Statistically Understanding Text

Function Words vs. Content Words

Function words have little lexical meaning but serve as important elements to the structure of sentences.

Function Words vs. Content Words

Function words have little lexical meaning but serve as important elements to the structure of sentences.

Example

- **Prepositions**: in, of, between, on, with, by, at, without, through, over, across, around, into, within
- **Pronouns**: she, they, he, it, him, her, you, me, anybody, somebody, someone, anyone
- **Determiners**: the, a, that, my, more, much, either, neither
- **Conjunctions**: and, but, for, yet, neither, or, so, when, although, however, as, because, before
- **Auxiliary verbs**: be (is, am, are), have, got, do
- **Particles**: no, not, nor, as

Function Words vs. Content Words

Function words have little lexical meaning but serve as important elements to the structure of sentences.

Example

- **Prepositions**: in, of, between, on, with, by, at, without, through, over, across, around, into, within
- **Pronouns**: she, they, he, it, him, her, you, me, anybody, somebody, someone, anyone
- **Determiners**: the, a, that, my, more, much, either, neither
- **Conjunctions**: and, but, for, yet, neither, or, so, when, although, however, as, because, before
- **Auxiliary verbs**: be (is, am, are), have, got, do
- **Particles**: no, not, nor, as

Function words are closed-class words

Function Words vs. Content Words

Content words contain more lexical meaning than function words.

Example

- **Nouns**: john, room, answer
- **Adjectives**: happy, new, large, grey
- **Full verbs**: search, grow, hold, have
- **Adverbs**: really, completely, very, also, enough

Function Words vs. Content Words

Examples

- Our friends called us yesterday and asked if we'd like to visit them next month.
- The best time to study is early in the morning or late in the evening.

Function Words vs. Content Words

Examples

- Our friends called us yesterday and asked if we'd like to visit them next month.
- The best time to study is early in the morning or late in the evening.

Tom Sawyer (by Mark Twain)

Text download: https://www.gutenberg.org/files/74/74-0.txt

Most Common Words in Tom Sawyer

Word	Freq.	Use
the	3332	determiner (article)
and	2972	conjunction
a	1775	determiner
to	1725	preposition, verbal infinitive marker
of	1440	preposition
was	1161	auxiliary verb
it	1027	(personal/expletive) pronoun
in	906	preposition
that	877	complementizer, demonstrative
he	877	(personal) pronoun
I	783	(personal) pronoun
his	772	(possessive) pronoun
you	686	(personal) pronoun
Tom	679	proper noun
with	642	preposition

Most words are smaller in length but have important grammatical roles.
They are determiners, prepositions, conjunctions, pronouns, etc.

What about other texts?

Shakespeare

https://ocw.mit.edu/ans7870/6/6.006/s08/lecturenotes/files/t8.shakespeare.txt

News articles

https://www.kaggle.com/uciml/news-aggregator-dataset

Amazon reviews

https://www.kaggle.com/bittlingmayer/amazonreviews

Type vs. Tokens

Types

Number of distinct words in the corpus (size of vocabulary).

Tokens

Total number of running words in the corpus.

Example

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

Tokens: 16

Types: 14

Type/Token Ratio

TTR

- The type/token ratio (TTR) is the ratio of the number of different words (types) to the number of running words (tokens) in a given text or corpus.
- This index indicates how often, on average, a new ‘word form’ appears in the text or corpus.

Comparison Across Texts

Mark Twain's Tom Sawyer

- 77,491 word tokens
- 8,486 word types
- TTR = 0.11

Complete Shakespeare work

- 928,012 word tokens
- 29,454 word types
- TTR = 0.032

Empirical Observations on Various Texts

Comparing Conversation, academic prose, news, fiction

Longman Grammar of Spoken and Written English, Biber et al. (1999).

- TTR scores the lowest value (tendency to use the same words) in conversation.
- TTR scores the highest value (tendency to use different words) in news.
- Academic prose writing has the second lowest TTR.

