Functional Vs Content Words

- Function words have little lexical meaning but serve as important elements to the structure of sentences.
- Prepositions, pronouns, auxiliary verbs, conjunctions, grammatical articles, etc.

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

The list is dominated by the little words of English, having important grammatical roles.

- Type-Token distinction
- Type-token distinction is a distinction that separates a concept from the objects which are particular instances of the concept
- 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.

71,370 word tokens

8,018 word types

TTR = 0.112

884,647 word tokens

29,066 word types

TTR = 0.032

- 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.

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.

NPTEL

Word	Freq. (f)	Rank (r)	$f \cdot r$	Word	Freq. (f)	Rank (r)	$f \cdot r$
the	3332	1	3332	turned	51	200	10200
and	2972	2	5944	you'll	30	300	9000
a	1775	3	5235	name	21	400	8400
he	877	10	8770	comes	16	500	8000
but	410	20	8400	group	13	600	7800
be	294	30	8820	lead	11	700	7700
there	222	40	8880	friends	10	800	8000
one	172	50	8600	begin	9	900	8100
about	158	60	9480	family	8	1000	8000
more	138	70	9660	brushed	4	2000	8000
never	124	80	9920	sins	2	3000	6000
Oh	116	90	10440	Could	2	4000	8000
two	104	100	10400	Applausive	1	8000	8000

NPTEL

- Word frequency is inversely proportional to their length.
- Stopwords account for a large fraction of text, thus eliminating them greatly reduces the number of tokens in a text.
- 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)

NPTEL