INFORMATION AND CONTROL 10, 386–393 (1967)

A "Law" of Occurrences for Words of Low Frequency

ANDREW D. BOOTH

Dean of the College of Engineering, University of Saskatchewan, Saskatoon, Canada; and Interdisciplinary Professor of Autonetics, Western Reserve University, Cleveland, Ohio

The way in which the number of words occurring once, twice, three times, and so on in a text is related to the vocabulary of the author has been investigated. It is shown that a simple relationship holds under more general conditions than those implied by Zipf's law.

Although Zipf's law is well known to linguists and to students of language statistics, there is another "law" enunciated by Zipf (1938) which holds for words of low frequency of occurrence and which is not well known. The interest of several linguistic friends suggested to the author that the following note might be of general interest, particularly because it shows that Zipf's second law is only partially true and can be replaced by a more general statement which has a greater range of validity.

ZIPF'S FIRST LAW

This law, first stated by Estoup (1916) and popularized by Zipf (1949), can be stated, briefly, as follows: The number of occurrences of each different word in a text is counted and the words are then arranged in a table in which the first word is the most frequent, the second word the second most frequent, and so on. The order of any word in the list is called its *rank* (r) and the number of occurrences of that word its *frequency* (f). Zipf's first law then states that:

$$rf = c,$$

where c is a constant for any particular text.

Table I, derived from a word frequency analysis of a paper by Stiles (1961), illustrates the sort of variation involved. It will seem that the "law" is by no means exactly obeyed but, considering its simplicity, that it is in fair agreement with the data. When more extensive data,

TABLE I

A TYPICAL RANK-FREQUENCY TABLE

Word	Rank (r)	Frequency (f)	rf
The	1	245	245
Of	2	136	272
Terms	3	98	294
To	4	81	324
A	5	65	325
And	6	61	366
In	7	55	385
We	8	52	416
Request	9	49	441
Documents	10	40	400
Which	20	26	520

such as those of Dewey (1923), are used better agreement results, and this extends to the four or five thousand different words of a 100,000-word sample. However, these things are well known, and it is the purpose of this note to illustrate and explain the second, and more general, "law" which holds for words of very *low* frequency of occurrence and not for those of high frequency.

LOW-FREQUENCY WORD OCCURRENCES

When the complete word frequency count is made for a text, it is found that words of high rank, that is of low frequency, occur in such a way that many words have the same frequency. Thus, Table II gives the analysis of three texts processed at the Center for Documentation and Communication Research, Western Reserve University, and of the sampling of newspaper English published by Eldridge (1911). We now calculate the ratios I_1/D, i.e., the ratio of the number of words occurring once to the number of different words for each of the texts; these are shown in Table III.

The next problem is to investigate the remarkable constancy of the ratio of single occurrences to the number of different words in the text, i.e., to the vocabulary. Assume that, in a sufficiently large corpus, the ranks of words do actually differ so that ranking is possible, and let the probability of occurrence of a word of rank r be $p(r)$. The composition of a text whose length is T is thus:

TABLE II

Word Counts for Four Texts

Text:	W.R.U.1	W.R.U.2	W.R.U.3	Eldridge
Total No. words (T):	4325	4409	8734	43,989
No. different words (D):	1001	1211	1698	6,002
Words occurring once (I_1)	541	710	887	2,976
Words occurring twice (I_2)	152	227	273	1,079
Words occurring 3 times (I_3)	94	91	151	516
Words occurring 4 times (I_4)	56	41	90	294
Words occurring 5 times (I_5)	36	32	62	212

TABLE III

Ratios of Single Occurrences to Vocabulary

Text	Different words, D	Words occurring once, I_1	I_1/D
W.R.U.1	1001	541	.54
W.R.U.2	1211	710	.58
W.R.U.3	1698	887	.52
Eldridge	6002	2976	.50

$T\,p(1)$ occurrences of the word of rank 1

$T\,p(2)$ occurrences of the word of rank 2

. . . .