Empirical Observations on Various Texts

Comparing Conversation, academic prose, news, fiction

Longman Grammar of Spoken and Written English, Biber et al. (1999).

- TTR scores the lowest value (tendency to use the same words) in conversation.
- TTR scores the highest value (tendency to use different words) in news.
- Academic prose writing has the second lowest TTR.

Not a valid measure of ‘text complexity’ by itself

- The value varies with the size of the text.
- For a valid measure, a running average is computed on consecutive 1000-word chunks of the text.

Word Distribution from Tom Sawyer

Frequency	Frequency of frequency
1	4222
2	1398
3	705
4	454
5	245
6	213
7	174
8	141
9	85
10	93
11	61
12	55
13	50
14	45
15	26

- TTR = 0.11 $\Rightarrow$ Words occuron average 9 times each.
- But words have a very uneven distribution.

Word Distribution from Tom Sawyer

Frequency	Frequency of frequency
1	4222
2	1398
3	705
4	454
5	245
6	213
7	174
8	141
9	85
10	93
11	61
12	55
13	50
14	45
15	26

- TTR = 0.11 ⇒ Words occur on average 9 times each.
- But words have a very uneven distribution.

Most words are rare

- 4222 (50%) word types appear only once
- They are called *happax legomena* (Greek for 'read only once')

Word Distribution from Tom Sawyer

Frequency	Frequency of frequency
1	4222
2	1398
3	705
4	454
5	245
6	213
7	174
8	141
9	85
10	93
11	61
12	55
13	50
14	45
15	26

- TTR = 0.11 ⇒ Words occur on average 9 times each.
- But words have a very uneven distribution.

Most words are rare

- 4222 (50%) word types appear only once
- They are called *happax legomena* (Greek for ‘read only once’)

But common words are very common

- 100 words account for 51% of all tokens of all text

Zipf's Law

- Count the frequency of each word type in a large corpus
- List the word types in decreasing order of their frequency

Zipf's Law

A relationship between the frequency of a word (f) and its position in the list (its rank r).

$$f \propto \frac{1}{r}$$

or, there is a constant k such that

$$f.r = k$$

i.e. the 50th most common word should occur with 3 times the frequency of the 150th most common word.

Zipf's Law

Let

- p_r denote the probability of word of rank r
- N denote the total number of word occurrences

$$p_r = \frac{f}{N} = \frac{A}{r}$$

The value of A is found closer to 0.1 for corpus

Empirical Evaluation from Tom Sawyer

Freq(f)	Rank(r)	f*r	Freq(f)	Rank(r)	f*r
3523	1	3523	43	243	10449
3052	2	6104	43	244	10492
1861	3	5583	43	245	10535
1797	4	7188	43	246	10578
1565	5	7825	43	247	10621
1165	6	6990	42	248	10416
1144	7	8008	41	249	10209
1018	8	8144	41	250	10250
975	9	8775	41	251	10291
970	10	9700	41	252	10332
929	11	10219	41	253	10373
869	12	10428	41	254	10414

Zipf's Other Laws

Correlation: Number of meanings and word frequency

The number of meanings m of a word obeys the law:

$$m \propto \sqrt{f}$$

Zipf's Other Laws

Correlation: Number of meanings and word frequency

The number of meanings m of a word obeys the law:

$$m \propto \sqrt{f}$$

Given the First law

$$m \propto \frac{1}{\sqrt{r}}$$

Zipf's Other Laws

Correlation: Number of meanings and word frequency

The number of meanings m of a word obeys the law:

$$m \propto \sqrt{f}$$

Given the First law

$$m \propto \frac{1}{\sqrt{r}}$$

Empirical Support

- Rank $\approx$ 10000, average 2.1 meanings
- Rank $\approx$ 5000, average 3 meanings
- Rank $\approx$ 2000, average 4.6 meanings

Zipf's Other Laws

Correlation: Word length and word frequency

Word frequency is inversely proportional to their length.

$$l \propto \frac{1}{f}$$

Zipf's Other Laws

Correlation: Word length and word frequency

Word frequency is inversely proportional to their length.

$$l \propto \frac{1}{f}$$

Given the First law

$$l \propto r$$

Impact of Zipf's Law

The Good part

Functional words account for a large fraction of text, thus eliminating them greatly reduces the number of tokens in a text.

