ZIPF'S LAW: THE HIDDEN FEATURES

Savanur, S. K.*

ABSTRACT:

{*An attempt is made to give mathematical and nonmathematical explanation of Zipf's law. Many hidden features and applications are illustrated. Zipf's law. The law is extended in relation to word-length*}

Keywords: *Zipf's Law*. *Word-Length Frequency Relationship Pareto's Law*

1. INTRODUCTION

George Kingsley Zipf (1902-1950) formulated one of the fundamental laws of Bibliometrics. He stated that the probability of occurrence of words or other items starts high and tapers off. In this context, the 'other items' imply that the law is applicable, not only to the words in a text, but also to other objects of study. These will be covered under the section 'Other Applications'.

2. MATHEMATICALLY SPEAKING…

Zipf gave the following formula to illustrate the law.

$P_n \approx 1/n^a$

> [Note: This type of laws, where n^a i.e. 'n to the power a' is involved, are called 'Power Laws'. Boyle's Law, Ohm's Law, inverse square laws used in optics, magnetism, gravity etc. follow power laws.]

**Department of Library and Information Science Joshi-Bedekar College, Thane*
sksavanur@rediffmail.com

Where, P_n is the frequency of occurrence of n^{th} ranked item and 'a' is nearly equal to 1. Thus, the formula is reduced to,

$$P_n \approx 1/n^1 = 1/n \text{ ---------- A}$$

Therefore, a few words occur very often and others may appear rarely. The law itself was too simple, like most of the fundamental laws: In fact it is so simple and obvious that one wonders, why it was stated in the first place! The statement of the law, in its simplest form, is as follows:

The frequency of occurrence of a word in text is roughly inversely proportional to the rank of that word in that text. If frequency of a word is 'F' its rank is 'R',

$$F \propto 1/R \text{---------- B}$$

Please notice the commonness between equations A and B. Different authors introduce this idea in different ways, but equation B is used for further discussions in this paper.

Consider equation B. If the 'proportionality' is to be replaced with 'equality', a constant of proportionality has to be introduced. In other words, now the equation B appears as the following equation C. [Such laws are also referred as Reciprocal Laws.]

$$F = K/R \text{ where K is a constant ---------- C}$$

By transposing R in the denominator to the left hand side of the equality sign, the equation D is obtained.

F*R =K i.e. the product of 'Frequency' and 'Rank' is a constant. ----D

To illustrate this, consider a document with 754 words with the following word frequencies and the corresponding ranks. (Please see Table 1.) According to the Zipf's law, as the 'Frequency' of a word increases, the numerical value of the corresponding 'Rank' decreases, and the product of the two is a constant. This appears too obvious but it is not.

When two variables [In this case 'Frequency' and 'Rank'.] are inversely proportional, their product is a constant (Equation D) which is approximately equal for all words in a text. The product is not numerically constant, but it is reasonably equal with least variation within a small range. The Standard Deviation in this range is minimal. This area is shown in grey shade in Table 1. At the extreme beginning and end of the Table, the product is varying in higher range.

The Zipf's law is shown graphically in Fig. 1.

There is another way in which Zipf's law can be shown graphically. Consider equation C.

F = K/R where K is a constant. The 'R' in the denominator can be written as a negative power. Therefore, equation C becomes equation E.

$$\mathbf{F = KR^{-1}} \text{ ---------- E}$$

$$\mathbf{Log\ F = Log\ K + (-1)\ Log\ R} \text{ ----------F}$$

[Based on the First Law of Logarithm: logAB = Log A+ logB

and the Third Law of Logarithm: $\log A^n = n \log A$]

Table 1 Frequency-Rank

Word	Frequency (F)	Rank (R)	Product F*R
the	189	1	189
to	97	2	194
of	67	3	201
and	60	4	240
a	48	5	240
will	40	6	240
in	34	7	238
is	30	8	240
oil	26	9	234
with	24	10	240
dollar	23	11	253
for	22	12	264
that	22	13	286
their	22	14	308
this	20	15	300
by	19	16	304
it	17	17	289
be	16	18	288
Total	**754**	-	-

