Zipf's Law in a Random Text from English With a New Ranking Method

Anurag Saxena[1], Monika Jauhari[2], B.M. Gupta[3]

1. Indira Gandhi National Open University, Maidan Garhi, New Delhi 110 068, India
anurags@ignou.ac.in
2. Associateship Graduate, INSDOC, New Delhi, India
3. National Institute of Science Technology and Developmental Studies,
Dr K.S.Krishnan Marg, New Delhi 110 012, India
bmgupta1@yahoo.com

Abstract

Zipf's law has attracted infometricians time and again. There have been many studies, which have explored the application of Zipf's law to various areas. However, there are a few parameters, which largely affect a study. These parameters are the power law embedded in Zipf's law, the ranking method, the type of text taken for the study and the behavior of extreme regions in the Zipf's curve. This communication tries to address all these points by taking a random text in English language from computer science literature. The selected text is called random because of its highly specific nature of technical words. The paper studies the properties of this text and compares the product of rank and frequency for three ranking procedures. It also analyses the performance of data in the extreme regions of the Zipf's curve. It is observed that ranking procedure and type of text have definite bearings on the performance of Zipf's curve.

1. Introduction

Zipf's law postulates that the frequency of occurrence of any word as a function of rank follows a power law with exponent close to unity. It has been applied to many areas like natural languages, monkey-typing texts, web-access statistics, informetrics, finance and business and ecological systems, etc. There is evidence of differences on whether the power law embedded in Zipf's law is actually a Yule distribution (Martindale, et al. 1996), lognormal distribution (Perline, 1996) or stretched exponential distribution (Laherrere, et al, 1998). There have been many applications of the law in natural languages, like English (Miller et. al. 1958), Chinese (Rousseau & Zhang, 1992), Voyanich manuscript (Landini, 1997), etc. However, there are few applications of the law to random texts. Li (1998) showed that the

Anurag Saxena, Monika Jauhari and B.M. Gupta, *Zipf's law in a random text from English with a new ranking method*, In: Guohua Jiang, Ronald Rousseau, Yishan Wu (Eds), Proceedings of the 9th International Conference on Scientometrics and Informetrics – ISSI 2003, Dalian University of Technology, Dalian, China, 2003, 271-279.

Zipf's law is applicable to random texts provided it has a very different word structure and length distribution than a natural language.

To investigate more into this area, the authors have selected a random text and have tried to find clues on the distribution of rank and frequency. An attempt has been made to evolve a new ranking method, based on tied-ranks and a comparison has been made with the random rank method, deployed by Zipf (1949) and maximum rank method, deployed by Chen & Leimkuhler (1987). According to Mandelbrot (1953), "The monkey language is, in the terminology of fractal geometry, self-similar and grows on infinite trees (any branch of the tree will be identical to the tree itself), thus needing an infinite dictionary. A natural language like English, on the other hand, is a massively geared down system that economizes on entropy in a number of ways, e.g., the interdependence—or redundancy—of words that seems necessary in order to make a text "meaningful." Most letter combinations (an uncountable set) in English are non-words". However, the random text taken for analysis in this communication is called "random" only because though it is in English, it follows a very subject specific usage of words, e.g. use of hyphenated words. Hence, in this communication, the random text used, differs from monkey typing text by only one virtue, i.e. every word in this random text has a definite meaning.

2. Methodology

To study the application of Zipf's law and the performance of the new ranking method on random texts, the authors have taken a text from a computer science " Operating System - Concepts and Design", by Milan Milenkovic , Second edition, 1997 (Tata McGraw Hill, New Delhi). The authors have counted the frequency of occurrence of each unique word in the text, and found 1775 unique or different words out of a total of 10,043 words in the full text. It was observed that the words of less than 9 characters in length were extensively used. However, one striking characteristic of computer science literature was the use of hyphenated words, which makes the word length vary over a large range. One can easily see from the table below that after words having 13 characters, there are a series of hyphenated words.

