WOLF

The PSYCHOLOGICAL RECORD

ol. I MAY, 1937 No. 6

THE DISTRIBUTION OF ASSOCIATED WORDS

B. F. SKINNER

The Principia Press, Inc.
Bloomington, Ind.

THE DISTRIBUTION OF ASSOCIATED WORDS*

B. F. SKINNER
University of Minnesota

In a sufficiently large sample of English speech Zipf[1] has shown that, when all words are arranged in order according to frequency of occurrence, there is a relation between the frequency of a word and its rank such that a straight line is obtained on logarithmic paper when frequency is plotted against rank. Zipf has used this "standard curve of distribution"[2] in comparing English with an inflected language, and I have used essentially the same relation (with frequency expressed as percentage of the total sample) in comparing standard English with verbal material obtained from subjects who "read words into" chance groupings of speech sounds.[3]

If a similar relation holds for samples of speech *selected on a semantic basis,* it should have a more direct bearing upon the dynamics of verbal behavior. A semantic selection could be made by choosing all the words in a sample that fall within some meaning-category, e.g., all words referring to color. Such a selection possesses a known internal bond upon which the distribution of frequencies may throw some light. But comparable selections of semantically related verbal responses are already available in the literature on free association, for example, in the early study of Kent and Rosanoff.[4] A stimulus-word is capable of evoking a large number of associated responses, although it is more likely to evoke some than others. The common property of being evoked by a single word serves to group these responses together semantically. Such a group is comparable with that of the color-words in a vocabulary—or, indeed, is nearly the same when the stimulus-word is "color." On the assumption that the relative strengths of associations in the vocabulary of an individual will be revealed in the relative frequencies with which responses are given by a fairly large group of subjects, the data reported by Kent and Rosanoff may be used to examine the frequency-distribution within a semantically selected sample of speech.

* Manuscript recommended for publication by Dr. J. R. Kantor, May 13, 1937.

1 Zipf, G. K. The Psycho-Biology of Language. 1935.

2 Op. cit., p. 47.

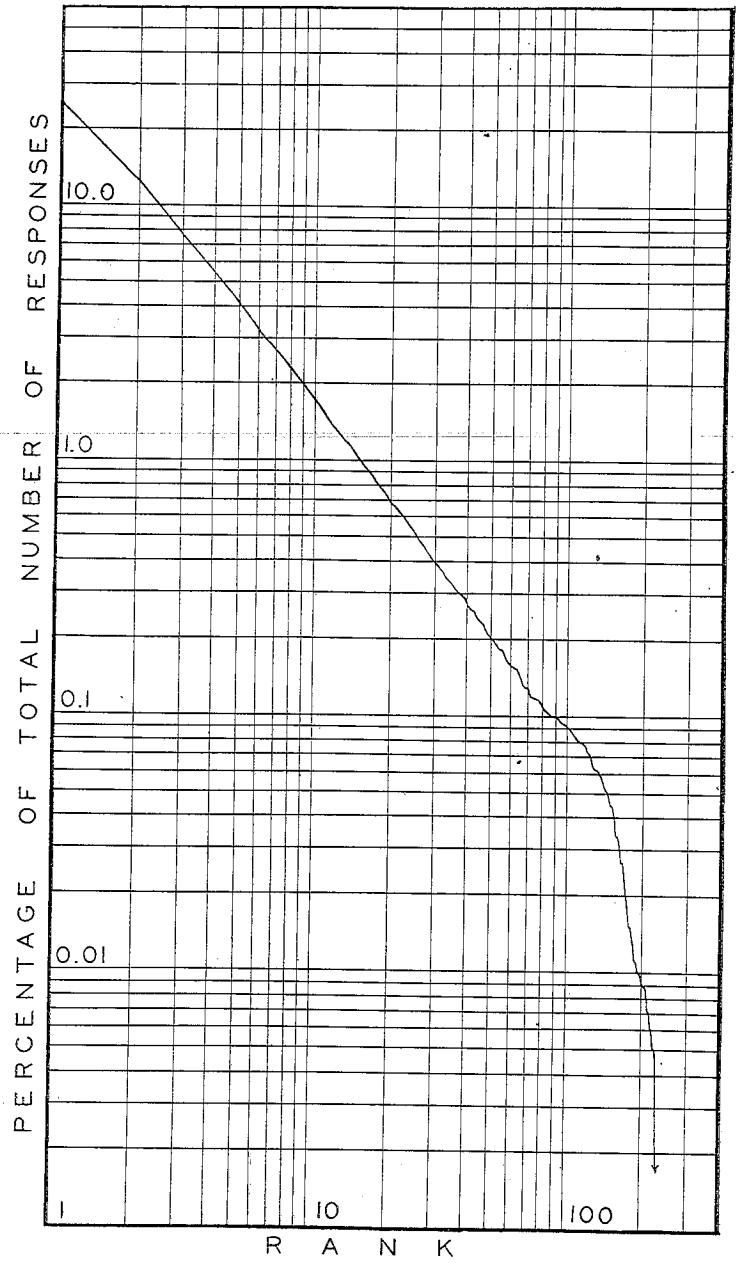
3 Skinner, B. F. The Verbal Summator and a Method for the Study of Latent Speech. Journ. Psychol., 1936, 2, 71-107.

