NTIS HC $4.25

Dupe 1 AD-743 894

Technical Report TR-187 March 1972

METHODS OF COMPUTING VOCABULARY SIZE FOR THE TWO-PARAMETER RANK DISTRIBUTION

H. P. Edmundson
G. Fostel
I. Tung
W. Underwood

(NASA-CR-129806) METHODS OF COMPUTING VOCABULARY SIZE FOR H.P. Edmundson, et al (Maryland Univ.) Mar. 1972 43 p N73-13197
CSCL 09B Unclas
G3/08 50321

UNIVERSITY OF MARYLAND
COMPUTER SCIENCE CENTER
COLLEGE PARK, MARYLAND

Technical Report TR-187 March 1972

METHODS OF COMPUTING VOCABULARY SIZE FOR THE TWO-PARAMETER RANK DISTRIBUTION

H. P. Edmundson
G. Fostel
I. Tung
W. Underwood

This research was sponsored in part by the Office of Naval Research under contract N00014-67-A-0239-0004, NR 049-261; in part by the National Aeronautics and Space Administration under grant NGL-21-002-008; and in part by the National Science Foundation Science Development Program under grant GU-2061. Reproduction in whole or in part is permitted for any purpose of the United States Government.

ABSTRACT

This paper describes a summation method for computing the vocabulary size for given parameter values in the 1- and 2-parameter rank distributions. Two methods of determining the asymptotes for the family of 2-parameter rank-distribution curves are also described. Tables are computed and graphs are drawn relating pairs of parameter values to the vocabulary size. The partial product formula for the Riemann zeta function is investigated as an approximation to the partial sum formula for the Riemann zeta function. An error bound is established that indicates that the partial product should not be used to approximate the partial sum in calculating the vocabulary size for the 2-parameter rank distribution.

TABLE OF CONTENTS

TABLE OF FIGURES

1. INTRODUCTION

1.1 Background

This paper is a continuation of the research reported by Edmundson [1972]. That paper included a historical summary of the controversy concerning the rank hypothesis. The rank hypothesis is based on the observation of the American philologist G. K. Zipf [1935, 1949] that the relative frequency f_r of a word type of rank r is approximately a constant c times the reciprocal of its rank r.

The model corresponding to Zipf's observation is that the probability of the occurrence of a word type of rank r is the product of a parameter c and the reciprocal of the rank of that word type. Hence the rank distribution formulated by Zipf has the density function

$$p_r = cr^{-1} \qquad c > 0$$

for r = 1,...,v where v is the theoretical vocabulary size.

The American linguist M. Joos [1936] observed that empirical data is not adequately fitted by Zipf's rank distribution, especially at the extremes where the rank is either very high or very low. Joos introduced a second parameter b as the exponent of the rank r. Thus the rank distribution formulated by Joos has the density function

$$p_r = cr^{-b} \qquad b \geq 1,\ c > 0$$

for r = 1,...,v. Let the cumulative distribution function by denoted by

$$F_r = \sum_{k=1}^{r} p_k$$

Since $F_v = 1$, it follows that

$$1 = c \sum_{r=1}^{v} r^{-b}$$

Note that the above equation is of the form $\phi(v,b,c) = 0$ and hence implies that v is a function of b and c.

1.2 Purpose

The purpose of this paper is to present several methods for computing the vocabulary size v, given values of the parameters b and c in the 2-parameter rank distribution. The linguistic motivation for this mathematical research is to provide linguists with a parameterized family of curves that will permit them to do the following:

(1) given any two of the three quantities v, b, and c, find the third.

(2) given any one of the three quantities v, b, and c, find the set of all possible pairs of the remaining two.

Of these possibilities perhaps the most linguistically interesting are the following:

(a) assuming a given vocabulary size v, find a pair of parameter values b and c that are linguistically satisfactory.

(b) assuming fixed values of the parameters b and c, compute the resulting vocabulary size v.

(c) assuming given values of the vocabulary size v and the parameter c, compute the resulting value of the parameter b.

(d) assuming given values of the vocabulary size v and the parameter b, compute the resulting value of the parameter c.

1.3 Scope

The remainder of this paper presents several methods of computing the vocabulary size v, given values of the parameters b and c. Section 2 discusses a direct summation method of calculating v for the 2-parameter rank distribution. Section 3 discusses a method for computing vocabulary size using a finite product involving primes. Section 4 presents two methods for determining asymptotes to the rank-distribution curves. This section contains, as the major result of the paper, a graph of the parameterized family of curves together with their asymptotes.

1.4 Results

Tables have been computed and graphs have been drawn for v satisfying the equation

$$\phi(v,b,c) = c \sum_{r=1}^{v} r^{-b} - 1 = 0$$

for certain values of the parameter b in the interval 0.90 to 1.14 and the parameter c in the interval 0.05 to 0.15. Asymptotes to the curves representing v vs. b have been determined for each value of c. A good error bound has been derived for the partial product formula for the Riemann zeta function as an approximation to the partial sum formula for the Riemann zeta function.

More extensive results covering approximation formulas for the vocabulary size for the 1-, 2-, and 3-parameter rank distributions are given in Edmundson et al. [1972].

2. SUMMATION METHOD

2.1 Program for the Summation Method

The most straight-forward way to solve for v, given b and c in the 2-parameter rank distribution where

$$1 = c \sum_{r=1}^{v} r^{-b}$$

is to add a sufficient number of terms until the sum multiplied by c first exceeds 1. The number v* of terms summed will be regarded as an approximation of the exact value v.

The values initially proposed for consideration were b = 0.90, 0.95, 0.99(.01)1.20 and c = 0.05(.01)0.15. Later, it was decided advisable to look at the fine structure in the range c = 0.065(0.001)0.100 when b = 1.00. However, v was not computed for all proposed values of b and c since either (1) the computation time is known to be excessive or (2) no such value of v exists. (See Section 4 on asymptotes.)

An ALGOL program for the summation method is presented in Fig. 1. In this program b and c are the parameters of the implicit function ϕ, r is the iterated variable, t is the reciprocal of r to the power b, log(v) is the common logarithm of v, s is the double-precision sum of the terms t, and q is the product of c and s. A value of b is read and c is initialized to 0.15. The program iterates through the loop, increasing r and computing q, until q exceeds 1.0. The value of r after q exceeds 1.0 is regarded as the value of v with respect to the parameters b and c. The common logarithm of v is computed to facilitate graphing the relationship of b, c, and v. The values of c, v, $\log_{10} v$, t, and q are then outputted.

The addition of terms **t** to form s causes some complication in this program. The UNIVAC 1108 computer used for these computations allows precision of up to 9 significant decimal digits. As r increases to the order 10^7, t is of the order 10^{-7}. When s becomes greater than 10, adding numbers of the order 10^{-7} to s would be meaningless on this computer. Therefore, s and q have been chosen to be double-precision variables, allowing 18 significant decimal digits for each. Double precision was not used for other variables to save computation time in arithmetic operations, especially for exponentiation.

```
begin comment summation method;
     real b,c,r,t,v;
     real procedure log(x);
     real x;
     log:=0.43429448*ln(x);
     comment use double precision for s and q;
     real 2 s,q;
     format val(4R15.8,R25.18,A1.0);
     read (b);
     s:=0.0&&0;
     r:=0.0;
     for c:=0.15 step -.01 until 0.05 do
     begin
     loop:  r:=r+1;
            t:=r**(-b);
            s:=s+t;
            q:=c*s;
            if q<1.0&&0 then go to loop;
            v:=r;
     write (val,c,v,log(v),t,q)
     end
end
```

Figure 1. ALGOL Program for Summation Method.

Instead of computing the sum for each value of c, considerable computer time is saved by the following procedure. For fixed b the c's are arranged in decreasing order. When the sum (multiplied by c) first exceeds 1.0, the calculation for the next smaller c may be started by using the current partial sum instead of restarting from its first term.

The computation time for each term in the sum has been found to be approximately 80 microseconds. The computation time for v is directly proportional to the size of v with a proportionality constant of 80 microseconds. For example, the value v = 898,515 calculated for b = 1.00 and c = 0.07 took approximately 70 seconds to compute on the UNIVAC 1108.

