Malaysian Journal of Library & Information Science, Vol.3, no.2, July 1998:93-98

ZIPF'S LAW AND WRITINGS ON LIS

B K Sen; Khong Wye Keen; Lee Soo Hoon; Lim Bee Ling;
Mohd Rafae Abdullah; Ting Chang Nguan; Wee Siu Hiang
MLIS Programme, Faculty of Computer Science and
Information Technology, University of Malaya
50603 Kuala Lumpur, Malaysia
E-mail: bksen@fsktm.um.edu.my

ABSTRACT

Presents the results of a study conducted to find out the validity of Zipf's law related to the word length and the frequency of its uses in the case of library and information science literature. The results obtained from the analysis of six different samples obey Zipf's law in all the cases with small deviations. The result provided by the sample comprising about 5,800 words fits the law best with just one deviation. The main exception is found to be one-letter words.

Keywords: Zipf's law; Library and information science literature; Bibliometrics.

DEFINITIONS

Alpha-numeric expression - an expression involving alphabet/s and number/s, e.g. 2nd.
Alpha-symbolic expression - an expression involving alphabet/s and symbol/s, e.g. au=.
Extra-textual references - References appearing outside the text in the form of footnotes or citations at the end of the text.
Intra-textual references - References figuring within the text.

INTRODUCTION

George Kingsley Zipf, a noted linguist, tried to examine the field of linguistics from the scientific point of view and discovered three different interesting laws. One of them is that the length of a word is very closely related to the frequency of its usage - the greater the frequency, the shorter the word (Zipf, 1935).

As far as literary writings are concerned such as fictions, short stories, and poems this law of Zipf holds good. Does this law apply to technical writings as well? The question arose because technical writing differs from literary or ordinary writing in a number of ways. In technical writing, more often than not, each term represents a particular concept which is used again and again whenever the author refers to that concept thus leading to the increase in the frequency of its use. For example, an astronomer writing about the moon will be obliged to use the word moon again and again whenever referring to the concept. On the other hand a poet writing about the same subject can use the words moon, Luna, Cynthia, sailor’s friend, orb of night, and so on to break the monotony or to rhyme with another word. This apart, in technical writing, we may find tables, charts, formulas, symbols, etc. which are usually not part of an ordinary literary writing. As library and information science (LIS) literature falls under the category of technical writing, it was intended to study the validity of Zipf’s finding in this particular field.

Another object of the study was to find out the optimum sample size in terms of the number of words required for the verification of this particular law. In the case of the law related to rank and frequency (Zipf 1949), the optimum sample size is found to be at least 5,000 words (Wyllys, 1981).

METHODOLOGY

To conduct the study, six papers of varying length (Teh and Wong, 1996; Kademani and Kalyane, 1996; Nor, 1996; Panigrahi and Panda, 1996; Parvathamma, 1996; Zainab and Nor, 1996) were chosen from the *Malaysian Journal of Library and Information Science* (July 1996).

The portions of an article which were excluded from our study comprised the names of the authors, author affiliations, abstract, keywords, alpha-numeric expressions like 2nd and F10, alpha-symbolic expressions like au=, and su=, abbreviations such as FDT and ISO, numbers written with digits, serial numbering, formulas, punctuation marks, intra- and extra-textual references, tables, figures, and appendices.

The rationale behind the exclusion of the names of authors and author affiliations is obvious because they do not represent the author’s style of writing and as such cannot be used for word counting. The keywords are sometimes chosen by consulting a thesaurus, where the author has little choice and sometimes they are added by the editor. Hence, it was felt safe to exclude them. An abstract is the condensed version of an article and does not necessarily represent the normal style of writing of an author. Moreover, sometimes the abstract is prepared by someone other than the author. Therefore, it was not considered. Alpha-numeric as well as alpha-symbolic expressions, abbreviations, numbers written with digits, serial numbering with a, b, c, etc., and formulas are not words, hence excluded. The references comprise certain fixed elements such as author, year, title of the article, and other bibliographical details, which are not the creation of the author. Tables and figures were to be excluded for the ease of sorting. Moreover, at times a table may contain numerous YESes, NOes, etc. representing the answers of respondents. which are again not representative of the style of the author. This is the case with appendices as well. In the articles analysed for this study one appendix listed the papers of a famous scientist, several appendices listed CD ROM databases, and so on. Thus, only the textual part of the article with the above exceptions was considered since that part seemed to be most relevant for judging the word use pattern of an author. For the study, it has been assumed that the changes in the frequency of the use of words during the editorial process is minimal and does not affect the overall result. The length of the six articles in terms of words vary between 862 to 5772 words.

Each word length was measured in terms of letters. This was done to avoid the problem of hyphenated words which at times give rise to different counts. For example, the word length of 'on-line' is seven and 'online' is six when counted in terms of characters. But, in both the cases, the word length is six when counted in terms of letters. The manual process was obviously cumbersome, laborious and time consuming. Hence, an alternative method (detailed below) amenable to computerised analysis was tried.

