`	1,045,733	2.63	0.053
3	`to`	968,882	2.44	0.073
4	`a`	892,429	2.25	0.090
5	`and`	865,644	2.18	0.109
6	`in`	847,825	2.13	0.128
7	`said`	504,593	1.27	0.089
8	`for`	363,865	0.92	0.073
9	`that`	347,072	0.87	0.079
10	`was`	293,027	0.74	0.074
11	`on`	291,947	0.73	0.081
12	`he`	250,919	0.63	0.076
13	`is`	245,843	0.62	0.080
14	`with`	223,846	0.56	0.079
15	`at`	210,064	0.53	0.079
16	`by`	209,586	0.53	0.084
17	`it`	195,621	0.49	0.084
18	`from`	189,451	0.48	0.086
19	`as`	181,714	0.46	0.087
20	`be`	157,300	0.40	0.079
21	`were`	153,913	0.39	0.081
22	`an`	152,576	0.38	0.084
23	`have`	149,749	0.38	0.087
24	`his`	142,285	0.36	0.086
25	`but`	140,880	0.35	0.089

r	w	frequency	$P \cdot 100$	$P \cdot r$
26	`has`	136,007	0.34	0.089
27	`are`	130,322	0.33	0.089
28	`not`	127,493	0.32	0.090
29	`who`	116,364	0.29	0.085
30	`they`	111,024	0.28	0.084
31	`its`	111,021	0.28	0.087
32	`had`	103,943	0.26	0.084
33	`will`	102,949	0.26	0.085
34	`would`	99,503	0.25	0.085
35	`about`	92,983	0.23	0.082
36	`i`	92,005	0.23	0.083
37	`been`	88,786	0.22	0.083
38	`this`	87,286	0.22	0.083
39	`their`	84,638	0.21	0.083
40	`new`	83,449	0.21	0.084
41	`or`	81,796	0.21	0.084
42	`which`	80,385	0.20	0.085
43	`we`	80,245	0.20	0.087
44	`more`	76,388	0.19	0.085
45	`after`	75,165	0.19	0.085
46	`us`	72,045	0.18	0.083
47	`percent`	71,956	0.18	0.085
48	`up`	71,082	0.18	0.086
49	`one`	70,266	0.18	0.087
50	`people`	68,988	0.17	0.087

Remarks:

- Collection statistics for AP89:

Total documents	84,678
Total word occurrences	39,749,179
Vocabulary size	198,763
Words occurring > 1000 times	4,169
Words occurring once	70,064

Text Statistics

Term Frequency: Zipf's Law (continued)

For relative frequencies, c can be estimated as follows:

$$1 = \sum_{i=1}^{n} P(w_i) = \sum_{i=1}^{n} \frac{c}{r(w_i)} = c \sum_{i=1}^{n} \frac{1}{r(w_i)} = c \cdot H_t, \quad \rightsquigarrow \quad c = \frac{1}{H_t} \approx \frac{1}{\ln(t)}$$

where t is the size $|V|$ of the vocabulary V, and H_n is the n-th harmonic number.

Constant c is language-dependent; e.g., for German $c = 1/ln(7.841.459) \approx 0.063$. [Wortschatz Leipzig]

Thus, the expected average number of occurrences of a word w in a document d of length m is

$$m \cdot P(w),$$

since $P(w)$ can be interpreted as a term occurrence probability.