2.2 Sample Output and Graph

The sample output in the case b = 1.0 is tabulated in Fig. 2 and its graph is plotted in Fig. 3. The outputted values v, t, and q are respectively those values of r, t, and q immediately after q has exceeded 1.0. Therefore v is the number of terms in the sum and the variable t is the last term in the sum, that is

$$t = v^{-b}$$

The table does not contain values of c less than 0.07 because the run was stopped after 75 seconds of execution.

c	v	$\log_{10} v$	t	q
0.15	441	2.6444385E+00	2.2675737E-03	1.00010907172776739D+000
0.14	710	2.8512583E+00	1.4084507E-03	1.00004584500300125D+000
0.13	1230	3.0899051E+00	8.1300813E-04	1.00001089167823920D+000
0.12	2336	3.3684728E+00	4.2808219E-04	1.00003499380429624D+000
0.11	4983	3.6974908E+00	2.0068232E-04	1.00002136503591327D+000
0.10	12367	4.0922642E+00	8.0860354E-05	1.00000429331210616D+000
0.09	37568	4.5748180E+00	2.6618399E-05	1.00000231334124586D+000
0.08	150661	5.1780007E+00	6.6374178E-06	1.00000052021645891D+000
0.07	898515	5.9535252E+00	1.1129475E-06	1.00000004305938254D+000

Figure 2. Computer Results for Summation Method for b = 1.00.

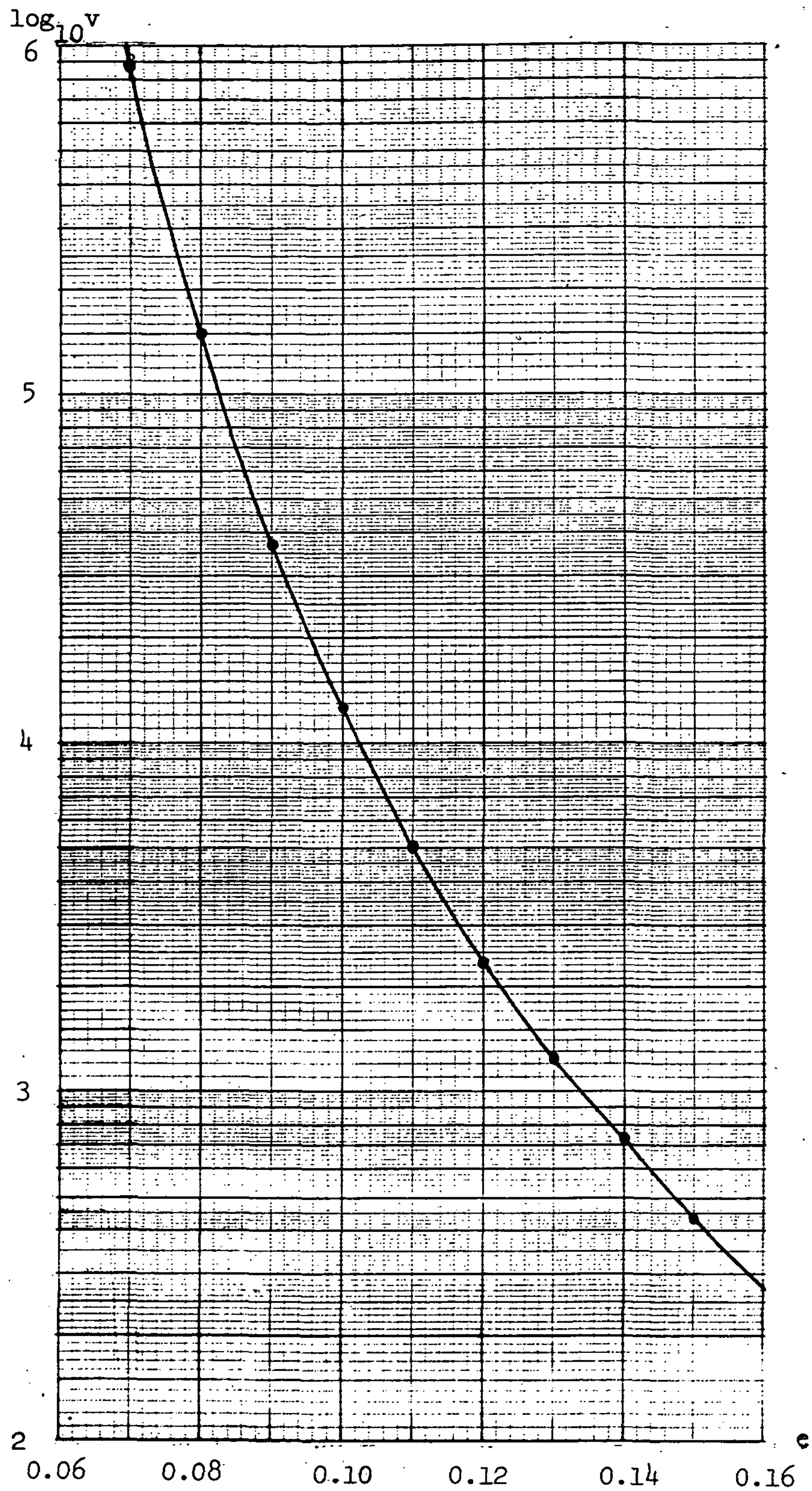

Figure 3. Curve Relating $\log_{10}v$ and c for b = 1.0.

2.3 Tables and Graph of the Results

The table of $\log_{10}v$ for certain values of b and c may be found in Fig. 4. More comprehensive tables are given in Appendix A for c at intervals of 0.01. For b = 1.0 the fine structure is given in Appendix B for c at intervals of 0.001.

The family of curves relating the values $\log_{10}v$, b, and c is presented in Fig. 5.

$\log_{10} v$

b \ c	0.05	0.06	0.07	0.08	0.09	0.10	0.11	0.12	0.13	0.14	0.15
0.90	4.6879	4.1659	3.7503	3.4104	3.1261	2.8848	2.6767	2.4955	2.3336	2.1931	2.0682
0.95	*	5.1279	4.5351	4.0624	3.6760	3.3541	3.0817	2.8476	2.6444	2.4669	2.3096
0.99	*	*	5.5796	4.8921	4.3498	3.9110	3.5487	3.2445	2.9854	2.7619	2.5670
1.00	*	*	5.9535	5.1780	4.5748	4.0923	3.6975	3.3685	3.0900	2.8513	2.6444
1.01	*	*	*	5.5133	4.8338	4.2978	3.8640	3.5059	3.2052	2.9489	2.7284
1.02	*	*	*	5.9144	5.1365	4.5336	4.0525	3.6595	3.3326	3.0561	2.8195
1.03	*	*	*	6.4069	5.4971	4.8083	4.2682	3.8329	3.4746	3.1744	2.9191
1.04	*	*	*	*	5.9375	5.1342	4.5185	4.0308	3.6346	3.3060	3.0294
1.05	*	*	*	*	*	5.5302	4.8142	4.2596	3.8165	3.4539	3.1514
1.06	---	*	*	*	*	6.0269	5.1717	4.5289	4.026	3.6216	3.2882
1.07	---	---	*	*	*	*	5.6170	4.8525	4.2720	3.8142	3.4431
1.08	---	---	---	*	*	*	6.1964	5.2531	4.5658	4.0389	3.6202
1.09	---	---	---	---	*	*	*	5.7691	4.9269	4.3062	3.8261
1.10	---	---	---	---	---	*	*	*	5.3878	4.6324	4.0696
1.11	---	---	---	---	---	---	*	*	6.0101	5.0449	4.3648
1.12	---	---	---	---	---	---	---	*	*	5.5941	4.7345
1.13	---	---	---	---	---	---	---	---	*	*	5.2202
1.14	---	---	---	---	---	---	---	---	*	*	5.9081

--- $\log_{10} v$ is undefined

* $\log_{10} v$ was not calculated because of excessive computation time

Figure 4. Table Relating $\log_{10} v$, b, and c.

