自然言語処理 46 − 2
(1984. 12. 7)

On Word Occurrence in Scientific and Technological Texts

Fumihiro Matsuo

Computer Center, Kyushu University 91
Hakozaki, Fukuoka 812, Japan

ABSTRACT

The rank-frequency characteristics of the words from the abstracts in a considerably large collection of INSPEC tapes are examined. The word probabilities are divided into three groups according to the word rank. In case of high frequency words the probability p(r) of the word with rank r is inversely proportional to r, i.e. follows Zipf's law. In case of medium frequency words, $p(r) \propto 1/r^{a \ln r - b}$, where a and b are constants. The probabilities of low frequency words are in inverse proportion to the square of r. The boundary between the high and the medium frequency words lies near rank 300, and the boundary between the medium and the low frequency words near rank 1000. This paper also discusses the number of low frequency words occurring n times.

1. INTRODUCTION

A word type is a character string that differs from other character strings [6]. By a word token we mean any word type occurring in a text. The rth most frequent word type in a text is said to be the word type with rank r and the number of occurrences of that word type is called its frequency. The rank-frequency characteristics of word types are of direct importance not only to linguistics but also to computer science, since the skewed distribution of word occurrence can be used to increase efficiency in textual data processing such as information retrieval and natural language processing. So far, this characteristics have been studied by Estoup [4], Bradford [3], Zipf [8], Mandelbrot [5] and others. However, their results do not agree well with the characteristics concerning large sets of word tokens from the abstracts in INSPEC tapes [1]. This paper discusses the rank-frequency characteristics of word types of the abstracts in INSPEC tapes issued from 1973 to 1982.

2. WORDS FROM ABSTRACTS IN INSPEC TAPES

To distinguish characteristic differences arisen from research fields, we will classify abstracts in INSPEC tapes issued from 1973 to 1982 into three sets, A, B and C, according to the sectional classification codes of the document records. The fields of three sets are as follows:

[A] Physics;
[B] Electrical Engineering and Electronics;
[C] Control Engineering and Computer Science.

Three sets are not mutually disjoint. About 20 % of abstracts belong to two or three sets. The number of word types and word tokens in these sets are shown in Table 1.

We select space and all special symbols as the delimiter for word tokens in the abstracts, therefore a word type is a alpha-numeric character string.

The log-log plots of word frequency versus word rank in Fig. 1 illustrate that there is little difference in the characteristics among three sets. The cumulative probabilities in Fig. 2 exhibit the strong resemblance more distinctly. The solid curved line in Fig. 2 shows the relation

$$\text{probability} = 0.092 \,/\, \text{rank}, \qquad (1)$$

which is Zipf's law [4,8] discussed below.

To see the effect of the sample size on the characteristics, we produce 14 sets of abstracts, A_1, A_2, ..., A_{14}, out of A. Each A_i, $1<i\leq 14$, is included in A_{i-1}, and A_i is ap-

Table 1. Word counts for three sets

	Set A	Set B	Set C
No. Word Tokens	87,354,577	37,606,323	23,391,467
No. Word Types	274,185	154,231	127,836