Each article was scanned with an optical character recognition (OCR) software i.e. OmniPage Pro. The resultant file was saved in Microsoft Word for Windows format. The file so generated was checked with the original article for accuracy. The portions of the article as detailed above were removed. Punctuation marks were then converted into spaces, and subsequently all spaces were converted into line breaks. This resulted in a pure word list which was saved in a text file.

The text file was run through the program written in Turbo-Pascal to count the frequencies according to the number of letters in the words. The program took the text file as its input and produced the corresponding frequencies as its output.

Program

```
program Character_Count;

const
      MAXLENGTH = 20;   {maximum length of
words }

var
   ln                  : string[255];
   infile, outfile     : text;
   i                   : integer;
   count               : array [0..MAXLENGTH] of
                           integer;

begin
   for i:=0 to MAXlength do { Initialise}
         count[i] := 0;
   writeln( 'ChrCount - Character Count');
   write('Input file? ');
   readln(ln);
   assign(infile, ln);
   reset(outfile);

  write('Output file? ');
  readln(ln);
  assign(outfile, ln);
  rewrite(outfile);

   readln(infile, ln);    {Analyse}
   while not eof(infile), do
   begin
       count[length(ln)] := count[length(ln)] + 1;
       readln(infile, ln)
   end;

   for i:=0 to MAXLENGTH do {Output to file}
        writeln(outfile, i, chr(9), count[i];

  close(outfile);
  close(infile);

  writeln('Completed.')
end.
```

RESULTS

The results of the study are presented in Table 1. The percentage of one-letter word varies from 0.67 to 2.23. This low percentage

Table - 1 Word Length vs Frequency of Occurrence

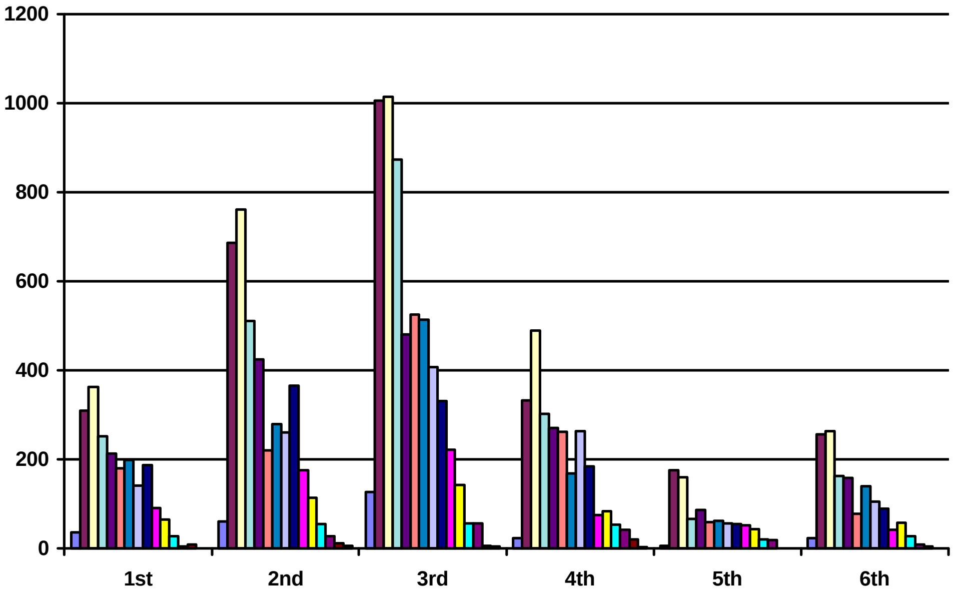
Word Length	Frequency & Percentage (Z-N)*		Frequency & Percentage (NEN)*		Frequency & Percentage (T-W)*		Frequency & Percentage (K-K)*		Frequency & Percentage (PA)*		Frequency & Percentage (P-P)*	
1	36	1.65	60	1.52	127	2.20	24	0.93	6	0.70	24	1.69
2	309	15.16	686	17.33	1006	17.45	333	12.87	176	20.42	256	18.04
3	362	17.47	761	19.23	1014	17.59	489	18.89	160	18.56	263	18.53
4	252	12.08	510	12.89	873	15.14	302	11.67	67	7.77	163	11.49
5	213	10.33	425	10.74	480	8.32	270	10.43	86	9.98	159	11.21
6	179	8.18	220	5.56	525	9.11	262	10.12	59	6.84	78	5.50
7	198	9.73	279	7.05	514	8.91	168	6.49	62	7.19	140	9.87
8	141	6.74	261	6.59	407	7.06	263	10.16	56	6.50	105	7.40
9	187	9.23	366	9.25	331	5.74	185	7.15	54	6.26	90	6.34
10	91	4.39	176	4.45	221	3.83	75	2.90	52	6.03	42	2.96
11	65	3.14	114	2.88	143	2.48	83	3.21	43	4.99	58	4.09
12	27	1.35	54	1.36	57	0.99	53	2.05	21	2.44	27	1.90
13	4	0.20	27	0.68	57	0.99	41	1.58	18	2.09	8	0.56
14	8	0.35	12	0.30	6	0.10	21	0.81	0	0.00	5	0.35
15	0	0.00	6	0.15	5	0.08	3	0.12	0	0.00	0	0.00
16	0	0.00	0	0.00	0	0.00	6	0.23	0	0.00	0	0.00
17	0	0.00	1	0.03	0	0.00	2	0.08	2	0.23	1	0.07
18	1	0.05	0	0.00	0	0.00	4	0.25	0	0.00	0	0.00
19	0	0.00	0	0.00	0	0.00	2	0.08	0	0.00	0	0.00
20	0	0.00	0	0.00	0	0.00	2	0.08	0	0.00	0	0.00
Total	2073	100.0	3958	100.0	5772	100.0	2588	100.0	862	100.0	1421	100.0