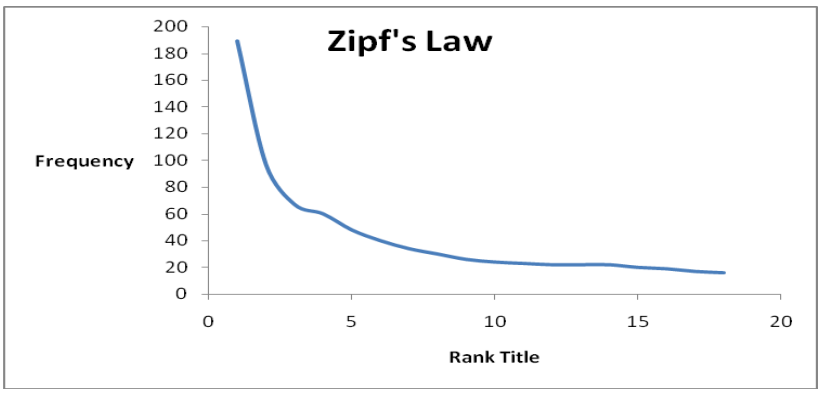


Fig. 1: Zipf's Law

Equation F is like a normal linear equation y = a + bx where 'a' is a constant or y-intercept and 'b' is the slope or coefficient of correlation or gradient. In equation F, -1 is coefficient of correlation and log k is a constant and is equal to y-intercept.

3. RATIONALE OF ZIPF'S LAW

Zipf thought the rationale for the law is in his '**Principle of Least Effort**', which he defined as follows:

The Principle of Least Effort means that a person...will strive to solve his problems in such a way as to minimize the total work that he must expend in solving both his immediate problems and his probable future problems. That, in turn, means that the person will strive to minimize the probable ***average*** rate **of** his work-expenditure (over **time). And in so doing, he will be minimizing his *effort,* by our definition of effort.* Least effort, therefore, is a variant of least work.** Italics in the original.) (Wyllys, 1981)

[http://www.hermetic.ch/wfca/zipf.htm]

[*The definition of 'effort' is built in the above paragraph. In that, 'effort' is the total work that one must expend in solving both his immediate and his probable future problems.]

According to this, an author avoids the strain or effort of searching for the specific words with the right shades and meanings. He/she relies on a limited vocabulary, hence certain words are used more frequently and words with specific meanings with less frequency.

3.1 Some Observations

After stating the rationale of Zipf's law, a few observations may be made at this stage to comprehend the features of the law.

- Zipf knew that the reciprocal law (i.e. Equation C) applies only to a large sample size. The maximum frequency and the maximum rank were equal to the intercepts of the line of best fit of slope (or gradient) equal to -1 (least squares in log scale). Of course, as the number of words in a text increases, the gradient tends to -1. Therefore, Zipf's idea was to select the size which gives gradient closest to -1 across a large number of samples of the texts. (Powers, 1998)

- Zipf's law shows that the findings are much better for the middle ranks than for the very extreme ends. This was shown as the grey shaded part of Table 1.

- Zipf worked with samples of various sizes. He suggests that the text should consist of at least 5000 words for the product **RF** to be reasonably constant, even in the middle ranks. (Wyllys, 1981)

- For most published works, the text should have around 10,000 words, while for children, at the time of starting school, it is around 2000 words. (Powers, 1998)

- In case of equally likely synonyms, one can think of replacing one with the other, so that only one word is used instead of the pair of synonyms. This will double its frequency and halve its rank. This is in line with Zipf's law. (Powers, 1998)

- Zipf's Law does not bother about the context and semantics. This means, the Zipf's Law does not distinguish the following,

    - The role of the word being counted. Eg: It does not consider whether it is a noun or a verb. Eg: Book, Press, Object, House etc.