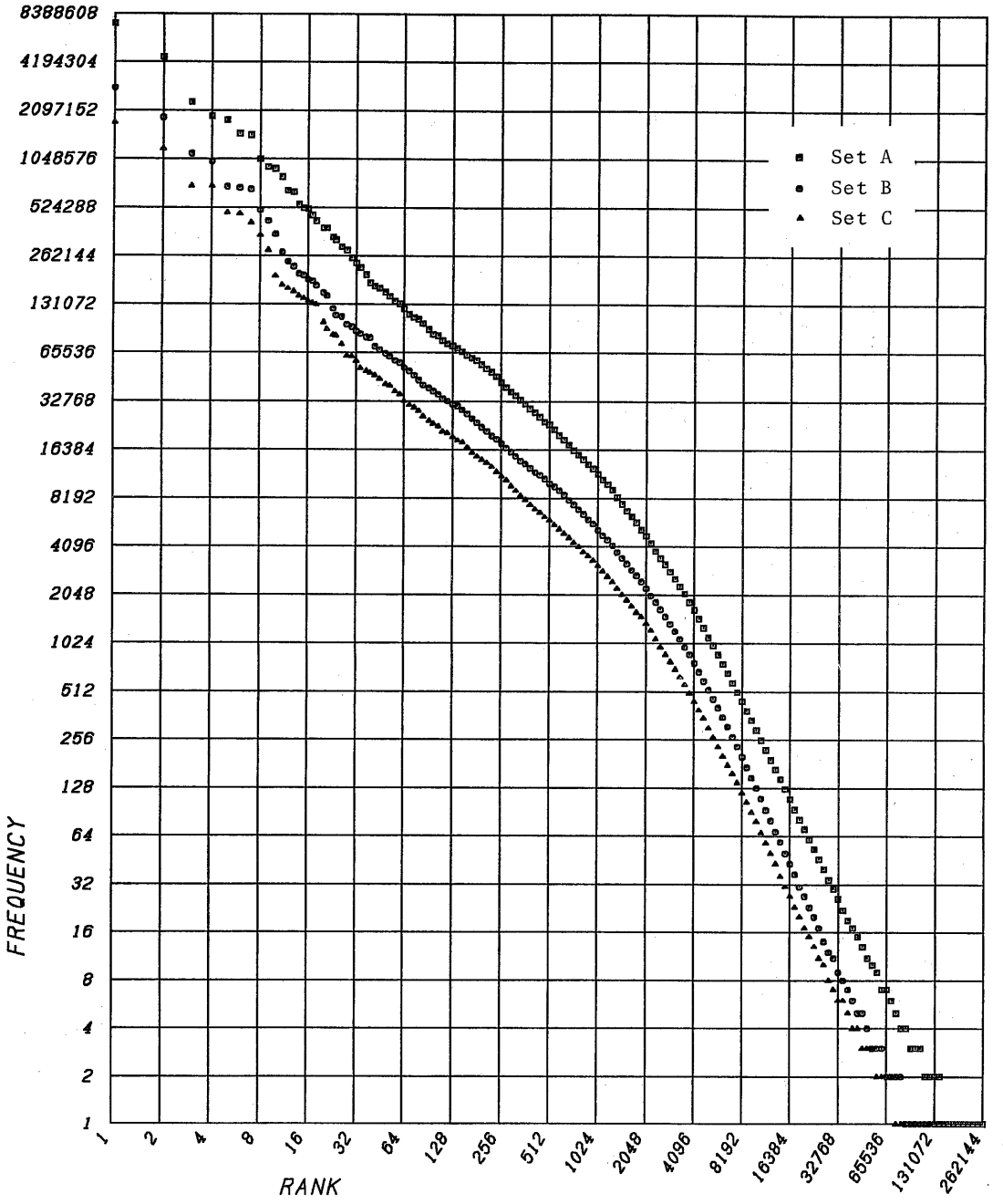

Fig. 1. Rank-frequency characteristics.

proximately half the size of A_{i-1}. The largest set A_1 contains all abstracts in the tapes from 1973 to 1980. Fig. 3 exhibits the characteristics of these sets.

3. LOW FREQUENCY WORD PROBABILITIES

Let f(r) be the frequency of the word type with rank r. Zipf's first law [2] asserts that:

$$r \cdot f(r) = k, \tag{2}$$

where k is a constant for any particular text.

Booth's modification of Zipf's second law [2] states that a word type will occur once if

$$1 \leq f(r) < 2. \tag{3}$$

From (2) and (3), we have

$$k/2 < r \leq k. \tag{4}$$

Thus, the number of word types in the text, which we denote by D, is

$$D = r_{max} = k. \tag{5}$$

For the number of word tokens, denoted by T, we obtain

$$T = \Sigma_{r=1}^{D} f(r) = k \Sigma_{r=1}^{D} 1/r \simeq k(\ln D + e), \tag{6}$$

where e = 0.5772156649... is Euler's constant.