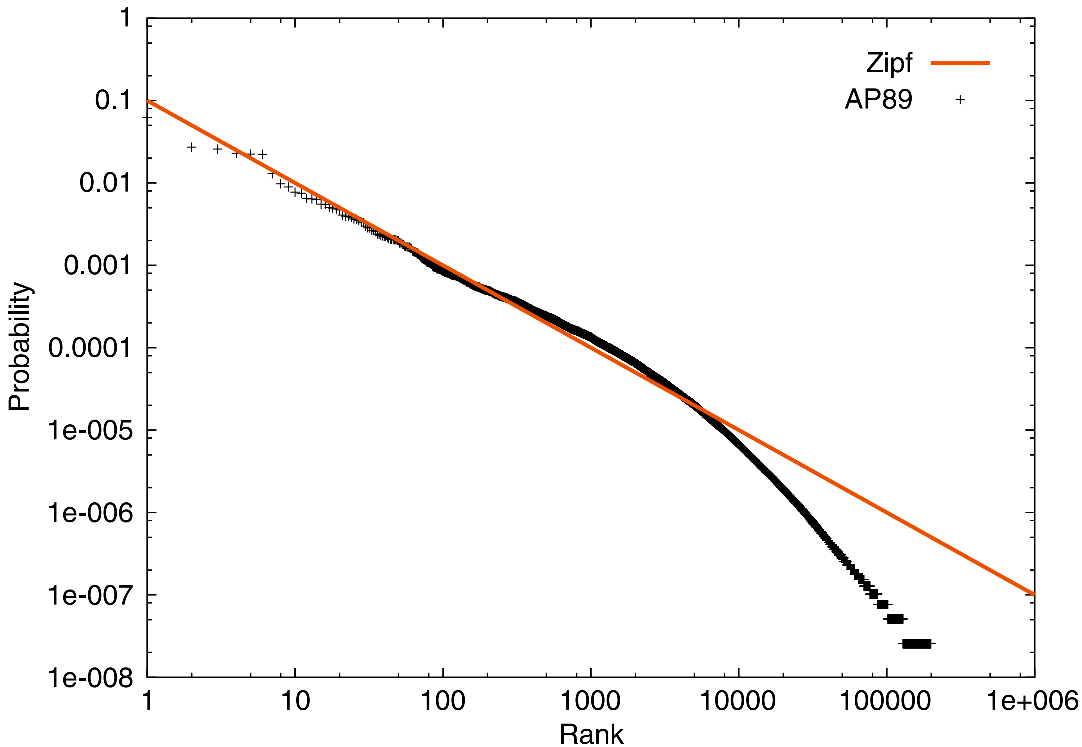
Text Statistics

Term Frequency: Zipf's Law (continued)

By logarithmization a linear form is obtained, yielding a straight line in a plot:

$$\log(P(w)) \;=\; \log(c) - a \cdot \log(r(w))$$

Example for AP89:



Remarks:

- As with all empirical laws, Zipf's law holds only approximately. While mid-range ranks of the frequency distribution fit quite well, this is less so for the lowest ranks and very high ranks (i.e., very infrequent words). The Zipf-Mandelbrot law is an extension of Zipf's law that provides for a better fit.

$$n \approx \frac{1}{(r(w) + c_1)^{1+c_2}}$$

- Interestingly, this relation cannot only be observed for words and letters in human language texts or music score sheets, but for all kinds of natural symbol sequences (e.g., DNA). It is also true for randomly generated character sequences where one character is assigned the role of a blank space. [Li 1992]

- Independently of Zipf's law, a special case is Benford's law, which governs the frequency distribution of first digits in a number.

Text Statistics

Term Frequency: Zipf's Law (continued)

For the vocabulary, t (types) is as large as the largest rank of the frequency-sorted list. For words with frequency 1:

$$P(w) = \frac{n_w}{N}, \quad t = r(n_w = 1) = c \times \frac{N}{1} = c \times N \approx \; e^{1/c}$$

Proportion of word forms that occur only n time. For $\mathbf{w}_n$ applies:

$$\mathbf{w}_n = r(n_w) - (r(n_w) + 1) = c \times \frac{N}{n} - c \times \frac{N}{n+1} = \frac{c \times N}{n(n+1)} = \frac{t}{n(n+1)}$$

For $\mathbf{w}_1$ applies in particular:

$$\mathbf{w}_1 = \frac{t}{2}$$

Half of the vocabulary in a text probably occurs only 1 time.

Text Statistics

Term Frequency: Zipf's Law (continued)

The ratio of words with a given absolute frequency n can be estimated by

$$\frac{\mathbf{w}_n}{t} = \frac{\frac{t}{n(n+1)}}{t} = \frac{1}{n(n+1)}$$

Observations:

- Estimations are fairly accurate for small x.
- Roughly half of all words can be expected to be unique.

Applications:

- Estimation of the number of word forms that occur n times in the text.
- Estimation of vocabulary size
- Estimation of vocabulary growth as text volume increases
- Analysis of search queries
- Term extraction (for indexing)
- Difference analysis (comparison of documents)