SPECIAL ARTICLES

STATISTICS OF VOCABULARY

WHILE studying some data on the relative frequency of use of different words in the English language, I noticed a rather interesting functional relationship which is here communicated. The note which follows, admittedly incomplete, is published in this form because the subject is one which I can not pursue and which may be of interest to those who are actively engaged in the study of language.

Suppose one takes a large representative sample of written English, counts the number of times each word appears and arranges the words in order of decreasing frequency of occurrence. The n^{th} word in such a list will then occur with an observed frequency which is a function of n, call it f(n). This function is clearly a monotonicly decreasing function of n, from the way the data have been arranged. But what is its form?

Two large published word counts are available. One is that of L. P. Ayres, "A Measuring Scale for Ability in Spelling," Russell Sage Foundation, 1915, and the other is that of G. Dewey, "Relativ Frequency of English Speech Sounds," Harvard University Press, 1923. Each of these writers analyzed samples of 100,000 words of written English.

In the accompanying figure is plotted the logarithm of the observed frequency of the n^{th} word against the

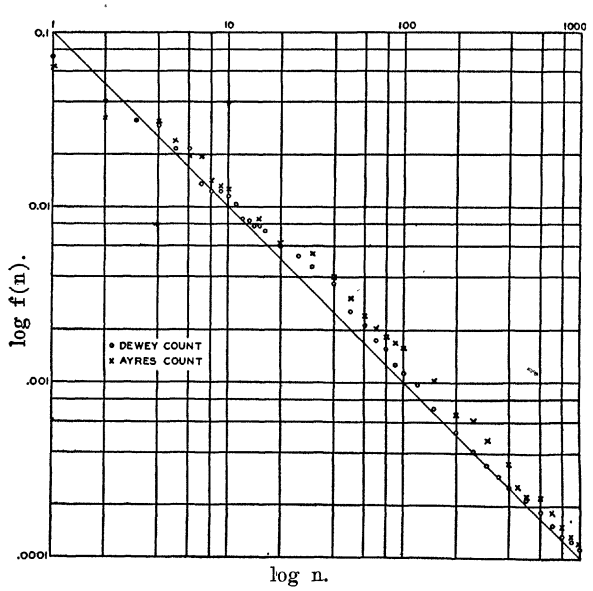

FIG. 1

logarithm of n. The circles are based on Dewey's count while the crosses are based on that of Ayres. The close approximation of the points to a straight line with unit negative slope is at once remarked. This suggests that there is something about the way in which man uses his language (Is the relation true for other languages?) which makes the frequency of occurrence of the n^{th} word be given by a formula of the form,

$$f(n) = \frac{k}{n}$$

On this form some comments will be made at the close of this letter. An interesting question concerns the value of the constant, k. Supposing the law to be valid over the entire range of the language, k must have such a value that the result of summation over all the different words in the sample will equal unity. That is, k is determined by the equation,

$$k \sum_{n=1}^{m} \frac{1}{n} = 1.$$

Since for large values of m (the number of different words in the sample) the summation can be replaced by $\lambda + \log_e m$, where $\lambda = 0.5772$, is Euler's constant and the logarithm is to the natural base, one has a ready means of computing the value of k from the total number of different words in the sample. Dewey found 10,161 different words in his sample, accordingly the value of k is 0.102.

A sort of check on the accuracy of this representation is given by assuming that the most infrequent word occurred just once and inferring from that fact and the value of k the total size of the sample. The value would clearly be 10,161 divided by k, or 99,500., which checks well with the actual size of sample counted, *i.e.*, 100,000.

On the figure the solid line has been drawn to represent the function with $k = 0.100$. It is seen to fit the data quite well, although there are systematic deviations.

Supposing the law to be substantially correct, the writer ventures to point out that it is perhaps a quantitative appearance in language of the Weber-Fechner law of psychology. In the language of the economist, it is a quantitative law of diminishing utility in vocabulary. The frequency of use of a word measures in some way its usefulness in transmitting ideas between individuals. Considering a vocabulary of in words it appears that the marginal increase in idea-transmitting power which can be accomplished by the addition of another word to the vocabulary is smaller the greater the value of n, according to the same law which governs the relation between the psychological increase in sensation accompanying an increase in the total intensity of the physical stimulus.

E. U. CONDON

BELL TELEPHONE LABORATORIES,
NEW YORK CITY