From (5) and (6), we can immediately see that

$$T \simeq D(\ln D + e). \tag{7}$$

However, (7) disagrees with Fig. 4, since it says that

$$T \propto D^2. \tag{8}$$

Circles in Fig. 4 denote the word count of A_1, A_2, ..., A_7 from set A. For sets B and C, (8) holds similarly. Using p(r) for the probability of the word type with rank r, we can express Zipf's first law as follows:

$$p(r) = c/r, \tag{9}$$

where c is a constant for any particular text. In (9), c must be

$$c = k/T \simeq 1/(\ln D + e), \tag{10}$$

from (6). In many literatures, however, c is considered to be a constant independent of T (or D) and nearly equal to 0.1 [8]. If so, from (6) we have

$$D \simeq \exp[1/c - e] = \text{constant}. \tag{11}$$

It is contradictory to (8).

To resolve this problem, let

$$p(r) = c_1/r^2, \tag{12}$$

for word types of low frequency of occurrence, here c_1 is a constant independent of T. In

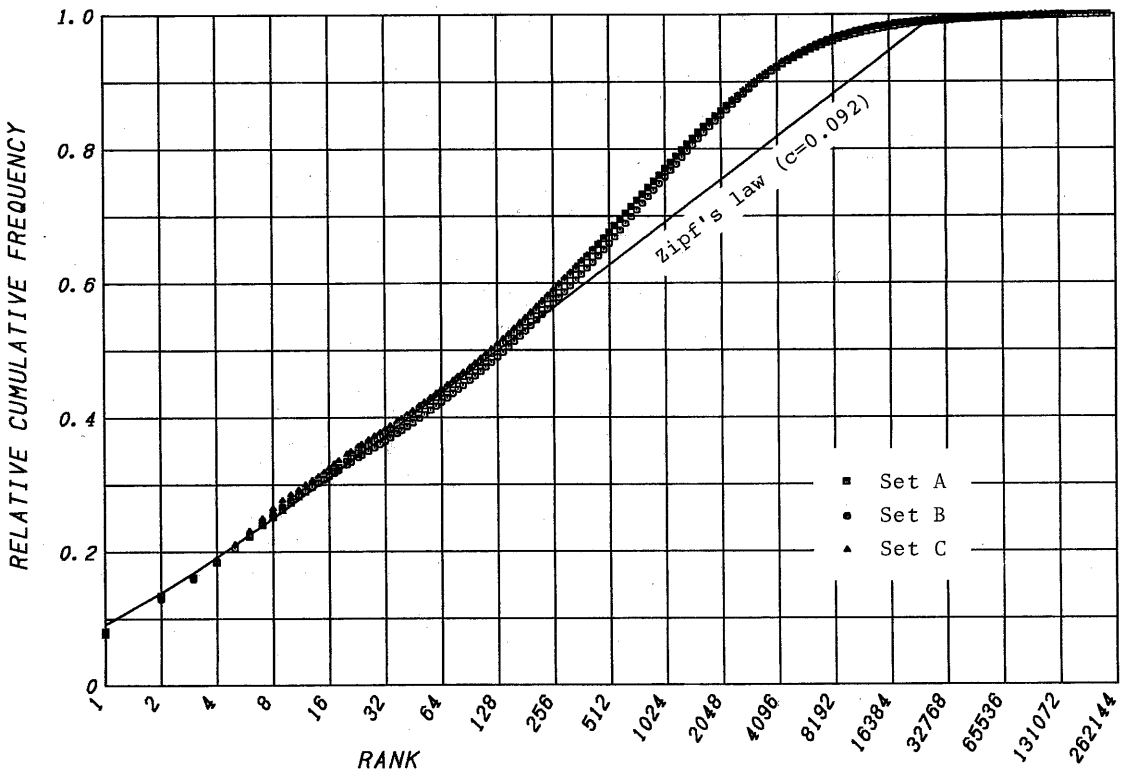

Fig. 2. Relative cumulative frequencies.