Table 1 Decription of words according to length and frequency

Word Example	Length	Frequency
A	1	205
AN	2	1765
CAD	3	1580
AREA	4	1100
LOGIN	5	730
DESIGN	6	856
ADDRESS	7	1076
LANGUAGE	8	844
INTERVALS	9	775
CONCURRENT	10	423
UTILIZATION	11	285
ABSTRACTIONS	12	165
COMMUNICATION	13	84
USER-SPECIFIED	14	37
CHANGE_PASSWORD	15	40
REMOTE-PROCEDURE	16	54
MEMORY-MANAGEMENT	17	7
PROGRAMMER-DEFINED	18	5
ADDRESS-TRANSLATION	19	4
LOWER-PRIORITY-BASED	20	3
COMPUTATION-INTENSIVE	21	2
TRANSACTION-PROCESSING	22	1
APPLICATION-PROGRAMMING	23	1

Use of hyphenated words can be taken as a special characteristic of the text taken, i.e. the computer science literature. It would thus be interesting to investigate the rank and frequency relationship as propounded by Zipf and other scientists in such a text. The authors have intentionally kept the hyphenated words as they are. One can also see that hyphenated words are typical in describing the very specific nature of the meaning they convey in the concerned literature. Some of them are the commands given to the computer to perform specific tasks.

All unique words were arbitrarily ranked according to their frequency of occurrence in a decreasing order. Words, which shared the same frequency, were arranged alphabetically and different ranks were assigned to each of them according to Zipf's approach of random-ranks. Thus, the words "able" got the rank(r) 868 and the word "writes" got the rank(r) 1775. One can see that two words contributing 1 occurrence each are assigned random ranks 868 and 1775, respectively according to Zipf's random rank approach. This leads to steps for large values of rank. This is one of the disadvantages with the random rank method. Chen and Leimkuhler (1987) had overcome this problem, by using the maximum rank for all the words with the same rank. Also their method helped in preserving the convertibility between frequency-rank distribution & frequency-count distribution and vice-versa, which was not possible in random rank approach. Another method proposed by us is based on the concept of "ties", which means, that if two observations are tied, i.e. they have the same frequency then they should be assigned the ranks according to the average of their random ranks. This was done in order to stabilize the product r x g(r), especially in the last rank-range. Here, r is the word rank and g (r) is the rank frequency i.e. the number of words of the same rank.

3. Analysis and Results

The authors had expected that the new ranking procedure based on “ties” would be able to minimize the dispersion of the product r x g(r) in all the rank range due to a simple logic that the maximum rank would always be greater than the average rank. A preliminary analysis of the product r x g(r) is as follows:

Table 2 Rank frequency relationships in different rank methods

Rank range	R x g(r) by Maximal Rank Method			r x g(r) by Tied Rank Method		
	Max	**Min**	**Std. Dev**	**Max**	**Min**	**Std. Dev**
1-10	1240	553	227.45	1377	553	227.4
11-51	1485	1239	57.79	1501	1239	62.85
52-99	1548	1352	56.23	1503	1352	46.15
108-228	1596	1512	30.99	1503	1456	16.29
276-1775	1775	1656	40.47	1538	1321.5	83.79

It can be seen from the above table that the r x g(r) is distributed with fairly less variability but for the rank-range (1-10). This is due to the fact that observation with rank 1 is a clear outlier. If we delete that observation from our calculation of standard deviation then the variability substantially reduces and comes down to 104.61 instead of 227.45. Also an interesting observation is that method of tied rank shows the same variability in the rank range (1-51), performs better in the rank range (52-228) and performs badly in the rank range (276-1775) when compared to the maximal rank method.

Table 3 Comparison of different models

Statistical Measure	Ranking Procedure		
	Zipf	***Chen***	***Tied***
std. Dev	223.76	99.14	86.47
Mean	1393.93	1718.16	1393.93
% c.v	16.052	5.77	6.20
min rank	1	1	1
max rank	1775	1775	1321.50

For linear fit y=a+bx
Parameters
	a =3.05
	b = -0.96
		a =2.99
		b = -0.91
			a =3.03
			b = -0.93

