DIVERSITY OF VOCABULARY AND THE HARMONIC SERIES LAW OF WORD-FREQUENCY DISTRIBUTION**

By J. B. Carroll
University of Minnesota

Samples of the speech or writing of individuals varying in age, intelligence, and background will be found to differ in what may be termed *diversity*, i.e., the relative amount of repetitiveness or the relative variety in vocabulary. Of two samples of equal length the one of low diversity has fewer different words, most of them common; the sample of higher diversity contains a greater number of different words, so that each word has a lower frequency. Some index of diversity is obviously desirable. A simple expedient would be to take the mean number of different words in samples of a standard size—say, 1,000 words—but since we cannot always obtain this information (where, for example, we know only the number of different words in some larger or smaller sample), it is desirable to make use of what may be called the *diversity curve*, which shows the relation between the number of words in the sample (N) and the number of different words (d).* A constant in the equation of this curve provides a useful index of diversity.

An attempt is made here to derive an equation for the diversity curve on the basis of certain hypotheses which have been empirically confirmed to a considerable extent in previous studies. E. V. Condon's early formulation of the harmonic series hypothesis (1) was inadequate because in effect it made diversity a function of the size of the sample. G. K. Zipf's essentially similar formulation (11, p. 45) provides that the most frequent word constitute 1/10 of all the words in a sample; that the next most frequent word constitute 1/20 of all the words in the sample; the next most frequent, 1/30 of all the words, etc. The harmonic series hypothesis also states that for each word in a given sample

* For our purposes every linguistic form is counted as a separate word, e.g., *do, does, did* are all discrete entities.

** Recommended for publication by Dr. B. F. Skinner, Nov. 10, 1938.

$fr^x = K$, where K and x are empirical constants for the sample; f = the no. of times a given word occurs in the sample; and r stands for the rank of a word in the given sample, the ranks being in order of descending frequency. The rank assigned to words of equal frequency is an average. In most cases $x = 1$, though it may vary slightly with the size of the sample. For the present it may be assumed to take the value 1.00. Zipf has further recognized (12) that account may be taken of diversity (Zipf's "average rate of repetitiveness") by assuming that the word of rank r will occur on the average once in every kr words, where k is an empirical constant whose value is normally 10, but varies inversely with the "rate of repetitiveness." For us k may serve as a direct index of diversity. In a sample of N words, the word of rank r will occur $\frac{N}{kr}$ times. That is, $f = \frac{N}{kr}$; $fr = \frac{N}{k} = K$. The number of different words in a sample of N words may then be derived as a function of N as follows:

If the statistical probability of the occurrence of word r is $\frac{1}{kr}$, then theoretically there will be a number of words which will certainly occur in a sample of size N, viz., all the words from $r = 1$ to $r = \frac{N}{k}$; these will theoretically occur one or more times. In samples of moderate size, these $\frac{N}{k}$ words will not make up the total sample, and the remainder of the sample is hence made up of words whose statistical chances of occurring are less than $\frac{1}{N}$, i.e., where $f = \frac{N}{kr} < 1$. Since these words would occur in a larger sample, their occurrences in a smaller sample are fortuitous. Hence this residue is (N — the total frequencies of the first $\frac{N}{k}$ words), or

$$N - \sum_{r=1}^{r=\frac{N}{k}>1} \frac{N}{kr} \qquad (1)$$

where Σ has one term corresponding to each different word. Then the number of different words in the sample, including the first $\frac{N}{k}$ words is

$$d = \frac{N}{k} + N - \frac{N}{k} \sum_{r=1}^{r=\frac{N}{k}>1} \frac{1}{r} \qquad (2)$$

since the last sum of $\Sigma \frac{1}{r}$ is approximately* $(0.577 + log_e \frac{N}{k})$, when $\frac{N}{k}$ > about 20,