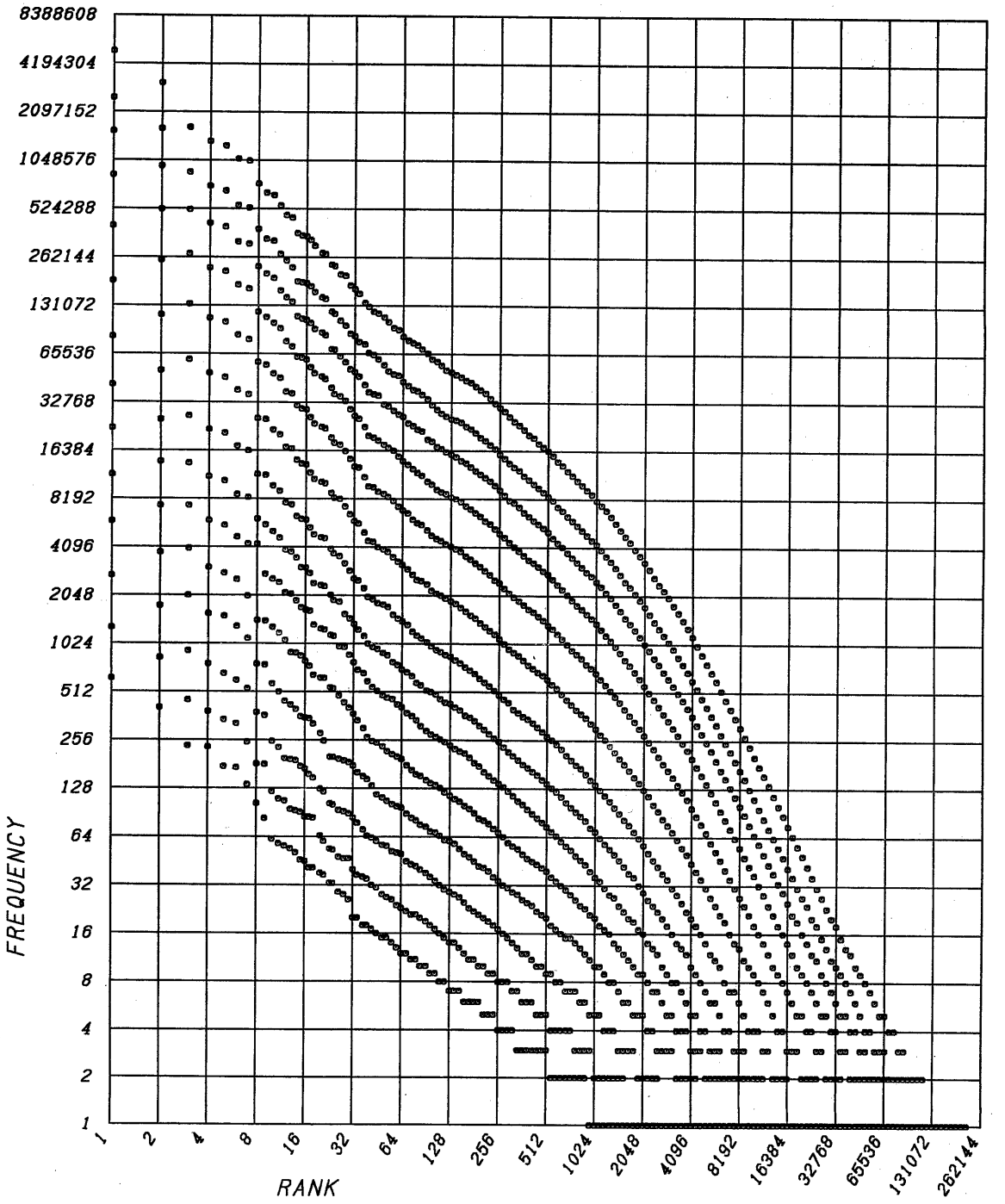

Fig. 3. Rank-frequency characteristics.