- The regional variations in language
- The Meaning (i.e. Synonyms) and Capitalization (Eg.: Earth, earth) have no effect
- Context, i.e. distinction between 'Red' in 'Red tape', 'Red Carpet', 'Red Fort'
- The law is independent of the language. It is true for Tamil, Marathi, Kannada, Hindi etc.

- It distinguishes Singular-Plural and Tenses Eg: Boy-Boys, Man-Men, Run-Ran, Drink-Drank etc. One wonders what happens in the following situations.
  - Are the numbers treated as words? 2015, 7 etc
  - Are the symbols treated as words? $, %, * etc.
  - Are the punctuation marks treated as words? Eg: !, :, ; etc

4. ZIPF'S LAW AND LINGUISTICS

The above observations show how close the Zipf's Law is to Linguistics. In fact, George Zipf was an American Linguist and his work is more a law in Linguistics than in any other discipline. Going a little deeper into linguistics, any word in a text is called a '**Token**' The individual or unique words in a text are called '**Types**'. This means that a 'type' may appear more than once in a 'Token'. 'Tokens' are instances and 'Types' are the distinct words. (Baroni, 2007)

Consider an example. 'I played with the player who subsequently played with the Australian team.' This sentence has 12 tokens, i.e. words, and it has nine types or unique words. [I, played, with, the, player, who, subsequently, Australian, team]'

Please note that the **'Token' increases with the size of the text** or the document, but the number of **'Types' increases with the vocabulary of the author**. Also, note that Zipf's Law deals with 'types', their 'frequencies' and 'ranks' in a text.

Variations in the vocabulary of a text is given by the 'Type/Token Ratio'. (Gatt)

5. WORD-LENGTH AND FREQUENCY RELATIONSHIP

Wentian Li in his paper 'Random Texts Exhibit Zipf's-Law-Like Word Frequency Distribution', published in *IEEE Transactions on Information Theory*, showed that words generated by randomly combining letters fit the Zipf's distribution. In his randomly generated text, the frequency distribution of word length was exponential — that is, words of length 1 occurred more than words of length 2 and so forth, with frequency declining exponentially with word length.

But, in case of English Language, it is not the word-length 1 which has maximum frequency, but it is word-length 3. [Maximum number words are of word-length 3. Eg: the, are, his, her, you, our, can, see, for, but, two, red, bed, bet, bat, cat, rat, dog, mat, boy, man, men etc.] In short, word length in English does not follow an exponential distribution like a randomly generated text.. (Li, nd)

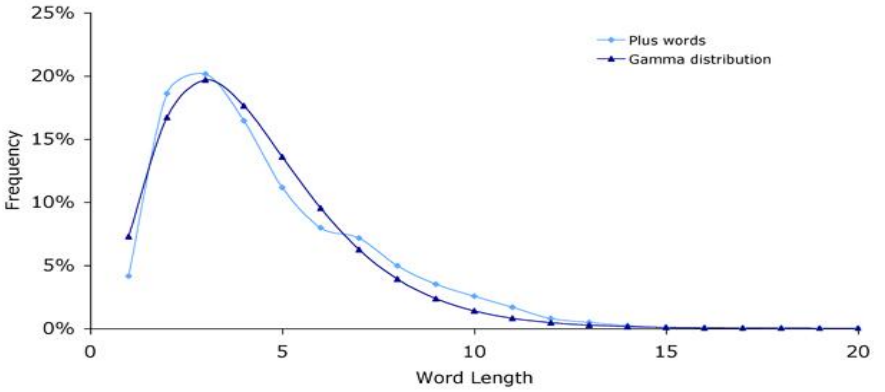


Fig. 2. Word Length-Frequency Relationship for English Language

Source: (The Mystery of Zipf, 2008)

6. APPLICATIONS IN LIBRARY AND INFORMATION SCIENCE (LIS)

LIS has the following select applications of the Zipf's law.