$$d = \frac{N}{k}(.423 + k - \log_e N + \log_e k) \qquad (3)$$

Or,

$$k(1 - \frac{d}{N}) + \log_e k = \log_e N - .423 \qquad (4)$$

By substituting the appropriate values of d and N in any given case from empirical data, equation (4) may be solved for k, the index of diversity, if the harmonic series is known to apply. A test of equation (3) is afforded by plotting its curve when k has been found in a particular sample, and by comparing the observed with the theoretical points. This has been done in the case of material which the writer has collected.** The mean number of different words per 200 words of this material was 137.8, with a standard deviation of 10.0. In order to make the points on the empirical diversity curve comparable, the blocks of 200 words were successively pooled with the cumulated sample in such a way that the mean d per 200 words was kept at approximately 137.8, the value obtained from the total sample. For this material k was found to be 8.76, where N = 8,000 and d = 2,155 for the total sample. The observed points and the theoretical diversity curve are plotted in Fig. 1. The virtual agreement between the two sets of values is shown numerically in Table I. This table also gives the observed values for a short word-count of Santayana's *Last Puritan*. It happens that these values are very close to those of the writer's material and to those calculated from the equation. The *Last Puritan* ma-

* Even though not all ranks are whole numbers.

** This sample was obtained by asking college students to fill in patterns of capital letters and asterisks with words of their own choice. For example, * Y F * might have elicited the response, *Is your father sick?* This method of collecting latent speech is somewhat analogous to the method of Skinner's verbal summator (8).

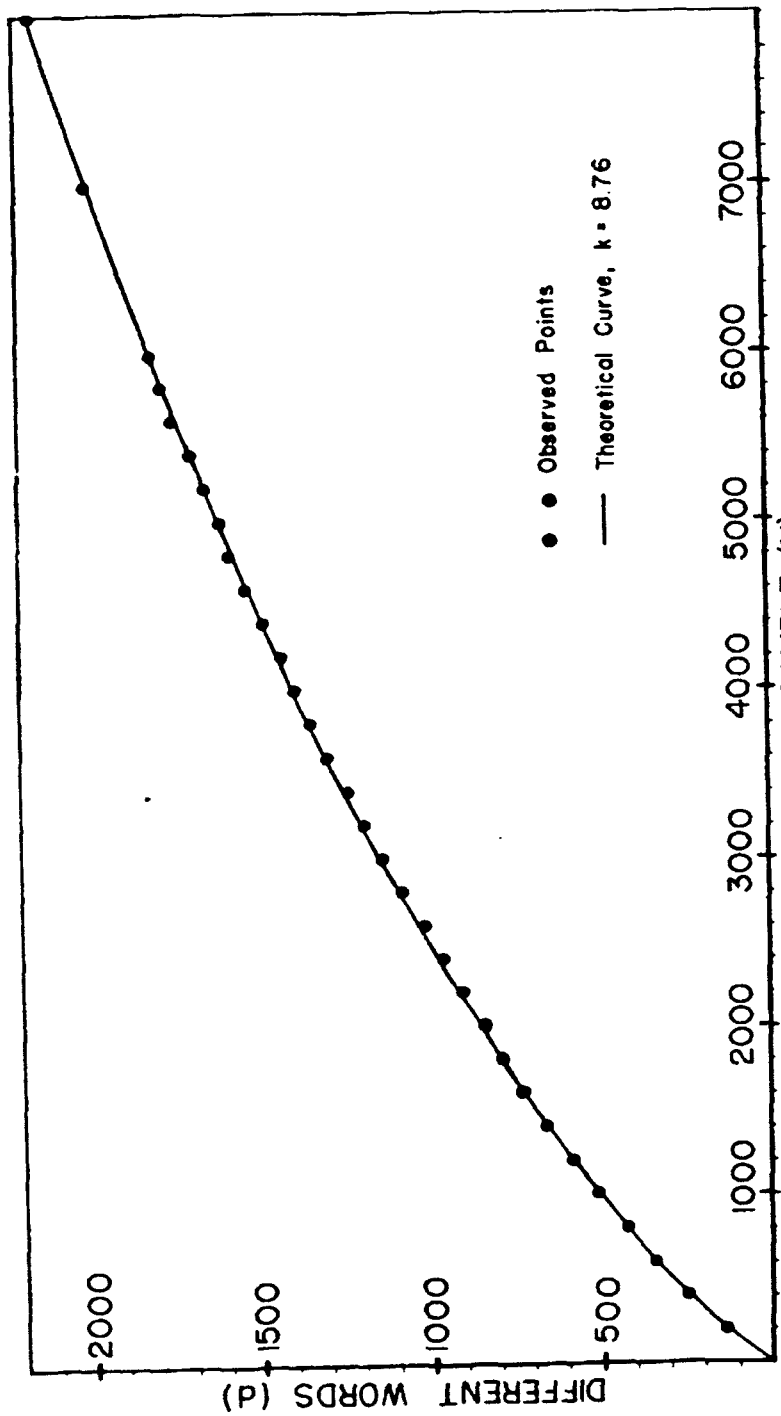


FIG. 1

Observed values of d for the Letter-Star material, with the theoretical curve of equation (3) where $k = 8.76$.