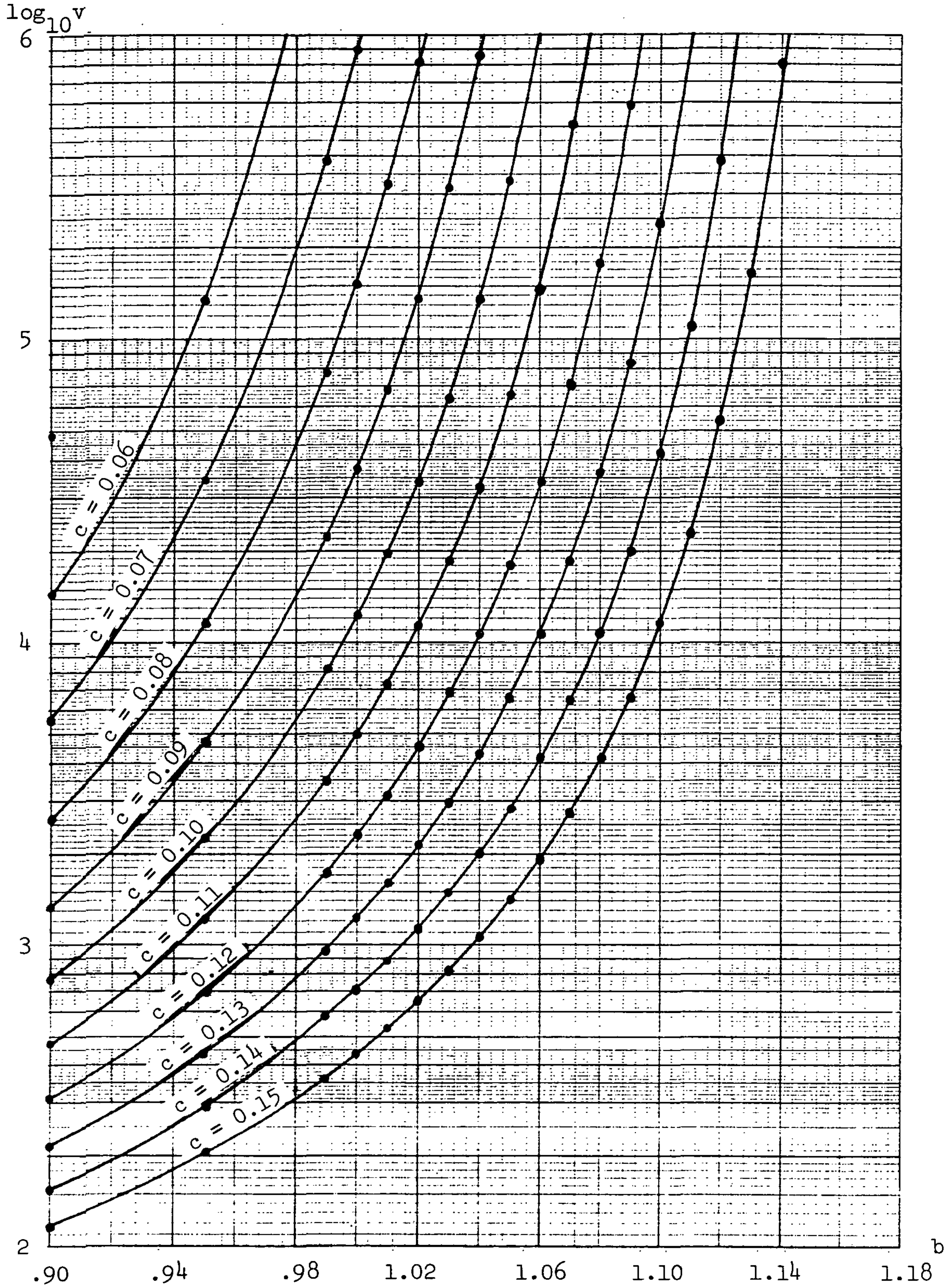

Figure 5. Family of Curves Relating $\log_{10}v$, b, and c.

3. THE TWO-PARAMETER RANK DISTRIBUTION AND THE RIEMANN ZETA FUNCTION

3.1 The Partial Sum and Partial Product Formulas for the Riemann Zeta Function

Since values of v greater than 10^6 could not be computed within reasonable computation times (as indicated in Fig. 4), another method for computing the vocabulary size must be found. Note that the function

$$f(v,b) = \sum_{r=1}^{v} r^{-b}$$

derived from the 2-parameter rank distribution is actually the partial sum of the Riemann zeta function defined by

$$\zeta(b) = \sum_{r=1}^{\infty} r^{-b} \qquad b > 1$$

One of the most important theorems concerning the Riemann zeta function is

$$\zeta(b) = \prod_{k=1}^{\infty} (1 - p_k^{-b})^{-1} \qquad b > 1$$

where p_k is the k-th prime number (see Apostol [1957, p. 389]; Jahnke, Emde, and Lösch [1960, p. 37]). Let

$$S_n = \sum_{r=1}^{n} r^{-b}$$

denote the n-th partial sum of the Riemann zeta function and let

$$P_n = \prod_{k=1}^{n} (1 - p_k^{-b})^{-1}$$

denote the n-th partial product of the Riemann zeta function. Because of the sparseness of prime numbers, consideration has been given to approximating the partial sum by the partial product.

For this approximation it is desirable to derive a bound on the difference between the partial product and the partial sum. Since

$$(1 - x)^{-1} = 1 + x + x^2 + x^3 + \dots$$

for $|x| < 1$, the partial product P_n may be written as

$$\prod_{k=1}^{n} (1 - p_k^{-b})^{-1} = \prod_{k=1}^{n} (1 + p_k^{-b} + p_k^{-2b} + p_k^{-3b} + \cdots)$$

$$= (1 + p_1^{-b} + p_1^{-2b} + \cdots) \cdots (1 + p_n^{-b} + p_n^{-2b} + \cdots)$$

After multiplication all terms are of the form

$$p_1^{-e_1 b}\, p_2^{-e_2 b} \cdots p_n^{-e_n b}$$

where the e_i are non-negative integers for $i = 1,\ldots,n$. Therefore the partial product may be expressed as the sum of all such terms

$$\prod_{k=1}^{n} (1 - p_k^{-b})^{-1} = \sum_{e_1=0}^{\infty} \sum_{e_2=0}^{\infty} \cdots \sum_{e_n=0}^{\infty} p_1^{-e_1 b}\, p_2^{-e_2 b} \cdots p_n^{-e_n b}$$

Since for every prime p_n every positive integer $r \leq p_n$ can be expressed as

$$r = p_1^{e_1}\, p_2^{e_2} \cdots p_n^{e_n}$$

for some integers $e_i \geq 0$ where $i = 1,\ldots,n$, it follows that

$$\sum_{r=1}^{p_n} r^{-b} \leq \sum_{e_1=0}^{\infty} \sum_{e_2=0}^{\infty} \cdots \sum_{e_n=0}^{\infty} p_1^{-e_1 b}\, p_2^{-e_2 b} \cdots p_n^{-e_n b}$$

Since by definition

$$\zeta(b) - P_n = \sum_{r=1}^{\infty} r^{-b} - \prod_{k=1}^{n} (1 - p_k^{-b})^{-1}$$

it follows that

$$\zeta(b) - P_n \leq \sum_{r=1}^{\infty} r^{-b} - \sum_{r=1}^{p_n} r^{-b} = \sum_{r=p_n+1}^{\infty} r^{-b}$$

Thus

$$0 \le \zeta(b) - P_n \le \sum_{r=p_n+1}^{\infty} r^{-b}$$

Multiplying by -1 and adding the term $\sum_{r=p_n+1}^{\infty} r^{-b}$ throughout, it follows that

$$0 \le P_n - S_{p_n} \le \sum_{r=p_n+1}^{\infty} r^{-b}$$

Since

$$\sum_{r=p_n+1}^{\infty} r^{-b} \le \int_{p_n}^{\infty} x^{-b} dx = \frac{p_n^{1-b}}{b-1} \qquad b > 1$$

the bounds for the difference between the partial product and the partial sum of the Riemann zeta function may be given by

$$(3.1) \qquad 0 \le P_n - S_{p_n} \le \frac{p_n^{1-b}}{b-1}$$

For example, if $b = 2.0$ and $p_n \doteq 10^6$, then (3.1) gives an error bound of the order 10^{-6}. Since $S_{p_n} \ge 1$, the relative error bound is

$$\left| \frac{P_n - S_{p_n}}{S_{p_n}} \right| \le \frac{p_n^{1-b}}{S_{p_n}} \le 10^{-6}$$

Hence for values of b and p_n of these magnitudes or larger, the partial product P_n is a good approximation to the partial sum S_{p_n}.

On the other hand, if $b = 1.1$ and $p_n \doteq 10^6$, then (3.1) gives an error bound of approximately 2.5. Since $S_{p_n} \leq \zeta(b) = 10.584$, the relative error bound is approximately 1/4. Hence, for values of b close to 1, the upper bound is too loose to approximate the difference. To estimate this difference better, the values of P_n and S_{p_n} will be calculated directly.