Standard Error	0.057	0.039	0.045
Correlation Coefficient	0.995	0.997	0.997

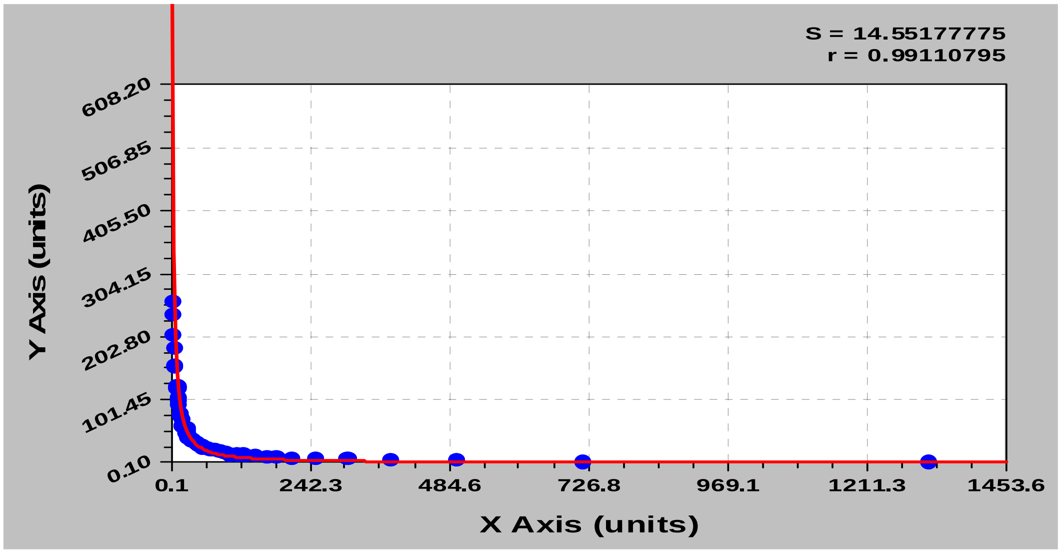
Here Standard Error (S) is the standard error of the estimate which quantifies the spread of data points around the regression curve and Correlation Coefficient (r) is the square-root of the normalized difference between the spread around mean and spread around the fitting function. As the regression model better describes the data, the correlation coefficient will approach unity. It can be seen that the random texts taken from the computer science literature do exhibit Zipf-like distribution with the slope of the linear regression touching unity. However, there is a marked difference in the performance of Maximal Rank and Tied Rank verses Random Rank of Zipf. There is a need to see whether the alternative ranking procedures perform better in other texts.

As far as the distribution of rank and frequency are concerned, it is found that the relation is a Shifted Power distribution (Mandelbrot Zipf's law) of the form

$$g(r) = a(r+b)^c$$

where the coefficients are estimated as a = 3301.44, b = -2.99 and c = -1.23

Table 4 Plot of tied-rank(x-axis) vs. frequency for the random text from computer science literature (where S and r are as defined above)



The authors have applied the fit on the Good's data used by Chen & Leimkuhler (1987) to check whether that also behaved in the similar manner. The fit behaved in the following manner:

$$g(r) = a(r+b)^{c}$$

The coefficient are estimated as a = 216.13, b=0 and c = -0.66

Table 5. Plot of maximal-rank(x-axis) vs. frequency for the Good's data (where S and r are as defined above)