terial was pooled in successive blocks of 200 words in the order in which they stood in the original text.

Equation (3), of course, only applies where $fr^{1.00} = \frac{N}{K}$ for most values of f. The writer's material yields this equation satisfactorily, except for that part of the distribution where r has values of 1 to about 30. This discrepancy, however, does not seem to invalidate equation (3).*

TABLE I

OBSERVED AND THEORETICAL VALUES OF d FOR THE WRITER'S LETTER-STAR MATERIAL, WITH SEVERAL VALUES OF d FROM SANTAYANA'S *LAST PURITAN*

N	Theoret. d ($k=8.76$)	Letter-Star d	Santayana d	N	Theoret. d ($k=8.76$)	Letter-Star d
200	138.2	138	145	3000	1144	1142
400	245	250	252	4000	1392	1394
600	339	345	342	5000	1616	1612
800	426	431	423	6000	1809	1807
1000	507	516	513	7000	1990	1992
1200	583	590	590	8000	(2155)	2155
1400	655	667	668			
1600	725	738	735			
1800	790	793	816			
2000	854	847	886			

The diversities of other samples which are known to follow the simple harmonic series law may easily be calculated. Table II gives N, d, and k for several such samples where the data are available. The data from Joyce's *Ulysses* are included with hesitation since the exponent of r is about 1.07.

* It has been suggested (9) that, in the distribution of associated words, for a discrepancy of this sort in the most frequent words there may exist a compensation in the form of an upward shift of the rest of the harmonic series curve without change of slope. This explanation may also hold here.

TABLE II

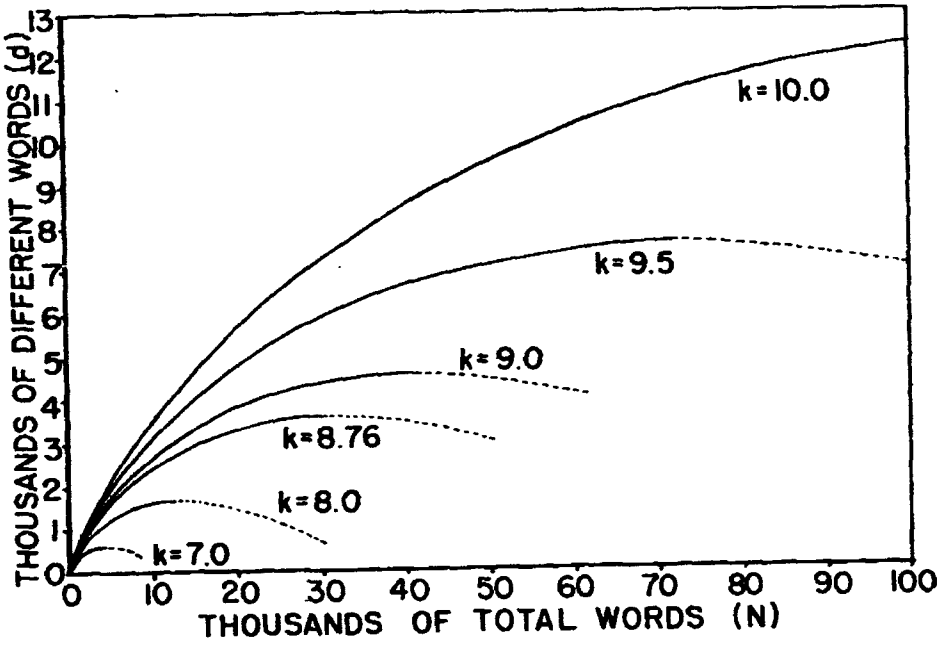
DIVERSITIES OF SEVERAL PUBLISHED SAMPLES