3.2 Comparison of the Partial Sum and Partial Product

This section is devoted to the calculation of the partial sum S_n and the partial product P_n for b in the interval (1.0, 1.2]. One problem with the latter calculation is the need to generate primes. The prime number generator presented by Chartres [1967] is used here to generate prime numbers less than 60,000. It has been rewritten in FORTRAN and appears in Appendix D. With these prime numbers the partial product P_n may be calculated by multiplying factor by factor. Graphs comparing the partial sums S_n and partial products P_n for b = 1.0, 1.1, and 1.2 are shown in Figs. 6, 7, and 8, respectively. Tables for these data points are given in Appendix C. For b = 1.0 in Fig. 6, the graphs of P_n and S_n appear to diverge and then converge. For b = 1.2 in Fig. 8, the partial product is a relatively good approximation to the partial sum. However, the main concern in this research is for b in the interval (1.0, 1.2]; even though the vocabulary size is undefined for b close to 1.2 in the chosen range of the parameter c, as is explained in Section 4 below.

It should be recalled that this research is concerned with finding the number of terms summed (that is, the vocabulary size), rather than the sum itself. Despite the fact that there may be a small difference between the partial sum S_n and the partial product P_n, there may still be a great difference between the number of terms summed in the partial sum and the largest prime p_n in the partial product. For example, in the case $b = 1.1$, if $p_n = 59{,}887$, then $P_n \doteq 8.78$ and $S_{p_n} \doteq 7.25$, giving a difference of only 1.53. However, P_n exceeds the value 7.25 when $p = 1{,}009$, while S_{p_n} exceeds this value when $p_n = 59{,}887$. Therefore, the partial product should not be used to approximate the partial sum in calculating the vocabulary size for the 2-parameter rank distribution.

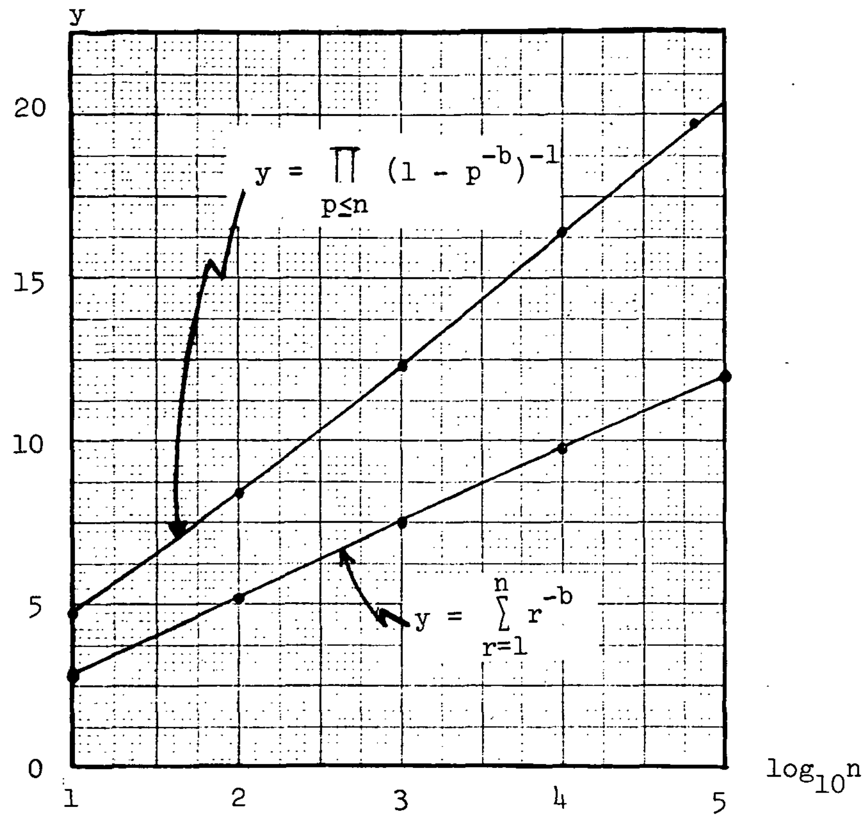

Figure 6. Comparison of Partial Sum and Partial Product for b = 1.0.

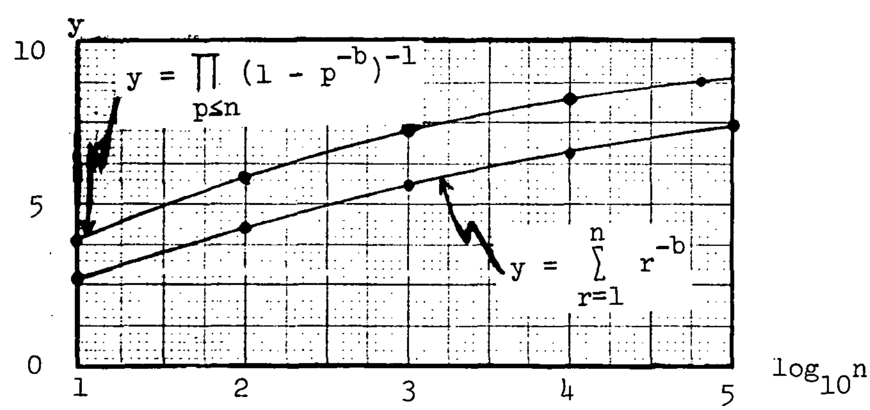

Figure 7. Comparison of Partial Sum and Partial Product for b = 1.1.

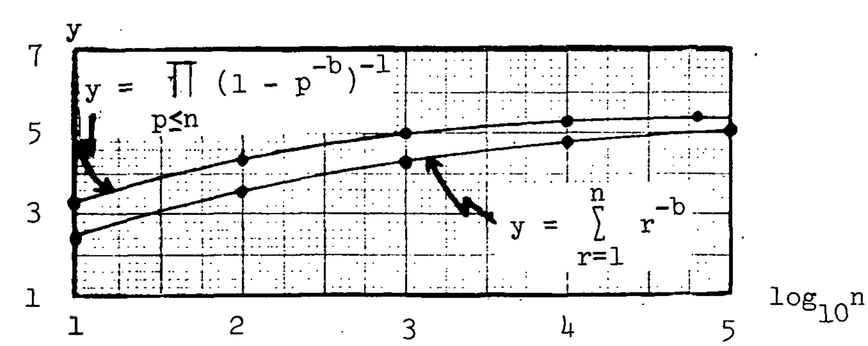

Figure 8. Comparison of Partial Sum and Partial Product for b = 1.2.

4. ASYMPTOTES OF THE RANK-DISTRIBUTION CURVES

4.1 Graphical Significance

In Section 2 the family of curves of v vs. b with c as a parameter was studied by investigating the implicit function

$$\phi(v,b,c) = c \sum_{r=1}^{v} r^{-b} - 1 = 0$$

There, the intervals of interest were [1.0, 1.2] for b and [0.05, 0.15] for c. Since the series

$$\sum_{r=1}^{\infty} r^{-b}$$

converges for $b > 1$, values of v do not exist that satisfy $\phi(v,b,c) = 0$ for those values of c such that

$$1/c > \sum_{r=1}^{\infty} r^{-b}$$

For fixed c, v tends to infinity as b increases. Therefore it is of interest to find the values of b that yield the asymptotes for these curves.

Since, for $b > 1$, v increases as b increases, the asymptotes will be the vertical lines $b = b^*$ where b^* satisfies

$$1 = c \sum_{r=1}^{\infty} r^{-b^*} = c\, \zeta(b^*)$$

That is, for each value of c the value b^* must be found such that

$$\zeta(b^*) = 1/c \tag{4.1}$$

Unfortunately, tables for the Riemann zeta function cannot be found that permit the calculation of b^* for $c = 0.05(0.01)0.15$. For example, $\zeta(b)$ jumps from 10.584 to ∞ as b goes from 1.1 to 1.0. Thus it is impossible to interpolate intermediate values of $\zeta(b^*)$.