$T\,p(r)$ occurrences of the word of rank r, etc.

To find the number of words which actually occur once we could proceed as suggested by Zipf (1938), who asserted that a word will occur once if

$$1.5 > T\,p(r) \geqq .5\,. \tag{1}$$

Zipf's law (first) suggests that

$$p(r) = k/r, \tag{2}$$

where k is constant for the text so that

$$1.5 > kT/r \geqq .5$$

or

$$r_{\max} = \frac{kT}{.5}, \qquad r_{\min} = \frac{kT}{1.5},$$

where the number of occurrences I_1, or $(r_{\max} - r_{\min})$, is

$$I_1 = kT(2 - 2/3) = 4/3\, kT. \tag{3}$$

To find the number of different words, D, which occur under these assumptions, we have

$$Tp(D) \geqq .5 \tag{4}$$

since D is simply the highest rank of any word.

Equations (2) and (4) lead immediately to

$$D = 2kT, \tag{5}$$

so that $I_1/D = 2/3$.

It is easy to show, following the same argument, that the condition for a word to occur n times is simply:

$$(n + 1/2) > Tp(r) \geqq (n - 1/2),$$

from which it follows that:

$$I_n = kT/(n^2 - 1/4). \tag{6}$$

This is the form in which Zipf stated his "second law;" and it leads immediately to:

$$I_n/I_1 = 3/(4n^2 - 1). \tag{7}$$

It is clear that, although this method of approach predicts the constancy of the ratio I_1/D for different text lengths, it produces a small discrepancy in actual value of this ratio: .67 predicted as compared with the values of .54, .58, .52, and .50 for our texts.

The values of I_n/I_1 also leave something to be desired, as shown by Table IV, where the values of this ratio, predicted by (7) and calculated from the Eldridge text are shown.

A MORE GENERAL "LAW"

It is trivial to extend the arguments of the previous section to "laws" of word frequency other than Zipf's. We shall take the more general

TABLE IV

PREDICTED AND CALCULATED VALVES OF I_n/I_1

Source	I_1/I_1	I_2/I_1	I_3/I_1	I_4/I_1	I_5/I_1
Predicted by Eq. (7)	1	.20	.086	.048	.030
Calculated from Eldridge	1	.36	.17	.10	.07

case:

$$p(r) = \frac{k}{r^\alpha}\ (\alpha > 0). \tag{8}$$

We shall also assume, in place of (1), that the condition for a single occurrence is

$$2 > Tp(r) \geqq 1,$$

or, in general, for n occurrences:

$$(n + 1) > Tp(r) \geqq n. \tag{9}$$

By using (8) and (9), and following our previous argument, we thus obtain:

$$I_n = (kT)^{1/\alpha}\left[\frac{1}{n^{1/\alpha}} - \frac{1}{(n+1)^{1/\alpha}}\right]. \tag{10}$$

Again, the number of different words, D, in the text is given by

$$Tp(D) \geqq 1$$

where

$$D = (kT)^{1/\alpha} \tag{11}$$

This leads to

$$I_1/D = \left(1 - \frac{1}{2^{1/\alpha}}\right), \tag{12}$$

which shows that, even when Zipf's law is replaced by the more general form (8), the ratio I_1/D is still independent of the text length.

The proximity of the calculated values of I_1/D to .5 suggests that Zipf's first law is, at least, a good first approximation to the true state of affairs, and, with $\alpha = 1$, we obtain from (12) $I_1/D = .5$, and from (10)

TABLE V

PREDICTED AND CALCULATED VALVES OF I_n/I_1

Source	I_1/I_1	I_2/I_1	I_3/I_1	I_4/I_1	I_5/I_1
Predicted by (13)	1	.33	.17	.10	.071
W.R.U.1	1	.28	.17	.10	.07
W.R.U.2	1	.32	.13	.06	.05
W.R.U.3	1	.31	.17	.10	.07
Eldridge	1	.36	.17	.10	.07