- The number of hits (visitors) on a websites
- Frequency of occurrence of different words in an article and the number of times an article is cited.
- The number of publications an author has contributed (Lotka's law)
- Number of articles and number of Journals (Bradford's Law)
- It is established earlier in this paper that that Zipf's law is more precisely observed in the middle range. All the common words like a, an, the, in, onto, for etc are in the initial range. These are actually 'Stop Words', in CDS/ISIS or WINISIS terminology. The upper range include certain less commonly used words with more length. The words in the middle range are those which indicate the subject content of the document. In other words, these are the

keyword. Thus, Zipf's law can be used in machine enabled indexing after scanning a paper with software meant for the purpose.

- As the Zipf's law is stated in terms of probability, whether one scans the entire paper or only its abstract, probability of occurrence of certain words remain the same. [Just as probability of getting 'Head' or 'Tail' is 0.5, for 10 or 100 or 1000 trials.] There is great economy of time and effort in getting the machine to identify keywords by merely scanning the author abstract instead of the entire document.

7. OTHER APPLICATIONS

Zipf's law has its roots in Linguistics. Researchers have applied the law to *the Bible* and many other types of texts.

- **"*The Bible*—**including both the Old and the New Testament—contains 17,859 different words, and a total of 796,888 words. We generated four sets of samples from the original book. Their respective sizes are: 40,000, 80,000, 160,000 and 320,000. As in the previous case, sampling proceeded according to the relative frequencies of the words." Tripp and Feitelson, Four sets of samples are made because, "the full data set diverges from Zipf's law. It drops much more steeply than predicted by the law. This trend becomes increasingly less pronounced as the population under consideration decreases in size. The other data sources … display varying instances of the same pattern." (Tripp & Feitelson, nd)

- Initially, Zipf's law was confined only to linguistics. Subsequently, it was realized that it is applicable to many social situations. Power laws relating rank to frequency have been demonstrated many situations.

- **Population and the size of cities**: Consider the **population** of a place and **size** are similar to **Frequency** and **Rank** respectively. To demonstrate this application Population of first 25 ranking cities was collected from *Wikipedia*. (Please see Appendix.) The Zipf's law can be seen in the fig.3

- With application of Zipf's Law to the population of Indian cities, one can realize that Chennai, Kolkata no longer belong to the top four cities. Bangalore, Hyderabad and Ahmedabad have overtaken theses cities. The population of Surat is very close to Kolkata and it is larger than Pune, in terms of population and its rank.

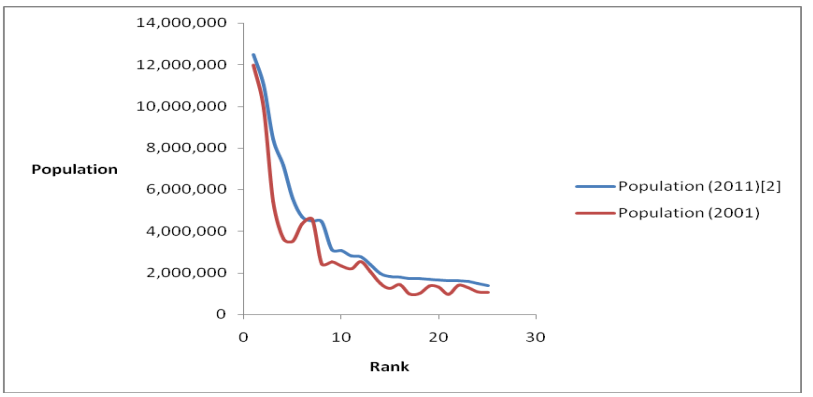


Fig 3: Population and Size of Indian Cities

- Similar, distribution can be seen in the US cities. The distribution is right skewed. This means that the "…bulk of the distribution occurs for fairly small sizes -most US cities have small populations -there is a small number of cities with population much higher than the typical value, producing the long tail to the right…." (Newman, 2005)

- Such applications help in arriving at new benchmarks in monitoring urban development and resource allocation for the purpose. It helps in developing new norms.Similarly, 'Income Distribution', 'Size of an organization/industry', 'Biological Genera and Species count' etc., follow Zipf's law.