Two methods are suggested here for determining the asymptotes. It turns out that they give similar values. Both of these methods are based on the graph of the curve of $\zeta(b) - \frac{1}{b-1}$ which is tabulated in Fig. 9 and plotted in Fig. 10; see also Walther [1926, p. 396] for a previous plot of this difference. The values $\zeta(b)$ are given in Dwight [1961].

b	$\frac{1}{b-1}$	$\zeta(b)$	$\zeta(b) - \frac{1}{b-1}$
1.1	10.00000 00	10.58444 85	0.58444 85
1.2	5.00000 00	5.59158 24	0.59158 24
1.3	3.33333 33	3.93194 92	0.59861 59
1.4	2.50000 00	3.10554 73	0.60554 73
1.5	2.00000 00	2.61237 53	0.61237 53
1.6	1.66666 67	2.28576 57	0.61909 90
1.7	1.42857 14	2.05428 88	0.62571 74
1.8	1.25000 00	1.88222 96	0.63222 96
1.9	1.11111 11	1.74974 64	0.63863 53
2.0	1.00000 00	1.64493 41	0.64493 41
2.5	0.66666 67	1.34148 73	0.67482 06
3.0	0.50000 00	1.20205 69	0.70205 69
3.5	0.40000 00	1.12673 39	0.72673 39
4.0	0.33333 33	1.08232 32	0.74898 99
4.5	0.28571 43	1.05470 75	0.76899 32
5.0	0.25000 00	1.03692 78	0.78692 78
5.5	0.22222 22	1.02520 46	0.80298 24
6.0	0.20000 00	1.01734 31	0.81734 31
6.5	0.18181 82	1.02100 59	0.83018 77
7.0	0.16666 67	1.00834 93	0.84168 26

Figure 9. Table of Values of $\zeta(b) - \frac{1}{b-1}$.

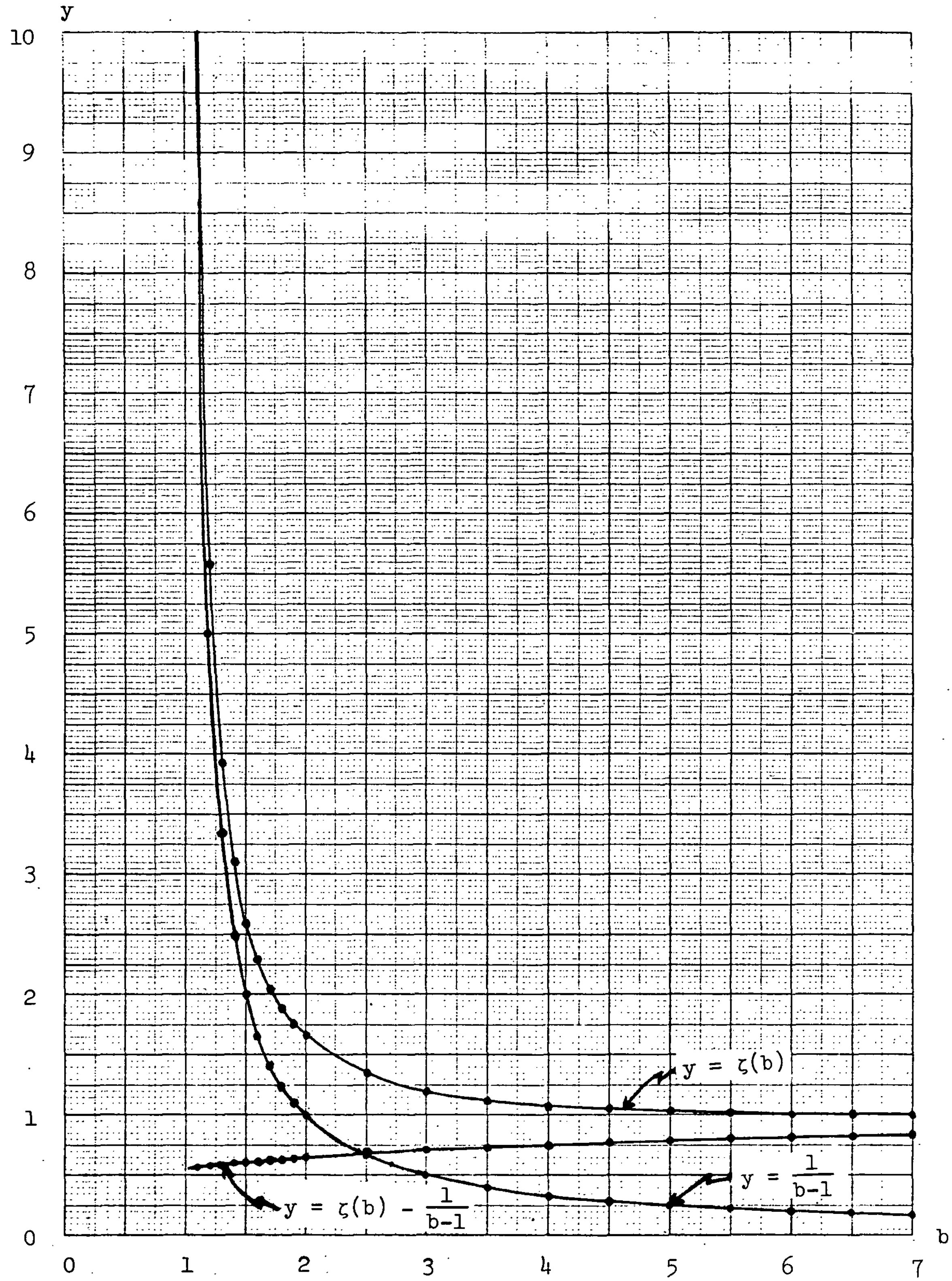

Figure 10. Graph of $\zeta(b) - \frac{1}{b-1}$.

4.2 Constant-value Method

The constant-value method assumes that the value $\zeta(b) - \frac{1}{b-1}$ is nearly constant when b is close to 1. This is confirmed by observing Fig. 10 for b in the interval (1.0, 1.2]; for example,

$$\zeta(1.1) - \frac{1}{1.1-1} = 0.584$$

and

$$\zeta(1.2) - \frac{1}{1.2-1} = 0.592$$

Let

$$a = \zeta(b) - \frac{1}{b-1} \tag{4.2}$$

Thus b* must satisfy both (4.1) and (4.2) and hence must satisfy

$$b^* = \frac{1}{1/c - a} + 1 \tag{4.3}$$

Because (1.0, 1.2] is the interval of b under consideration, the mid-point b = 1.1 is chosen. For this point, a = 0.584 448 464 since $\zeta(1.1)$ = 10.584 448 464.

Fig. 11 is a table of the asymptotes b = b* given by (4.3)

c	b*
0.05	1.051 505
0.06	1.062 180
0.07	1.072 986
0.08	1.083 924
0.09	1.094 997
0.10	1.106 207
0.11	1.117 558
0.12	1.129 051
0.13	1.140 689
0.14	1.152 476
0.15	1.164 414

Figure 11. Asymptotes Obtained by Constant-value Method.

4.3 Straight-line Method

As a generalization of the constant-value method, the straight-line method assumes that the graph of $\zeta(b) - \frac{1}{b-1}$ is close to a straight line when b is close to 1.

Let

$$g(b) = \zeta(b) - \frac{1}{b-1}$$

Under the assumption that g(b) is a straight line, $g''(b) = 0$. Hence it follows from Taylor's formula that

$$g(b) = g(a) + g'(\theta)(b-a) \tag{4.4}$$

where a is some given point and θ is some point between b and a.

Again, a = 1.1 is chosen as the given point. Since it is assumed that $g'(b)$ is constant, the value of $g'(\theta)$ may be calculated as follows:

$$g'(\theta) = \frac{g(1.2) - g(1.1)}{1.2 - 1.1} = 0.071\ 339\ 763$$

Thus b* must satisfy both (4.1) and (4.4) and hence b* must satisfy

$$\frac{1}{c} - \frac{1}{b^*-1} = \zeta(1.1) - \frac{1}{1.1-1} + g'(\theta)(b^* - 1.1)$$

or, equivalently, b* must satisfy

$$Ab^{*2} + Bb^* + C = 0 \tag{4.5}$$

where

$$A = g'(\theta) = 0.071\ 339\ 763$$

$$B = \zeta(1.1) - 2.1\ g'(\theta) - \frac{1}{c} - 10$$

$$C = \frac{1}{c} - \zeta(1.1) + 1.1\ g'(\theta) + 11$$

Fig. 12 is a table of the asymptotes b = b*, given by solving (4.5).

c	b*
0,05	1.051 496
0.06	1.062 171
0.07	1.072 976
0.08	1.083 916
0.09	1.094 994
0.10	1.106 213
0.11	1.117 575
0.12	1.129 085
0.13	1.140 747
0.14	1.152 563
0.15	1.164 538

Figure 12. Asymptotes Obtained by Straigt-line Method.