TABLE VI

MORE EXTENSIVE DATA FOR I_n/I_1

Source	I_6/I_1	I_7/I_1	I_8/I_1	I_9/I_1	I_{10}/I_1
Calculated from Eq. (13)	.048	.036	.028	.022	.018
Eldridge	.051	.035	.028	.029	.015

$$I_n/I_1 = 2/n(n + 1). \tag{13}$$

The values of I_n/I_1 for the fours texts of Table II, and the predicted values, calculated from (13) are shown in Table V. It is clear that the agreement is remarkable. The smallness of the data sample makes it unprofitable to calculate values of I_n/I_1 for $n > 5$ in the case of the three W.R.U. samples, but the series can be continued to $n = 10$ for the Eldridge material and leads to results shown in Table VI, which are still in good agreement.

MANDELBROT'S LAW

Mandelbrot has shown (1957), under quite general conditions, that word frequency should follow a law of the type:

$$p(r) = (B - 1)V^{B-1}(r + V)^{-B}, \tag{14}$$

where B and V are constants. Inserting this in Eq. (9) we obtain:

$$(n + 1) > T(B - 1)V^{B-1}(r + V)^{-B} \geqq n,$$

whence

$$r_{\max} = \left[\frac{T(B - 1)V^{B-1}}{n}\right]^{1/B} - V$$

$$r_{\min} = \left[\frac{T(B - 1)V^{B-1}}{n + 1}\right]^{1/B} - V,$$

so that

$$I_n = (r_{\max} - r_{\min}) = [T(B - 1)V^{B-1}]^{1/B}\left[\frac{1}{n^{1/B}} - \frac{1}{(n+1)^{1/B}}\right], \quad (15)$$

which is identical to (10) if we put $\alpha = B$ and $k = (B - 1)V^{B-1}$.

It follows that the statistical evidence, presented in Tables V and VI, affords no indication which would enable us to discriminate between Zipf's law, the more general form given by Eq. (8), and Mandelbrot's revision.

CONCLUSION

The revised form of Zipf's second law seems in excellent accord with the observed facts. There is no reason to suppose, however, that the rather arbitrary assumption used to deduce I_1 would be equally valid in languages other than English. Nevertheless, the important feature of the demonstration is that the general form of the law of occurrence for low frequency words is independent of the detailed validity of Zipf's law for the distribution as a whole.

Equation (11) gives a means of estimating the Zipf constant, k, for a given author, and this in turn provides a measure of his richness of vocabulary. The implications of the Mandelbrot relationship (14) for vocabulary size estimation have been previously discussed by Mandelbrot himself (1960).

ACKNOWLEDGMENT

The author is indebted to Mr. Tefko Saracevic for raising the problem discussed in this note and to the Documentation Center, Western Reserve University, for producing the first three textual analyses. He is also grateful to the referee for drawing his attention to Mandelbrot's excellent account of the theory of word frequencies and related Markovian models.

RECEIVED: June 21, 1966

REFERENCES

DEWEY, G. (1923). "Relative Frequency of English Speech Sounds." Harvard Univ. Press.

ELDRIDGE, R. C. (1911). "Six Thousand Common English Words." Clement Press, Buffalo.

ESTOUP, J. B. (1916). "Gammes stenographiques," 4th edition., Gauthier-Villars, Paris.

MANDELBROT, B. (1957). "Théorie mathématique de la loi d'Estoup-Zipf." Institut de Statistique de l'Universite, Paris.

MANDELBROT, B. (1960). *Am. Math. Soc. Symp. Appl. Math.*, pp. 180–219.

STILES, H. E. (1961). *Assoc. Computing Machinery* **8,** 271–279.

ZIPF, G. K. (1935). "The Psycho-Biology of Language," p. 40. Houghton-Mifflin, Boston.

ZIPF, G. K. (1938). *Psychol. Record* **2,** 347–367.

ZIPF, G. K. (1949). "Human Behaviour and the Principle of Least Effort." Addison-Wesley, Cambridge, Mass.