19CSE453 – Natural Language Processing

Empirical Laws

By

Ms. Kavitha C.R.

Dept. of Computer Science and Engineering

Amrita School of Computing, Bengaluru

Amrita Vishwa Vidyapeetham

Function Words vs. Content Words

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

Example

Tom Sawyer lives with his Aunt Polly and his half-brother Sid.

Summer arrives, and Tom and Huck go hunting for buried treasure in a haunted house.

Function words are closed-class words

prepositions, pronouns, auxiliary verbs, conjunctions, grammatical articles, particles 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.**
- **These are usually referred to as *function words*, such as determiners, prepositions, complementizers 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 one really exceptional word is *Tom*, whose frequency reflects the text chosen.

How many words are there in this text?

Type vs. Tokens

Type-Token distinction

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

Type/Token Ratio

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.

Type vs. Tokens

Type-Token distinction

There has been *a* lot of fuzz on the topic. A solution has been arrived at.

Type: 13
Tokens: 16
TTR=13/16

Comparison Across Texts

Mark Twain's Tom Sawyer

71,370 word tokens

8,018 word types

TTR = 0.112

Complete Shakespeare work

884,647 word tokens

29,066 word types

TTR = 0.032

Empirical Observations on Various Texts

- ***Comparing Conversation, academic prose, news, fiction***
  - 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

Word Frequency	Frequency of Frequency
1	3993
2	1292
3	664
4	410
5	243
6	199
7	172
8	131
9	82
10	91
11-50	540
51-100	99
> 100	102

Sum : 8018

TTR = 0.11 $\Rightarrow$ Words occur on average 9 times each.

But words have a very uneven distribution.

Most words are rare

3993 (50%) word types appear 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).

$$f \propto 1/r$$

or, there is a constant k such that

$$f.r = k$$

The	an	and	Tom
4000	1000	2000	1500

In descending order

the	4000	1
and	2000	2
Tom	1500	3
an	1000	4

Zipf's Law

Let

p_r **denote the probability of word of rank** r

N **denote the total number of word occurrences**

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

$$f \,.\, r = A.N$$

The value of A is found closer to 0.1 for corpus

Empirical Evaluation from Tom Sawyer

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

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…….

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

Correlation: Word length and word frequency

Word frequency is inversely proportional to their length.

Impact of Zipf’s Law

The Good part

Stopwords 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 in the corpus.

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

Typically, K $\approx$ 10-100

$\beta \approx$ 0.4 - 0.6 (roughly square root of N)

Thank You