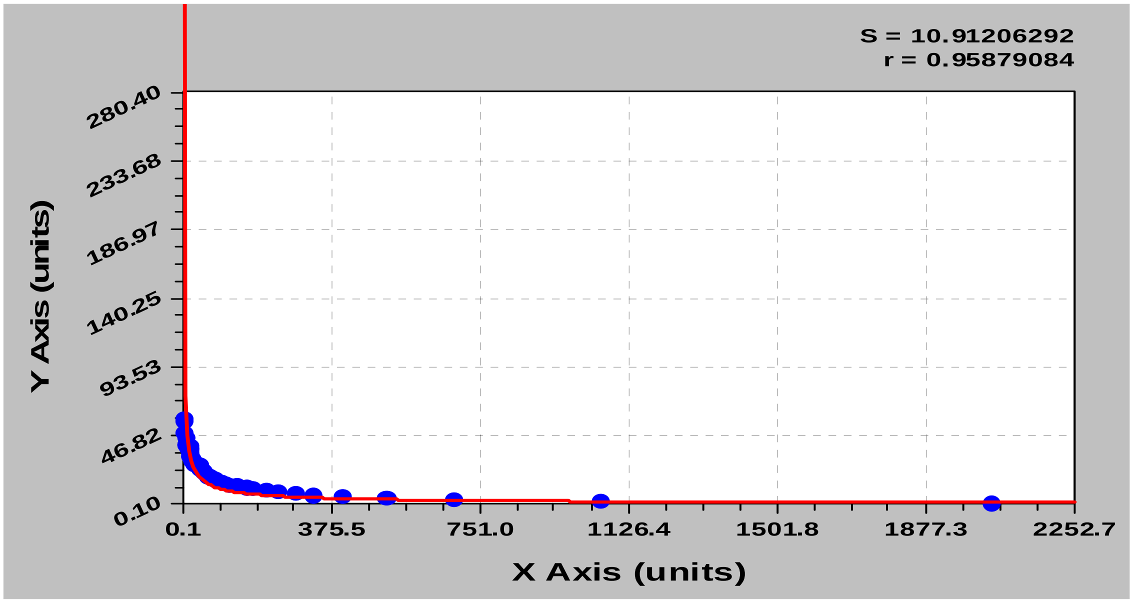
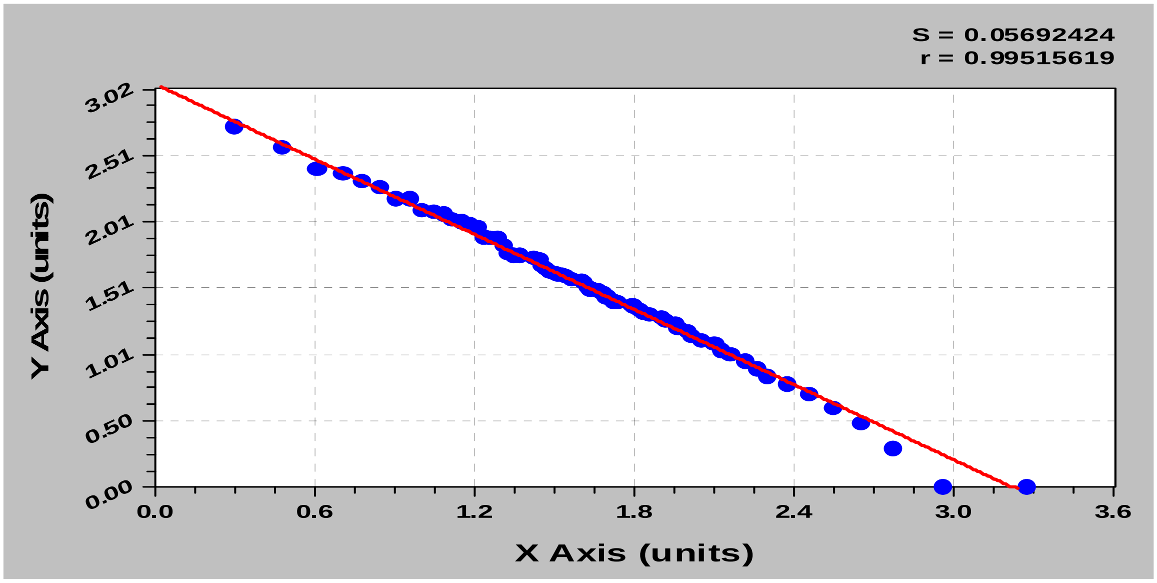
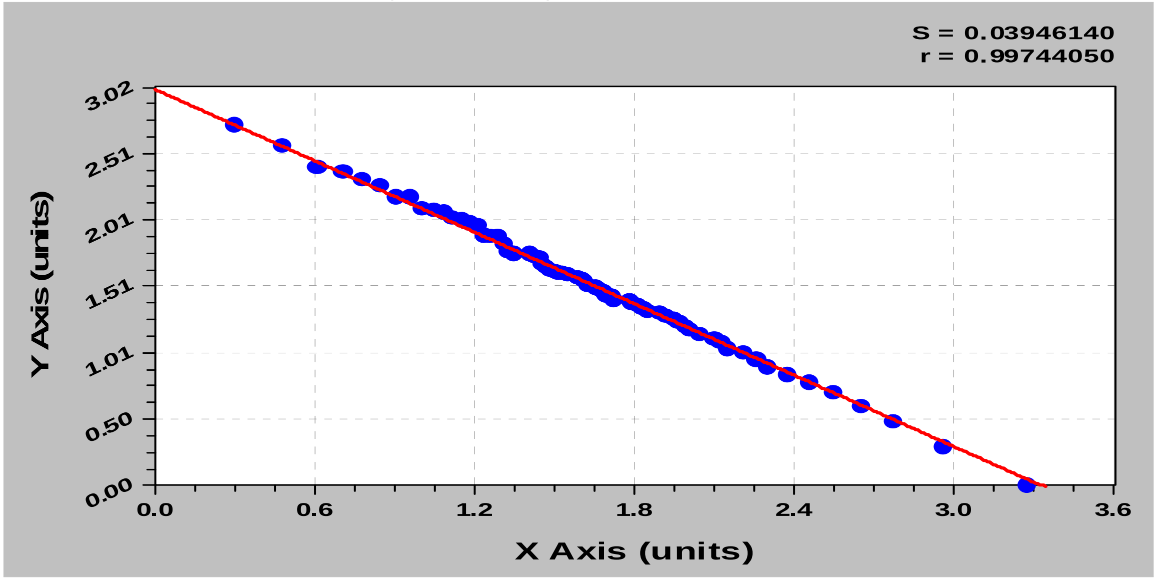