- Z-N - (Zainab 1996); NEN - (Nor 1996); T-W (Teh 1996): K-K (Kademani 1996); PA (Parvathamma 1996); P-P (Panigrahi 1996) was expected since in English language there are only two one-letter words , i.e., 'a' and 'I' that are used quite frequently. In technical writing, the use of the word 'I' is not very common. In the six papers analysed for the study, we have not encountered the word 'I' even once. Hence, the percentage given above reflects only the occurrence of 'a'. At any rate, this low percentage does not follow Zipf's law.

However, the situation changes from two-letter words onwards. The percentage of occurrence varies from 15 to 20 in the case of two-letter words, which remains more or less constant around 18% in the case of three-letter words. After this in all cases the fall is gradual as well as uniform as can be seen from Fig. 1. The percentage of occurrence starts falling below 1 with thirteen-letter words. The number of words comprising 15 to 18 letters is negligibly small. Words comprising more than eighteen letters were not encountered. If we ignore the case of one-letter words, then the findings obey Zipf's law quite well.

As to the optimum sample size it may be observed from Fig. 1 that the decrease is most uniform in the case of the third chart where the sample size is reasonably large (5772 words). With lesser sample size certain deviations are observed. The fifth chart with the least sample size (862 words) also show more or less uniform decrease. Still, it is felt that for better results a sample size of about 5,000 words may be considered adequate.

Figure 1: Frequency Distribution of Word Lengths of Six Sample Articles



CONCLUSION

This study indicates that the LIS writings also follow the Zipf's law when only the textual part of the writing is considered omitting alpha-numeric and alpha-symbolic expressions, abbreviations, headings of illustrations, intra-textual references, words figuring within tables, keywords. However, just from this study it is not possible to generalise that the law will be applicable to all types of technical writings, for which separate studies need to be undertaken.

ACKNOWLEDGEMENT

The authors are profoundly thankful to Prof. Mashkuri Yaacob, Dean, Faculty of Computer Science and Information Technology, University of Malaya, for providing all facilities in the preparation of the paper.

REFERENCES

Kademani, B. S. and V. L. Kalyane. 1996. Outstandingly cited and most significant publications of R Chidambaram, a nuclear physicist. *Malaysian Journal of Library and Information Science* Vol. 1, no. 1: 21-36.

Nor Ehzan, N. 1996. The use of CD-ROM databases by Malaysian postgraduate students in Leeds. *Malaysian Journal of Library and Information Science* Vol. 1, no. 1: 37-55.

Panigrahi, C. and K. C. Panda. 1996. Reading interests and information sources of school going children: a case study of two English medium schools of Rourkela, India. *Malaysian Journal of Library and Information Science* Vol. 1, no. 1: 57-65.

Parvathamma, N.1996. The coverage of Indian literature in social science bibliographic databases on CD-ROM. *Malaysian Journal of Library and Information Science* Vol. 1, no. 1:85-92.

Teh, K. H. and S. F. Wong. 1996. Developing a CDS/ISIS-based online cataloguing and information retrieval interfaces for use in small libraries. *Malaysian Journal of Library and Information Science* Vol. 1, no. 1: 1-20.

Wyllys, R. E. 1981. Empirical and theoretical bases of Zipf's law. *Library Trends* Vol. 30, no.1: 53

Zainab, A. N. and Nor Eliza. M. Z.1996. Introducing MAKLUM the general reference expert adviser developed for a university library. *Malaysian Journal of Library and Information Science* Vol. 1, no. 1: 93-107

Zipf, G. K.1935. *The psycho-biology of language: an introduction to dynamic philology*. Houghton Mifflin, Boston; Riverside Press, Cambridge. p. v.

Zipf, G. K. 1949. *Human behavior and the principle of least effort: an introduction to human ecology*. Reading, Mass.: Addison-Wesley Press.