	N	d	k
Writer's Material	8,000	2,155	8.76
Verbal Summator (8)	3,046	806	7.60
Eldridge (3)	43,989	6,002	9.3
Dewey (2)	100,000	10,161	9.8
Joyce's *Ulysses* (4)	260,430	29,899	[10.9]

An important consequence of equation (3) is that the harmonic series law will not apply to very large samples unless an extremely high diversity exists. The maximum value of d for any value of k is where $\frac{1}{k} \Sigma \frac{1}{r} = 1$, i.e., where the first $\frac{N}{k}$ words exhaust the total sample. The higher the diversity, the larger the sample for which $fr = K$ can hold. For $k = 10$, the maximum value of d = about 12,100. Beyond the maximum value of d the harmonic series may still hold for a part of the distribution, but a compensation must exist either as a diminution of the frequencies of the commonest words or as a change in the exponent of r from the "normal" value of 1.00. Further study of the (probably independent) variation of x and k in $fr^x = \frac{N}{k}$ is needed. Unfortunately, such study is hindered by the lack of adequate tables to accomplish the summation of series of the type $\frac{1}{r^x}$ where x is not integral.

When curves of different values of k are plotted on coordinates of d and N (Fig. 2), there are large areas where points could not occur on the ascending portion of any diversity curve, but could only occur on a portion of a curve as it descends towards a zero value of d. This would be true, for example, of the Thorndike sample (10) where N = 4,500,000; d = approximately 55,000; k = 12.5; also of the Brandenburg 4-year-old child G sample (6) where N = 14,930; d = 999; k = 7.67. These values of k are meaningless, for wherever N is larger than its value at the maximum value of d (here

FIG. 2



Theoretical curves of equation (3), plotted for several values of k.

$N = dk$), d is equal to or smaller than its maximum value, and any point will be on the descending portion of a curve whose ascending portion is not truly characteristic of the sample. The simple harmonic series law, at least, cannot hold for such samples.

Though the diversity equation derived here cannot hold when the curve is extended out to large values of N, there is a suggestion that the diversity of a small sample has a definite relation to the total vocabulary represented by the sample, since if we assume the validity of the harmonic series law there is a maximum value of d for a given k, as has been already indicated. The study of vocabulary size would possibly be facilitated by the application of a measure of diversity such as has been suggested here, together with tables of the expected total vocabulary associated with different values of k, the tables having been constructed as a result of both theoretical and empirical investigations. A test of individual vocabulary might possibly be devised on this basis. An index of diversity might also be used to differentiate linguistic materials with respect to stylistic or other characteristics.

BIBLIOGRAPHY

1. Condon, E. V., Statistics of Vocabulary. *Science,* 1928, *67,* 300.
2. Dewey, Godfrey, *Relativ Frequency of English Speech Sounds.* Cambridge, Mass., 1923. Pp. xii + 148.
3. Eldridge, R. C., *Six Thousand Common English Words, Their Comparative Frequency, and What can be done with Them.* Niagara Falls, N. Y., 1911.
4. Hanley, Miles L., *Word-Index to James Joyce's Ulysses.* Madison, Wisc., 1937. Pp. xxiii + 392.
5. Joos, Martin, Review of (11). *Language,* 1936, *12,* 196-210.
6. Nice, M. M., Concerning All-Day Conversations. *Ped. Sem.,* 1920, *27,* 166-177.
7. Santayana, George, *The Last Puritan, a memoir in the form of a novel.* New York, 1936.
8. Skinner, B. F., The Verbal Summator and a Method for the Study of Latent Speech. *Journ. of Psychol.,* 1936, *2,* 71-107.
9. Skinner, B. F., The Distribution of Associated Words. *Psychol. Record,* 1937, *1,* 71-76.
10. Thorndike, E. L., On the Number of Words of Any Given Frequency of Use. *Psychol. Record,* 1937, *1,* 399-406.
11. Zipf, G. K., *The Psycho-Biology of Language.* Boston, 1935. Pp. ix + 336.
12. Zipf, G. K., Observations of the Possible Effect of Mental Age upon the Frequency-Distribution of Words from the Viewpoint of Dynamic Philology. *Journ. of Psychol.,* 1937, *4,* 239-244.