- **Pareto's Law**: If cumulated frequencies are considered instead of plain frequency, it represents Pareto's law. Pareto's law established 80:20 formulae. For eg: 80% of circulation comes from 20% collection; 80% of library budget is spent for 20% collection etc. It has other connotations in management science. 80% of the work is carried out by 20% of the staff; 80% of problems are created by 20% workers etc.

All of these—word frequency, citation frequency, and publication frequency-obey a distribution called Zipf's law. Zipf's law applies as well to such diverse subjects. The law has been shown to encompass many natural phenomena, and is equivalent to the distributions of Yule, Lotka, Pareto, Bradford, and Price. (Fedorowicz, 1982) Thus, a simple, obvious law has many hidden features. Besides, there are still many unexplored areas of applications of Zipf's Law.

Appendix

Top 25 Cities and Their Population

City	Rank	Population (2011)	Population (2001)
Mumbai	1	12,478,447	11,978,450
Delhi	2	11,007,835	9,879,172
Bangalore	3	8,425,970	5,438,065
Hyderabad	4	7,170,545	3,637,483
Ahmedabad	5	5,570,585	3,520,085
Chennai	6	4,681,087	4,343,645
Kolkata	7	4,486,679	4,572,876
Surat	8	4,462,002	2,433,835
Pune	9	3,115,431	2,538,473
Jaipur	10	3,073,350	2,322,575
Lucknow	11	2,815,601	2,185,927
Kanpur	12	2,767,031	2,551,337
Nagpur	13	2,405,421	2,052,066
Indore	14	1,960,631	1,474,968
Thane	15	1,818,872	1,262,551
Bhopal	16	1,795,648	1,437,354
Visakhapatnam	17	1,730,320	982,904
Pimpri-Chinchwad	18	1,729,359	1,012,472
Patna	19	1,683,200	1,366,444
Vadodara	20	1,666,703	1,306,227
Ghaziabad	21	1,636,068	968,256
Ludhiana	22	1,613,878	1,398,467
Agra	23	1,574,542	1,275,134
Nashik	24	1,486,973	1,077,236
Faridabad	25	1,404,653	1,055,938

REFERENCES

Baroni, M. (2007). Distributions in Text. In A. Ludeling, & M. Kyto, Corpus Linguistics: An International Handbook (p. np). Berlin: Mouton de Gruyter. From http://sslmit.unibo.it/~baroni/publications/hsk_39_dist_rev2.pdf

Fedorowicz, J. (1982). The Theoretical Foundation of Zipf's Law and Its Application to the Bibliographic Database Environment. Journal of the American Society for Information Science , 285-293.

Gatt, A. (n.d.). Corpora and the Lexicon. Retrieved July 27, 2015, staff.um.edu.mt/albert.gatt/teaching/dl/lin3098-zipf.ppt

Li, W. (nd). Random Texts Exhibit Zipf's Law-Like Word Frequency Distribution. IEEE Transactions on Information Theory , np.

Newman, M. (2005). Power Laws, Pareto Distributions and Zipf' Law. Contemporary Physics , 323-351.

Powers, D. M. (1998). Applications and Explanations of Zipf's Law. In D. M. Powers, New Methods in Language Processing and Computational Natural Language Learning (pp. 151-160). np: ACL.

Tripp, O., & Feitelson, D. (nd). Zipf's Law Revisited. Retrieved July 26, 2015, leibniz.cs.huji.ac.il/tr/1048.pdf

West, M. (2008). The Mystery of Zipf. + Plus Magazine , file:///F:/Zipf's%20Law/Zipf%20Home/The%20mystery%20of%20Zipf%20_%20plus.maths.org.html.

West, M. (2008). The Mystery of Zipf. + Plus Magazine . https://plus.maths.org/content/mystery-zipf

Wyllys, R. E. (1981). Emperical and Theoretical Bases of Zipf's Law. Library Trends , 53-64. From

Zipf's Law (N.D.).Retrieved August 12, 2015, From Https://Www.Ideals.Illinois.Edu/Bitstream/Handle/2142/7182/Librarytrendsv30i1g_Opt.Pdf?Sequence=1