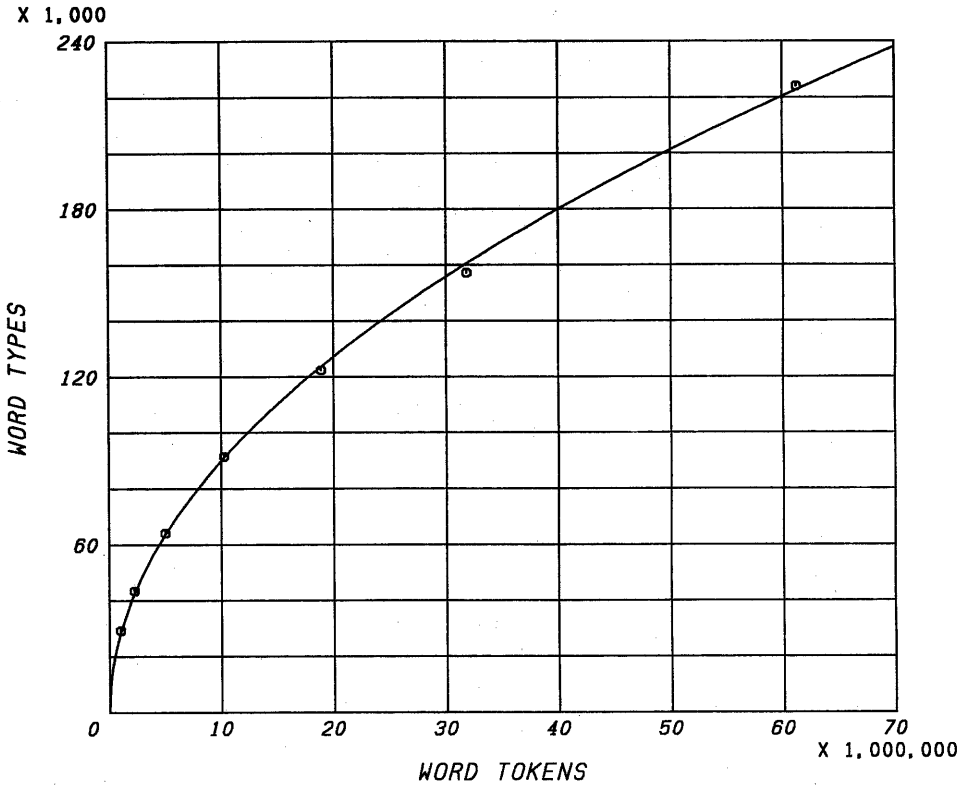

Fig. 4. Number of word tokens (T) versus number of word types (D).

fact, Tc_1/r^2 seems to approximate to the frequency of a word type with a rank lower than about 3000 in Fig. 1. From (3) and (12), we have

$$D = \sqrt{c_1 T}, \tag{13}$$

in much the same way as we have derived (5) from (3) and (4). Equation (13) agrees well with Fig. 2. The plots about low frequency word types in Fig. 1 also appear to agree approximately with (12). It is another advantage from (12) that $\Sigma_{r=1}^{\infty} p(r)$ could be equal to 1 whatever the probabilities of word types of high or medium frequency might be, since $\Sigma_{r=1}^{\infty} 1/r^2 = \pi^2/6$.

4. HIGH FREQUENCY WORD PROBABILITIES

From (1) and Fig. 3, the probability p(r) of the word type of high frequency whose rank is r seems to be

$$p(r) = c_h/r, \tag{14}$$

where c_h is a constant independent of T.

It is natural to consider that the probabilities of high frequency word types obey Zipf's law, for most of high frequency word types are common words. In Fig. 1, (14) seems to hold for the word type with a rank higher than about 100.

5. MEDIUM FREQUENCY WORD PROBABILITIES

The author can not account for the characteristics of medium frequency word types of which word ranks are between about 100 and about 3000. If we make an approximation with parabolas for the log-log plots in Fig. 1, we have

$$p(r) = c_m/r^{a \ln r - b}, \tag{15}$$

where c_m, a and b are constants.

6. NUMBER OF WORD TYPES OCCURRING N TIMES

The number of word types occurring n times in some text is more important than their probabilities of occurrence for low frequency word