Note that the above values of b* agree with those in Fig. 11 to 3 decimal places. On the other hand, values of b are considered only in increments of 0.01. Hence the constant-value method is good enough for determining the asymptotes $b = b^*$ for various values of c.

The asymptotes of the parameterized family of curves $\phi(v,b,c)$ are plotted in Fig. 13.

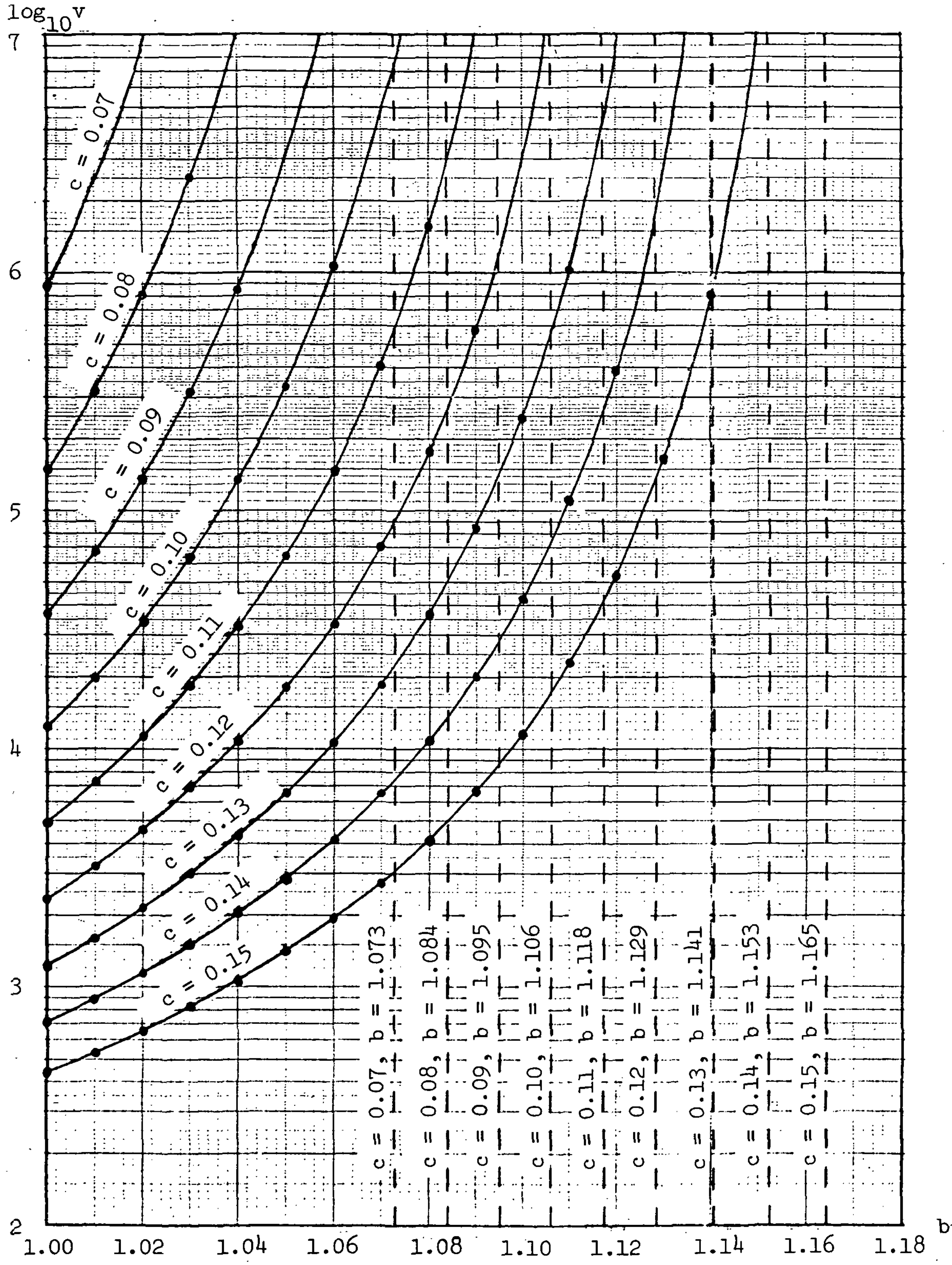

Figure 13. Parameterized Family of Vocabulary Curves and Their Asymptotes.

5. SUMMARY

The major result of this paper has been the computation of the vocabulary size v, given the values of the linguistic parameters b and c, which appear in the 2-parameter rank distribution

$$p_r = cr^{-b} \qquad b \geq 1, \; c > 0$$

for $r = 1,\ldots,v$. This result provides linguists with a parameterized family of curves, shown in Fig. 5, which will permit them to do the following:

(1) given any two of the three quantities v, b, and c, find the third

(2) given any one of the three quantities v, b, and c, find the set of all possible pairs of the remaining two.

Assume for the sake of example that the 130,000 entries contained in Webster's Seventh New Collegiate Dictionary [1967] represent the vocabulary size v of English. Then from Fig. 5 it may be seen that any one of the following pairs of values of the parameters b and c will yield this value $v \doteq 130{,}000$: (1.02, 0.09), (1.04, 0.10), and (1.06, 0.11).

A second result of this paper has been the determination of values of the parameters b and c for which v is undefined. These values are represented in Fig. 13 as asymptotes to the family of vocabulary-size curves. The two methods used to determine these asymptotes yield very close results. Hence the simpler constant-value method suffices.

Finally, an error bound has been determined for the partial product of the Riemann zeta function as an approximation to the partial sum of the Riemann zeta function. For values of the parameter b considered in this research, the error bound indicates that the partial product is a poor approximation of the partial sum. However, for other values of the parameter b, the approximation is good.

Comprehensive tables of the vocabulary size v for the 2-parameter rank distribution are given in Appendices A and B.

REFERENCES

M. Apostol (1957), Mathematical Analysis, Addison-Wesley, Cambridge, Mass.

B. A. Chartres (1967), "Algorithm 310: prime number generator 1", Comm. ACM, vol. 10, no. 9, p. 569.

H. B. Dwight (1961), Mathematical Tables of Elementary and Some Higher Mathematical Functions, 3rd ed., Dover, New York, pp. 210-213.

H. P. Edmundson (1972), The Rank Hypothesis: A Statistical Relation Between Rank and Frequency, Technical Report TR-186, Computer Science Center, University of Maryland.

H. P. Edmundson, G. Fostel, I. Tung, and W. Underwood (1972), Approximation Formulas for Vocabulary Size for the One-, Two-, and Three-Parameter Rank Distributions, Technical Report TR-188, Computer Science Center, University of Maryland.

M. Joos (1936), "Review of Zipf's 'The Psycho-Biology of Language'", Language, vol. 12, pp. 196-210.

E. Jahnke, F. Emde, and F. Lösch (1960), Tables of Higher Functions, McGraw-Hill, New York.

A. Walther (1926), "Anschauliches zur Riemannschen Zetafunktion", Acta Mathematica, vol. 48, pp. 393-400.

Webster's Seventh New Collegiate Dictionary (1967), Merriam, Springfield, Mass.

G. K. Zipf (1935), The Psycho-Biology of Language, Houghton-Mifflin, New York, reprinted M.I.T. Press, Cambridge, Mass., 1965.

G. K. Zipf (1949), Human Behavior and the Principle of Least Effort, Addison-Wesley, Cambridge, Mass.

Appendix A.