Table 6 Plot of log rank with log frequency for random rank method (here S and r are as defined above)



It can been seen that the power distribution (Mandelbrot Zipf's law) is fitting this type of data fairly well but with a slight modification in the form and parameters for different texts.. Besides this, the authors plotted the log rank with log frequency to see how the ranking methods fare. Here, the x-axis refers to the log-rank and y-axis to the log frequency.

Table 7 Plot of log rank with log frequency for maximal rank method



Tabel 8 Plot of log rank with log frequency for Tied rank method

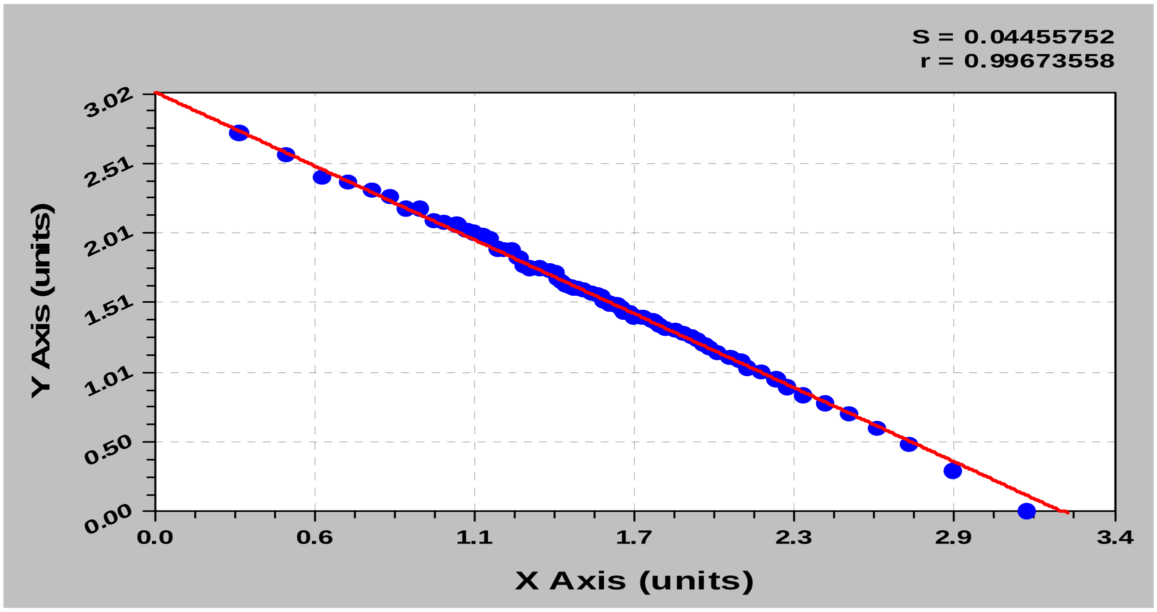
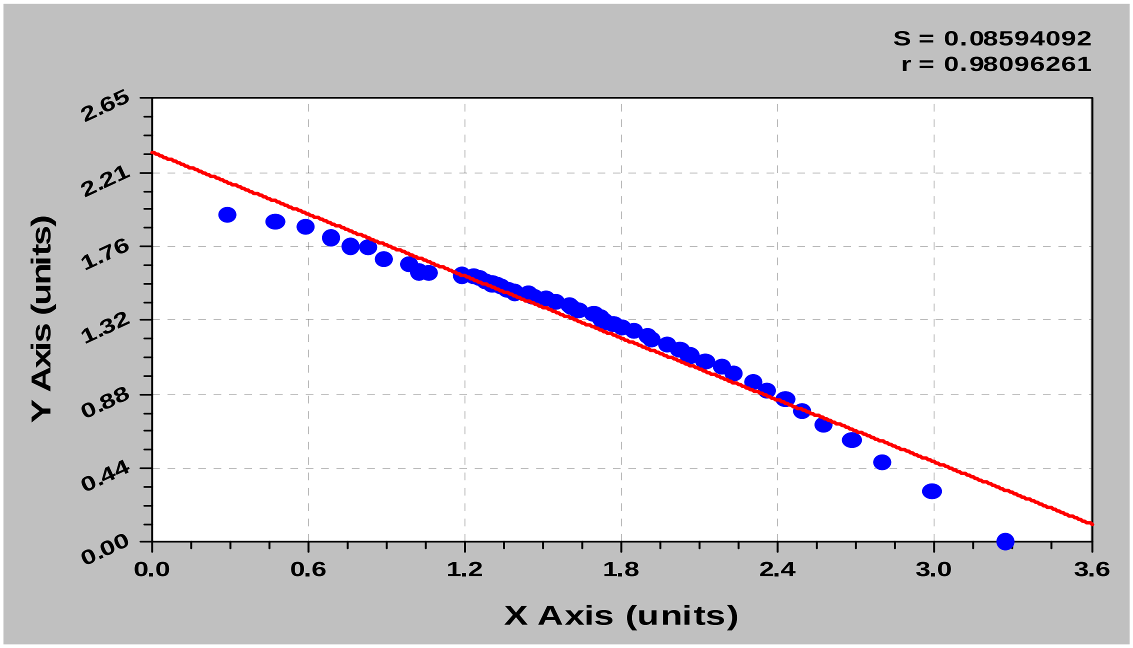


Table 9 Plot of log rank with log frequency for Good's data



It could be seen very clearly that both the Maximal rank method and Tied rank method perform better than the Random rank method of Zipf. It can be seen from the fits of the rank-range at the end. The purpose of analyzing Good's data here was just to give a picture that it did not fit the Zipf' law properly. The exponent in the fit of Good's data comes out to be -0.61, which was not close to –1 as propounded by Zipf.

4. Discussion

From the figures given in the earlier section it is evident that the lower tail (containing lower ranks) of the plot of log rank vs log frequency behaved in the best possible manner in the case of Maximal rank. The scatter in tied rank method was better than that in random rank method but not better than that in the maximal rank method. The question that naturally arises is whether the ranking method had a bearing on the type of text in question.

The analysis of Good’s data done by Chen and Leimkuhler was revisited in the earlier section and it is evident from the figure that The curve of log rank vs log frequency was not linear, specifically for Region I, as defined by Chen and Leimkuhler (Region I comprises higher ranks). This was a departure from their corollary 1 which says “In Region I Zipf-curve is linearly degreasing iff b=0”. The same concept if applied to our data gave the result– “ Curve is linearly decreasing even if b ≠ 0”.

5. Conclusions

There are two basic issues, which come out of this exercise. Firstly, random texts do follow Zipf’s law, however the exponent varies from text to text. The method of random rank performs inferiorly to the maximal rank method and the tied rank method proposed by authors, however there is a need for further investigation in this area as to ascertain whether the ranking method has a bearing on the type of text in question.