The Bad part

Most words are extremely rare and thus, gathering sufficient data for meaningful statistical analysis is difficult for most words.

Vocabulary Growth

How does the size of the overall vocabulary (number of unique words) grow with the size of the corpus?

Heaps' Law

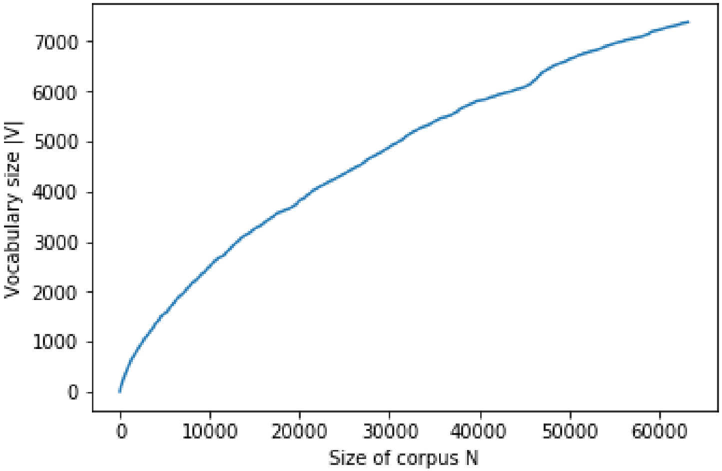
Let $|V|$ be the size of vocabulary and N be the number of tokens.

$$|V| = KN^{\beta}$$

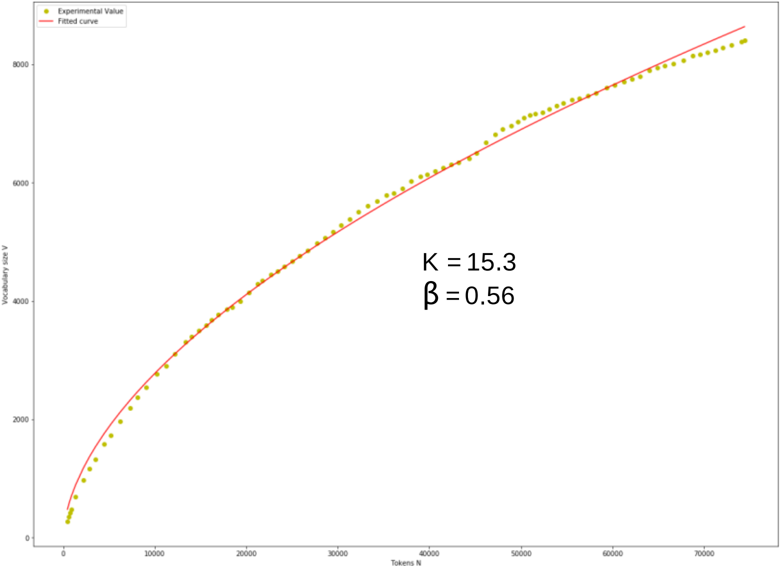
Typically

- K $\approx$ 10-100
- $\beta \approx$ 0.4 - 0.6 (roughly square root)

Heaps' Law: Empirical evidence from last year assignments



Heaps' Law: Empirical evidence from last year assignments



Take Home Exercise

Tom Sawyer

- Download Tom Sawyer dataset.
- Compute tokens, types, and TTR.
- Check if Zipf's law holds true for meanings and length.
- Plot Heaps' law