Table of Vocabulary Size: Gross Structure for Various Values of b

Legend:

$$c \sum_{r=1}^{v} r^{-b} \doteq 1$$

$$t = v^{-b}$$

$$q = c \sum_{r=1}^{v} r^{-b}$$

c	v	$\log_{10} v$	t	q
		b = 0.90		
0.15	117	2.0681 859	1.3760 453E-2	1.0014 731
0.14	156	2.1931 246	1.0621 549E-2	1.0002 732
0.13	217	2.3364 597	7.8919 852E-3	1.0009 267
0.12	313	2.4955 443	5.6755 783E-3	1.0004 738
0.11	475	2.6766 936	3.8992 022E-3	1.0002 350
0.10	767	2.8847 954	2.5332 857E-3	1.0001 454
0.09	1,337	3.1261 314	1.5363 202E-3	1.0000 129
0.08	2,573	3.4104 398	8.5232 299E-4	1.0000 466
0.07	5,628	3.7503 541	4.2138 721E-4	1.0000 037
0.06	14,651	4.1658 673	1.7812 281E-4	1.0000 046
0.05	48,744	4.6879 212	6.0376 923E-5	1.0000 021
		b = 0.95		
0.15	204	2.3096 302	6.3951 590E-3	1.0003 510
0.14	293	2.4668 676	4.5339 401E-3	1.0002 603
0.13	441	2.6444 386	3.0745 628E-3	1.0000 564
0.12	704	2.8475 727	1.9715 418E-3	1.0000 627
0.11	1,207	3.0817 073	1.1813 488E-3	1.0001 101
0.10	2,260	3.3541 085	6.5102 402E-4	1.0000 172
0.09	4,743	3.6760 531	3.2192 121E-4	1.0000 049
0.08	11,545	4.0623 940	1.3826 931E-4	1.0000 011
0.07	34,287	4.5351 295	4.9161 716E-5	1.0000 018
0.06	134,241	5.1278 852	1.3443 400E-5	1.0000 006
		b = 0.99		
0.15	369	2.5670 264	2.8750 402E-3	1.0000 352
0.14	578	2.7619 278	1.8437 052E-3	1.0000 983
0.13	967	2.9854 265	1.1077 144E-3	1.0000 927
0.12	1,756	3.2445 245	6.1365 002E-4	1.0000 479
0.11	3,538	3.5487 578	3.0671 133E-4	1.0000 201
0.10	8,148	3.9110 510	1.3429 491E-4	1.0000 031
0.09	22,379	4.3498 407	4.9392 133E-5	1.0000 038
0.08	78,007	4.8921 336	1.4347 881E-5	1.0000 007
0.07	379,814	5.5795 709	2.9938 139E-6	1.0000 002
		b = 1.00		
0.15	441	2.6444 385	2.2675 737E-3	1.0001 091
0.14	710	2.8512 583	1.4084 507E-3	1.0000 458
0.13	1,230	3.0899 051	8.1300 813E-4	1.0000 109
0.12	2,336	3.3684 728	4.2808 219E-4	1.0000 350
0.11	4,983	3.6974 908	2.0068 232E-4	1.0000 214
0.10	12,367	4.0922 642	8.0860 354E-5	1.0000 043
0.09	37,568	4.5748 180	2.6618 399E-5	1.0000 023
0.08	150,661	5.1780 007	6.6374 178E-6	1.0000 005
0.07	898,515	5.9535 252	1.1129 475E-6	1.0000 000

c	v	$\log_{10} v$	t	q
		b = 1.01		
0.15	535	2.7283 537	1.7553 460E-3	1.0001 732
0.14	889	2.9489 017	1.0510 158E-3	1.0000 437
0.13	1,604	3.2052 043	5.7908 677E-4	1.0000 544
0.12	3,206	3.5059 634	2.8772 454E-4	1.0000 335
0.11	7,312	3.8640 361	1.2511 908E-4	1.0000 049
0.10	19,850	4.2977 604	4.5631 207E-5	1.0000 036
0.09	68,201	4.8337 907	1.3118 115E-5	1.0000 001
0.08	326,049	5.5132 828	2.7013 720E-6	1.0000 001
		b = 1.02		
0.15	660	2.8195 439	1.3306 542E-3	1.0001 624
0.14	1,138	3.0561 422	7.6336 970E-4	1.0000 646
0.13	2,151	3.3326 403	3.9875 563E-4	1.0000 518
0.12	4,556	3.6595 358	1.8504 333E-4	1.0000 052
0.11	11,285	4.0525 015	7.3527 269E-5	1.0000 043
0.10	34,167	4.5336 068	2.3753 144E-5	1.0000 007
0.09	136,926	5.1364 858	5.7648 026E-6	1.0000 002
0.08	821,128	5.9144 108	9.2747 234E-7	1.0000 000
		b = 1.03		
0.15	830	2.9190 781	9.8480 358E-4	1.0000 489
0.14	1,494	3.1743 505	5.3755 011E-4	1.0000 210
0.13	2,983	3.4746 532	2.6369 824E-4	1.0000 213
0.12	6,807	3.8329 557	1.1273 420E-4	1.0000 083
0.11	18,543	4.2681 799	4.0158 244E-5	1.0000 040
0.10	64,316	4.8083 190	1.1154 022E-5	1.0000 005
0.09	314,124	5.4971 010	2.1776 394E-6	1.0000 000
0.08	2,552,052	6.4068 894	2.5171 200E-7	1.0000 000
		b = 1.04		
0.15	1,070	3.0293 837	7.0703 497E-4	1.0000 935
0.14	2,023	3.3059 958	3.6455 608E-4	1.0000 035
0.13	4,311	3.6345 779	1.6597 358E-4	1.0000 127
0.12	10,735	4.0308 020	6.4263 737E-5	1.0000 052
0.11	32,999	4.5185 007	1.9987 536E-5	1.0000 008
0.10	136,216	5.1342 281	4.5751 232E-6	1.0000 001
0.09	866,023	5.9375 293	6.6829 692E-7	1.0000 000
		b = 1.05		
0.15	1,417	3.1513 698	4.9097 762E-4	1.0000 178
0.14	2,844	3.4539 295	2.3625 117E-4	1.0000 196
0.13	6,554	3.8165 064	9.8325 976E-5	1.0000 023
0.12	18,182	4.2596 416	3.3680 328E-5	1.0000 018
0.11	65,200	4.8142 475	8.8113 020E-6	1.0000 004
0.10	338,995	5.5301 932	1.5606 198E-6	1.0000 001

c	v	$\log_{10} v$	t	q
		b = 1.06		
0.15	1,942	3.2882 492	3.2693 082E-4	1.0000 107
0.14	4,184	3.6215 916	1.4491 486E-4	1.0000 078
0.13	10,623	4.0262 471	5.3973 189E-5	1.0000 003
0.12	33,796	4.5288 652	1.5827 155E-5	1.0000 002
0.11	148,485	5.1716 825	3.2962 228E-6	1.0000 002
0.10	1,064,009	6.0269 415	4.0873 512E-7	1.0000 000
		b = 1.07		
0.15	2,774	3.4431 064	2.0695 512E-4	1.0000 159
0.14	6,520	3.8142 475	8.2938 300E-5	1.0000 110
0.13	18,706	4.2719 808	2.6852 238E-5	1.0000 033
0.12	71,211	4.8525 470	6.4235 443E-6	1.0000 004
0.11	414,033	5.6170 349	9.7672 586E-7	1.0000 000
		b = 1.08		
0.15	4,171	3.6202 401	1.2306 672E-4	1.0000 021
0.14	10,937	4.0388 981	4.3450 023E-5	1.0000 020
0.13	36,797	4.5658 123	1.1719 867E-5	1.0000 013
0.12	179,091	5.2530 737	2.1216 824E-6	1.0000 001
0.11	1,571,650	6.1963 557	2.0320 564E-7	1.0000 000
		b = 1.09		
0.15	6,700	3.8260 747	6.7542 725E-5	1.0000 043
0.14	20,239	4.3061 889	2.0242 028E-5	1.0000 018
0.13	84,512	4.9269 183	4.2624 472E-6	1.0000 005
0.12	587,690	5.7691 482	5.1478 806E-7	1.0000 001
		b = 1.10		
0.15	11,738	4.0695 940	3.3376 944E-5	1.0000 021
0.14	42,895	4.6324 065	8.0232 953E-6	1.0000 006
0.13	244,233	5.3878 043	1.1841 733E-6	1.0000 000
		b = 1.11		
0.15	23,162	4.3647 760	1.4292 184E-5	1.0000 003
0.14	110,882	5.0448 610	2.5130 683E-6	1.0000 002
0.13	1,023,645	6.0101 492	2.1317 401E-7	1.0000 000
		b = 1.12		
0.15	54,267	4.7345 357	4.9810 398E-6	1.0000 004
0.14	392,703	5.5940 641	5.4281 043E-7	1.0000 000

c	v	$\log_{10} v$	t	q
		$\boxed{b = 1.13}$		
0.15	166,038	5.2202 074	1.2623 088E-6	1.0000 002
		$\boxed{b = 1.14}$		
0.15	809,261	5.9080 885	1.8398 361E-7	1.0000 000

Appendix B.