Secondly, the analysis of Good’s data forces us to raise some doubts about the generalizations of regions and the Mandelbrot-Zipf law (Chen and Leimkuhler 1987) which says “In region I. The Zipf-type curve is linearly decreasing iff b=0”. However, in region I of the plot of Good’s data the curve is not linearly decreasing even if b=0

References

1. Colin Martindale, Andrzej K Konopka (1996). Oligonucleotide frequencies in DNA follow a Yule distribution. *Computer & Chemistry* 20(1): 35-38.
2. Richard Perline (1996). Zipf's law, the central limit theorem, and the random division of the unit interval. *Physical Review* E 54(1): 220-223.
3. Jean Laherrere, D Sornette (1998). Stretched exponential distributions in nature and economy: 'Fat tails' with characteristic scales. *European Physical Journal* B2: 525-539.

4. GA Miller, EB Newman (1958). Tests of a statistical explanation of the rank-frequency relation for words in written English. *American Journal of Psychology*, 71, 209-218
5. R Rousseau, Qiaoqiao Zhang (1992). Zipf's data on the frequency of Chinese words revisited. *Scientometrics* 24(2): 201-220.
6. G Landini (1997). Zipf's laws in the Voynich manuscript. http://sun1.bham.ac.uk/G.Landini/evmt/zipf.htm"
7. W Li (1992). Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory* , 38(6): 1842-1845.
8. GK Zipf (1949). *Human Behavior and the Principle of Least Effort.* Addison-Wesley, 1949.
9. Chen Ye-Sho, Leimkuhler Ferdinand F. (1987). Analysis of Zipf's Law : An index approach. *Information Processing and Management* 23(3): 171-182.
10. Mandelbrot, B. An information theory of the statistical structure of language. *Proc. Symposium on Applications of Communication Theory*; September 1952; London: Butterworths; 1953; 486-500

Annex

rank(ran)	**g(r)**	**r(max)**	**g(rmax)**	**r(tied)**	**g(rt)**
1	553	1	553	1	553
2	545	2	545	2	545
3	375	3	375	3	375
4	259	4	259	4	259
5	238	5	238	5	238
6	204	6	204	6	204
7	184	7	184	7	184
8	155	8	155	8	155
9	153	9	153	9	153
10	124	10	124	10	124
11	121	11	121	11	121
12	118	12	118	12	118
13	105	13	105	13	105
14	103	14	103	14	103
15	99	15	99	15	99
16	92	16	92	16	92
17	79	17	79	17 18	79 77
18	77	18	77		
19	76	19	76	19	76
20	68	20	68	20	68
21	59	21	59	21	59
22	58	22	58	22	58

23	57	25	57	24	57
26	54	26	54	26	54
27	53	27	53	27	53
28	47	28	47	28	47
29	45	29	45	29	45
30	43	30	43	30	43
31	42	31	42	31	42
32	41	32	41	32	41
33	40	33	40	33	40
34	39	35	39	34.5	39
36	37	38	37	37	37
39	36	39	36	39	36
40	35	40	35	40	35
41	33	41	33	41	33
42	32	44	32	43	32
45	31	45	31	45	31
46	30	46	30	46	30
47	29	47	29	47	29
48	28	48	28	48	28
49	27	51	27	50	27
52	26	52	26	52	26
53	25	59	25	56	25
60	24	60	24	60	24
61	23	63	23	62	23
64	22	66	22	65	22
67	21	69	21	68	21
70	20	77	20	73.5	20
78	19	80	19	79	19
81	18	86	18	83.5	18
87	17	89	17	88	17
90	16	96	16	93	16
97	15	99	15	98	15
100	14	108	14	104	14
109	13	121	13	115	13
122	12	128	12	125	12
129	11	138	11	133.5	11
139	10	158	10	148.5	10
159	9	175	9	167	9
176	8	193	8	184.5	8
194	7	228	7	211	7
229	6	276	6	252.5	6
277	5	338	5	307.5	5
339	4	430	4	384.5	4
431	3	568	3	499.5	3
569	2	867	2	718	2
868	1	1775	1	1321.5	1

A revised version of this contribution has been published as:

Saxena, Anurag, Jauhari, Monika, Gupta, B. M. (2007). Zipf's law in a random text from english with a new ranking method. DESIDOC Bulletin of Information Technology, 27(4), 51-58