4 Kent, G. H., and Rosanoff, A. J. A Study of Association in Insanity. Amer. Journ. Insanity, 1910, 67, 37-96.

Kent and Rosanoff presented 100 words to 1000 normal subjects, who were told to respond in each case with the first word that occurred to them other than the stimulus-word itself. The words obtained with each stimulus-word, together with their frequencies, were published in the paper cited. I have treated these data in the following way.[5] The frequencies under each stimulus-word were arranged in rank order, separate positions being given to repeated frequencies, and the values in each rank were then averaged. The mean frequency per thousand of the word most likely to be evoked by a given stimulus-word was found to be 258; that of the second most frequent word, 121; that of the third, 77; and so on. All the resulting mean frequencies expresed as percentages of 1000 are plotted against rank in Fig. 1. The curve is approximately linear for the 100 responses most likely to occur. Words beyond this point occur too infrequently (less than once per thousand) to give reliable information. The set of stimulus-words used by Kent and Rosanoff evoked from 71 to 280 different responses, with a median of 145. The curve in Fig. 1 breaks at about the point at which only one-half the stimulus-words are still evoking responses. Had 2000 or 3000 subjects been studied, the linear part of the curve might have been extended further, since longer lists for each stimulus-word would have been obtained.

The principal deviation in the valid section of the curve is a slight upward convexity, most noticeable in the relatively low value for the word occurring most frequently (rank=1). The source of this curvature is apparent when the set of 100 words is divided into four groups according to the frequency of the first word. The limiting initial frequencies for the four groups are: 650-328, 326-250, 246-180, and 180-96 responses per thousand. Means for the twenty-five values in each rank in each group are plotted in Fig. 2. It will be seen that the curvature in Fig. 1 is contributed almost entirely by the group with the lowest frequency in Rank 1. It is significant that this group contains nearly all the abstract stimulus-words in the set (*e.g.*, *religion*, *justice*) and many terms with no sharply defined referents (*e.g.*, *comfort*, *sickness*, *hungry*). There are no single strong associations for words of this sort, and it is clear that the linear relation does not hold rigorously. This may mean, not that the relation is invalid, but that words of this sort are not properly to be regarded as simple units in the dynamics of verbal

[5] I am greatly indebted to Mr. Nathan Gewirtz, a Federal Aid Student, for assistance in tabulating this material.