Table 2. Values of I_n/D

n	Abstracts in INSPEC-tapes Set A	Set B	Set C	Matsuo Eq.(23)	Booth Eq.(16)
1	.460076	.453307	.440893	.422650	.5
2	.135999	.134616	.137833	.130137	.166667
3	.069391	.067328	.071420	.069249	.083334
4	.042471	.041879	.043618	.044631	.05
5	.030071	.028976	.030453	.031822	.033333
6	.022394	.021922	.022513	.024161	.023810
7	.017313	.017182	.017515	.019151	.017857
8	.014049	.013421	.013924	.015663	.013889
9	.011383	.011197	.011804	.013120	.011111
10	.009515	.009881	.010521	.011198	.009090
20	.003370	.003294	.003395	.003954	.002381
30	.001765	.001822	.001721	.002152	.001075
40	.001218	.001193	.001306	.001398	.000610
50	.000821	.000901	.000892	.001000	.000392
100	.000281	.000305	.000313	.000354	.000099

types. Let I_n be this number in some set of sentences. Booth [2] has derived

$$I_n/D = 1/n(n-1) \qquad (16)$$

from (9) and the generalization of (3), which says that a word types will occur n times if

$$n \leq Tp(r) < n+1. \qquad (17)$$

Zipf's original form corresponding to (17) is

$$n-1/2 \leq Tp(r) < n+1/2. \qquad (18)$$

From (9) and (18), Zipf has obtained

$$I_n/D = 1/2(n^2-1/4). \qquad (19)$$

Equation (19), however, is out of accord with the observed facts, so that Booth has modified (18). Although (16) agrees nicely the observations for relatively small sentence sets, it could not be applicable to the large set in which (9) does not hold. Furthermore, (17) is unnatural and tricy in comparison with (18).

Assuming that (12) and (18) hold instead, we now follow the same argument that has deduced (16) or (19). Equations (12) and (18) lead immediately to

$$\sqrt{c_1T/(n+1/2)} < r \leq \sqrt{c_1T/(n-1/2)}\ . \qquad (20)$$

Hence,

$$I_n = \sqrt{c_1T/(n-1/2)} - \sqrt{c_1T(n-1/2)}$$
$$= \sqrt{2c_1T}(1/\sqrt{2n-1} - 1/\sqrt{2n+1}\ ). \qquad (21)$$

In the same manner as we have obtained (13), we have

$$D = \sqrt{2c_1T}\ , \qquad (22)$$

so that

$$I_n/D = 1/\sqrt{2n-1} - 1/\sqrt{2n+1}\ . \qquad (23)$$

The values of I_n/D for INSPEC tapes issued from 1973 to 1982 and the predicted values calculated from (23) and (16) are shown in Table 2. Closed agreement between the observed and the calculated values of (23) is obtained. On the other hand, (16) become out of accord with the observed values when n become more than 20.

7. CONCLUSION

The probability of word types with rank r p(r) is summarized as follows:

$$p(r) = \begin{cases} c_h/r & \text{for high frequency word types;} \\ c_m/r^{a \ln r - b} & \text{for medium frequency word types;} \\ c_l/r^2 & \text{for low frequency word types,} \end{cases}$$

where c_h, c_m, c_l, a and b are constants. In Fig. 1, the boundary between the high and the medium frequency words appears to lie near rank 300, and the boundary between the medium and the low frequency words near rank 1000.

Equation (23) gives the number of low frequency word types occurring n times.

The relation between the number of word types D and the number of word tokens T is given by (22).

REFERENCES

1. Aitchson, T. M., Martin, M. D., and Smith, J. R. : Developments towards a Computer-Based Information Service in Physics, Electrotechnology and control, Inform. Stor. Retr., **4**, 2, 1968, 177-186.
2. Booth, A. D. : A "Law" of Occurrences for Words of Low Frequency, Inform. Control, **10**, 4, 1967, 386-393.
3. Bradford, S. C. : Documentation, Crosby Lockwood, London, 1948.
4. Estoup, J. B.: Gammes stenographiques, 4th Ed., Gauthier-Villars, Paris, 1916.
5. Mandelbrot, B.: Théorie mathématique de la loi d'Estoup-Zipf, Institut de Statistique de l'Université, Paris, 1957.

6. Schwartz, E. S.: A Dictionary for Minimum Redundancy Encoding, J. ACM, **10**, 4, 1963, 413-439.
7. Shannon, C.E.: Prediction and Entropy of Printed English, Bell Syst. Tech. J., **30**, 1951, 50-64.
8. Zipf, G. K.: Human Behaviour and the Principle of Least Effort, Addison-Wesley, Cambridge, Mass., 1949.