Table of Vocabulary Size: Fine Structure when b = 1.0

Legend:

$$c \sum_{r=1}^{v} r^{-b} \doteq 1$$

$$t = v^{-b}$$

$$q = c \sum_{r=1}^{v} r^{-b}$$

c	v	$\log_{10}v$	t	q
0.100	12,367	4.0922 642	8.0860 354E-5	1.0000 043
0.099	13,681	4.1361 178	7.3094 072E-5	1.0000 005
0.098	15,167	4.1808 996	6.5932 616E-5	1.0000 043
0.097	16,849	4.2265 741	5.9350 703E-5	1.0000 013
0.096	18,759	4.2732 096	5.3307 745E-5	1.0000 005
0.095	20,933	4.3208 314	4.7771 461E-5	1.0000 006
0.094	23,414	4.3694 755	4.2709 490E-5	1.0000 027
0.093	26,251	4.4191 458	3.8093 787E-5	1.0000 006
0.092	29,506	4.4699 103	3.3891 411E-5	1.0000 016
0.091	33,249	4.5217 785	3.0076 092E-5	1.0000 000
0.090	37,567	4.5748 065	2.6619 107E-5	1.0000 000
0.089	42,563	4.6290 321	2.3494 584E-5	1.0000 012
0.088	48,360	4.6844 862	2.0678 246E-5	1.0000 017
0.087	55,107	4.7412 066	1.8146 515E-5	1.0000 004
0.086	62,987	4.7992 508	1.5876 292E-5	1.0000 001
0.085	72,221	4.8586 634	1.3846 388E-5	1.0000 004
0.084	83,079	4.9194 912	1.2036 736E-5	1.0000 007
0.083	95,892	4.9817 823	1.0428 399E-5	1.0000 006
0.082	111,069	5.0455 928	9.0034 123E-6	1.0000 006
0.081	129,115	5.1109 766	7.7450 335E-6	1.0000 001
0.080	150,660	5.1779 978	6.6374 618E-6	1.0000 002
0.079	176,489	5.2467 175	5.6660 755E-6	1.0000 003
0.078	207,585	5.3171 959	4.8173 037E-6	1.0000 001
0.077	245,192	5.3895 062	4.0784 365E-6	1.0000 001
0.076	290,884	5.4637 197	3.4377 965E-6	1.0000 002
0.075	346,666	5.5399 112	2.8846 209E-6	1.0000 001
0.074	415,109	5.6181 620	2.4090 058E-6	1.0000 000
0.073	499,525	5.6985 571	2.0019 018E-6	1.0000 000
0.072	604,207	5.7811 356	1.6550 619E-6	1.0000 001
0.071	734,753	5.8661 413	1.3610 016E-6	1.0000 000
0.070	898,514	5.9535 247	1.1129 487E-6	1.0000 001
0.069	1,105,200	6.0434 408	9.0481 360E-7	1.0000 001
0.068	1,367,733	6.1360 012	7.3113 685E-7	1.0000 000
0.067	1,703,432	6.2313 247	5.8705 014E-7	1.0000 000
0.066	2,135,683	6.3295 367	4.6823 428E-7	1.0000 000
0.065	2,696,317	6.4307 709	3.7087 627E-7	1.0000 000

Appendix C.

Table of Partial Sums and Partial Products of the Riemann Zeta Function

Legend:

$$n \doteq 10^m$$

$$S_n = \sum_{r=1}^{n} r^{-b}$$

$k = \pi(n) =$ number of primes $\leq n$

$$P_k = \prod_{p \leq n} (1-p^{-b})^{-1}$$

m	n	$\log_{10} n$	S_n
		$\boxed{b = 1.0}$	
1.0	10	1.0000 000	2.9289 682
2.0	100	2.0000 000	5.1873 756
3.0	1,000	3.0000 000	7.4854 442
4.0	10,000	4.0000 000	9.7870 694
5.0	100,000	5.0000 000	12.0842 53
		$\boxed{b = 1.1}$	
1.0	10	1.0000 000	2.6801 551
2.0	100	2.0000 000	4.2780 222
3.0	1,000	3.0000 000	5.5727 979
4.0	10,000	4.0000 000	6.6030 995
5.0	100,000	5.0000 000	7.4191 992
		$\boxed{b = 1.2}$	
1.0	10	1.0000 000	2.4677 133
2.0	100	2.0000 000	3.6030 320
3.0	1,000	3.0000 000	4.3357 395
4.0	10,000	4.0000 000	4.7988 505
5.0	100,000	5.0000 000	5.0886 065

m	k	p_k	$\log_{10}p_k$	P_k
		$\boxed{b = 1.0}$		
1.0	4	7	0.8450 780	4.3749 998
	5	11	1.0413 927	4.8124 996
2.0	25	97	1.9867 717	8.3113 550
	26	101	2.0043 214	8.3944 684
3.0	168	997	2.9986 952	12.3509 49
	169	1,009	3.0038 912	12.3632 02
4.0	1,229	9,973	3.9988 258	16.4242 35
	1,230	10,007	4.0003 039	16.4258 76
4.8	6,050	59,887	4.7773 325	19.6015 65
		$\boxed{b = 1.1}$		
1.0	4	7	0.8450 980	3.6504 009
	5	11	1.0413 927	3.9316 164
2.0	25	97	1.9867 717	5.7867 887
	26	101	2.0043 214	5.8231 302
3.0	168	997	2.9986 952	7.2474 486
	169	1,009	3.0038 912	7.2510 470
4.0	1,229	9,973	3.9988 258	8.2392 852
	1,230	10,007	4.0003 039	8.2396 128
4.8	6,050	59,887	4.7773 325	8.7888 710
		$\boxed{b = 1.2}$		
1.0	4	7	0.8450 980	3.1306 292
	5	11	1.0413 927	3.3173 169
2.0	25	97	1.9867 717	4.3664 417
	26	101	2.0043 214	4.3836 863
3.0	168	997	2.9986 952	4.9652 038
	169	1,009	3.0038 912	4.9664 379
4.0	1,229	9,973	3.9988 258	5.2615 612
	1,230	10,007	4.0003 039	5.2616 444
4.8	6,050	59,887	4.7773 325	5.3873 186

Appendix D.

FORTRAN Program for the Prime Number Generator

```
COMMENT PRIME NUMBER GENERATOR
      INTEGER PRIMES(10000),Q(100),DQ(100)
      LOGICAL LT
C   THIS IS THE UPPER LIMIT OF THE PRIMES TO BE GENERATED
      L=60000
      J=2
      K=2
      PRIMES(1)=2
      PRIMES(2)=3
      Q(2)=9
      DQ(2)=6
      DO 1 N=5,L,2
      LT=.TRUE.
      DO 2 I=2,J
      IF (N.NE.Q(I)) GO TO 2
      Q(I)=N+DQ(I)
      LT=.FALSE.
      IF (I.NE.J) GO TO 2
      J=J+1
      Q(J)=PRIMES(J)**2
      DQ(J)=2*PRIMES(J)
      GO TO 1
    2 CONTINUE
      IF (.NOT.LT) GO TO 1
      K=K+1
      PRIMES(K)=N
      KS=K-9
      IF ((K/10)*10.EQ.K) PUNCH 100,(PRIMES(I),I=KS,K)
  100 FORMAT (10I8)
    1 CONTINUE
      END
```

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)* Computer Science Center University of Maryland College Park, Md. 20742	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP

3. REPORT TITLE

Methods of Computing Vocabulary Size for the Two-parameter Rank Distribution

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

Technical Report

5. AUTHOR(S) *(First name, middle initial, last name)*

Edmundson, H. P.; Fostel, G.; Tung, I.; and Underwood, W.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
March 1972	43	11

8a. CONTRACT OR GRANT NO. N00014-67-A-0239-0004 b. PROJECT NO. c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) Technical Report TR-187 9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*

10. DISTRIBUTION STATEMENT

Distribution of this document is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY Information Systems Branch Office of Naval Research Arlington, Virginia

13. ABSTRACT

This paper describes a summation method for computing the vocabulary size for given pairs of the parameter values of the 2-parameter rank distribution. Two methods of determining the asymptotes of the rank-distribution curves are also described. Tables are computed and graphs are drawn relating pairs of parameter values to vocabulary size. The partial product formula for the Riemann zeta function is investigated as an approximation to the partial sum formula for the Riemann zeta function. An error bound is established that indicates that the partial product should not be used to approximate the partial sum in calculating the vocabulary size for the 2-parameter rank distribution.

DD FORM 1 NOV 65 1473

UNCLASSIFIED

Security Classification

UNCLASSIFIED
Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Statistical Linguistics Rank Distribution Vocabulary Size Riemann Zeta Function Mathematical Modeling Error Bounds Asymptotes						

UNCLASSIFIED
Security Classification