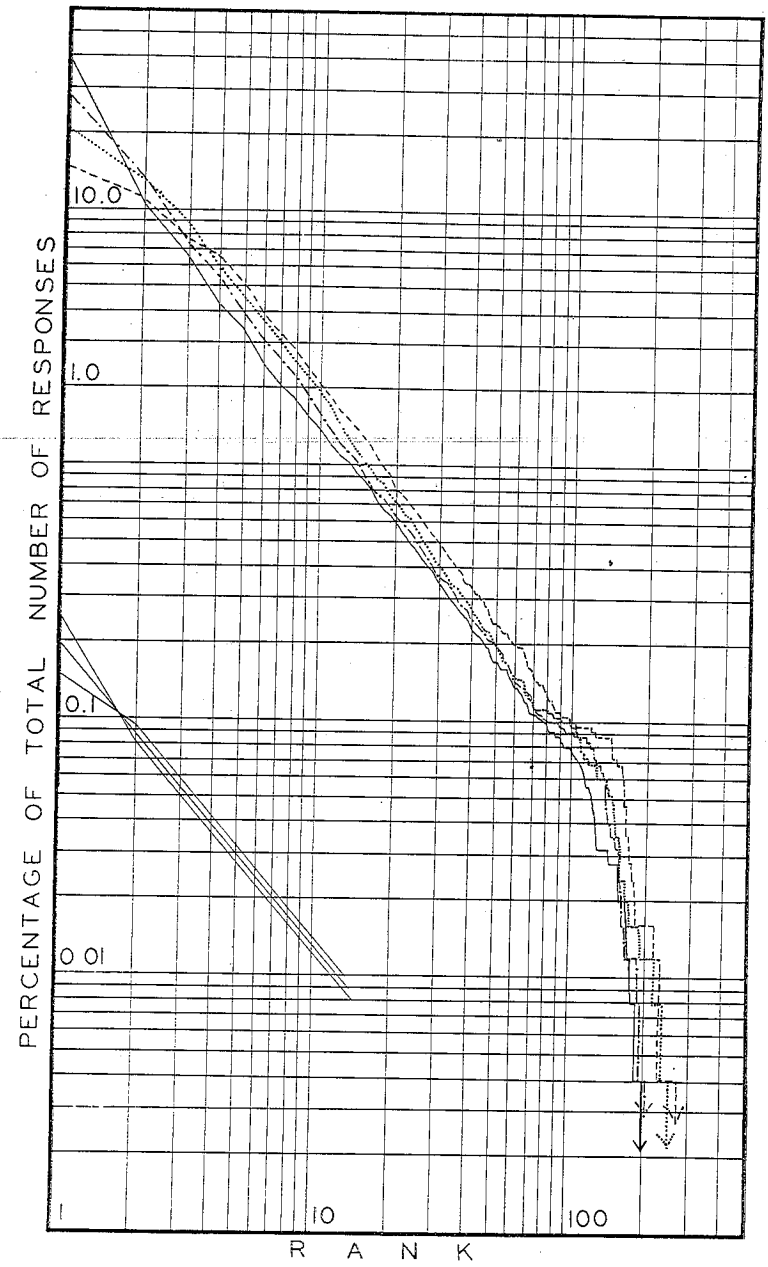

PERCENTAGE OF TOTAL NUMBER OF RESPONSES
10.0
1.0
0.1
0.01
1
10
100
RANK

behavior. The question obviously calls for further data. The other three groups show only slight differences in the character of their words; the separation in Fig. 2 is apparently merely the result of the grouping. Of the three groups the highest and the lowest have opposite curvatures of about the same degree, and the middle is approximately linear.

From Fig. 2 it may be seen that if the most frequent word tends to be especially strong, the additional responses from the total of 1000 that it commands are subtracted from the rest of the curve proportionately throughout its length. The body of the curve is shifted downward without change of slope. Similarly, if the initial response is relatively weak, the additional responses that it fails to command are distributed proportionately throughout the rest of the curve which shifts upward without change of slope. Thus, the segregated curves in Fig. 2 cross each other at the second or third word and remain essentially parallel thereafter. The order of the curves at Rank 1 is the reverse of that throughout the body of the graph. A theoretical case has been entered in the lower part of Fig. 2. It is assumed that a linear relationship is valid when the frequency in Rank 1 is 200. In the curve beginning at 250 the additional 50 responses have been subtracted proportionately from the rest of the curve. In the curve beginning at 150, they have been added.

The hump in the curve at Rank=100+ is due to an excess of unique responses. It would occur if a small number of responses were not determined in any way whatsoever by the stimulus-word, and this is doubtless the case. Such responses are random and only rarely duplicated by other subjects; they are attributable to the failure of the stimulus-word to evoke a response and the pressure put upon the subject to respond with something at any cost. The fact that the hump is especially pronounced for the group containing many abstract words confirms this interpretation, since an abstract stimulus is most likely to produce no associated response and hence to force out a random response extraneously determined.

The data do not reveal the relative strengths of associations to different words in the same rank. Although *man* evokes *woman* as often as *cabbage* evokes *vegetable* when *man* and *cabbage* are each presented 1000 times and 1000 responses demanded, it is quite possible that the "absolute" tendency for *man* to evoke *woman* is greater or less than that of *cabbage* to evoke *vegetable*. If the frequencies had been proportional to absolute strength rather than


PERCENTAGE OF TOTAL NUMBER OF RESPONSES
10.0
1.0
0.1
0 01
1
10
100
R A N K

summing to 1000, there would presumably have been a family of curves of the same slope but with different intercepts on the y-axis. The curves could have been brought together by plotting percentages as in the present case.

A simple statement of the result of this analysis is that for the 75 words having the strongest first associations in the Kent-Rosanoff list (or, slightly less accurately, for the total list), the frequency (f) with which a given association will occur in 1000 responses may be determined from its rank (R) according to the relation

$$f = \frac{300}{R^{1.29}}$$

A comparison of calculated and observed frequencies at a few points along the curve is given in the following table:

Rank	1	4	10	30	50	80
Calculated	300	50.4	15.1	3.7	1.90	1.04
Observed	295	52.3	16.5	3.7	1.92	1.05

This equation has little practical significance because the frequency must be determined before ascertaining the rank, but the simplicity of the relation is nevertheless important for the study of the dynamics of speech. Its bearing upon theories of speech could be stated only at greater length than this report will allow.