Psychol Res (1982) 44:97–103

Psychological Research

Zipf's Law in Transcribed Speech[1]

Dennis R. Ridley

Cornell Institute for Occupational Education, Department of Education, Stone Hall, Cornell University, Ithaca, New York 14853

Summary. Based on large samples of written text. 'Zipf:s Law' holds that the logarithms of frequencies of words, and the number of different words at those frequencies, have an inverse linear relationship. This study tested this law on an oral sample by analysis of a transcribed interview which had been conducted for an unrelated study. A second purpose was to test whether the law holds differently for open- vs closed-class words. The words in the interview were tabulated by frequencies and numbers of words at each frequency, and separated by syntactic class. As Zipf found for written samples, the law provides a reasonable summary of the data for the less frequently used words but predicts fewer different words among the most frequently used words than were actually found. Linear regressions performed for each syntactic class revealed that the regression equation for open-class words provided the best fit to Zipf's law. The sample of closed-class words, containing a much smaller proportion of low frequency words, revealed a poor fit to Zipf's law.

It is well known that most text contains a few words that occur often (e.g., a, the, have) and many different words that appear a small number of times (e.g., autochthonous, soteriological). Writers apparently communicate their meaning adequately for most purposes using a small number of common words. However, this inverse relationship between the frequencies of words and the number of different words at those frequencies has long been known to have a surprisingly precise relationship. Zipf (1935) was the first to find that this regularity could be expressed in a mathematical formula which has been referred to as 'Zipf's law.'[2] The formula is

1 The author wishes to thank Diane Beyerchen, Wendy Calla, Edward Kulick, and Clinton B' Walker for their cooperation and suggestions.

2 'Zipf's law' can also refer to another closely related formula in addition to $ab^2 = k$. On the basis of Zipf's observations that this relationship varies with sample size, he proposed another formula relating the frequencies of words with the ranks in terms of frequency and suggested this would hold for all samples provided they are large enough to give statistical significance.

00340–0727/82/0044/0097/$ 01.40

$$ab^2 = k,$$

where *a* is the number of different words at a given frequency, *b* is the frequency, and *k* is a constant. No constant could be specified by Zipf because, as his data revealed, its magnitude increases with the sample size. Zipf maintained, however, that the same relationship between *a* and *b* holds across a wide range of sample sizes and languages, and is entirely independent of the subject matter of the speech.

Zipf restricted his attention to samples of connected written discourse containing many thousands of words. He maintained that samples of about 5,000 words are large enough to demonstrate the law (1935 p. 46). Possibly because of the difficulty of recording verbatim oral samples at the time he did his work, Zipf did not report data on such samples. He did state, however, that oral samples tend to be distorted because vocal and manual idiosyncrasies or gestures peculiar to the speaker are often used in place of words (1935 p. 215).

Significantly, Zipf apparently assumed that the above law holds without respect to the syntactic class of individual words. Zipf's studies merely sought to demonstrate the validity of this law throughout the speech system, ignoring the syntactic distinctions among words. That these distinctions may be significant is suggested by their prominence in first language learning.

Purpose of the Study

The present study is a test of Zipf's law with a transcribed sample of one person's speech. The major purpose was to test whether the precise mathematical relationship summarized by Zipf's law still holds for a small sample of speech. A second purpose was to test whether a crucial syntactic distinction among words (i.e., open- vs closed-class words) has any effect on Zipf's law.

Oral samples are of interest for at least three reasons. 1) A check of the literature revealed no such tests of Zipf's law on oral samples. 2) Zipf believed that oral samples are poor sources of data for this type of investigation owing to the presence of vocal and manual gestural distortion as noted above. 3) Finally, differences between the frequency distributions for oral and written speech might shed light on psychological differences between speaking and writing; for example, different strategies might be used for accessing one's lexicon, which would be reflected in frequency distributions. Also, writers may write to a richer recognition vocabulary than speakers speak to, affecting the relative frequency of unusual words in written text compared with their use in speech.

Method

The sample was taken from the interview protocol of a 25-year old female participant in an evaluation research study which was totally unrelated to the purpose of this study.[3] The sample comprised 2,823 words. Although the informant's speech was in-

3 The interview was conducted for the Center for the Study of Developing Nations, University of California at Santa Barbara, as part of an evaluation of a program to encourage the teaching of non-Western history in secondary social science.

terrupted a number of times by probes, it was nevertheless treated as a distinct sample; the interviewer's speech was excluded.

As a preliminary step, the recorded interview was transcribed verbatim without editing. Each word in the typescript was then recorded in alphabetical order and each repetition of the same word was tallied. Proper names, hyphenated words, and homonyms were treated as separate words. Contractions (can't, wasn't, etc) were separated into their constituent parts and each part was treated as a separate word.

The next step was to separate the sample into the open-class words (nouns, verbs, adjectives, and adverbs) and closed-class words, defined conveniently as the remainder of the sample. This work was carried out by the author and his assistants. A predoctoral student in linguistics at Cornell University provided a cross-check of the accuracy of syntactic sorting, finding an extremely small number of errors.[4]

The identical procedures, described below, were performed for each syntactic class and for the total sample.

The data were summarized in the same way as Zipf summarized his data for three separate samples. These data are presented in Zipf's (1935) work in which he discussed the law at length. The three samples were: 'Peiping Chinese,' the 'Latin of Lautus,' and 'American Newspaper English.' Zipf provided two kinds of summaries for each of these samples. The first was a simple tabulation of frequencies compared with the number of words at each of those frequencies. For example, in the Chinese sample, this tabular presentation reveals that there were 2,046 words that occurred only one time. The other method of summarizing the results was to plot the data on double logarithmic graph paper with frequencies of occurrence as the ordinate (y-axis) and number of different words at each frequency as the abscissa (x-axis). Zipf also drew a straight line approximately fitting the data points, with a slope of -0.5 as is consistent with the law.

Results

As a first step in the analysis, linear regressions between the logs of *a* (number of different words) and *b* (frequency), for each syntactic class and for the total sample, were carried out.

The following regression equations (with $y' = \log a$ and $x' = \log b$) were obtained:

(1) *Open-class words:* $y' = -0.593x' + 2.449$
(2) *Closed-class words:* $y' = -1.68x' + 1.676$
(3) *Total sample:* $y' = -0.776x' + 1.676$

The regression lines representing the above regression equations are presented in Fig. 1, 2, and 3 together with the data points for open-class words, the closed-class words, and the combined (i.e., total) sample. (The linear equation, representing the logarithmic form of Zipf's law applied to this sample, which is discussed below, also appears in Fig. 3.)

4 The reliability check revealed that syntactic classifications were in agreement in 94.5% of the cases.

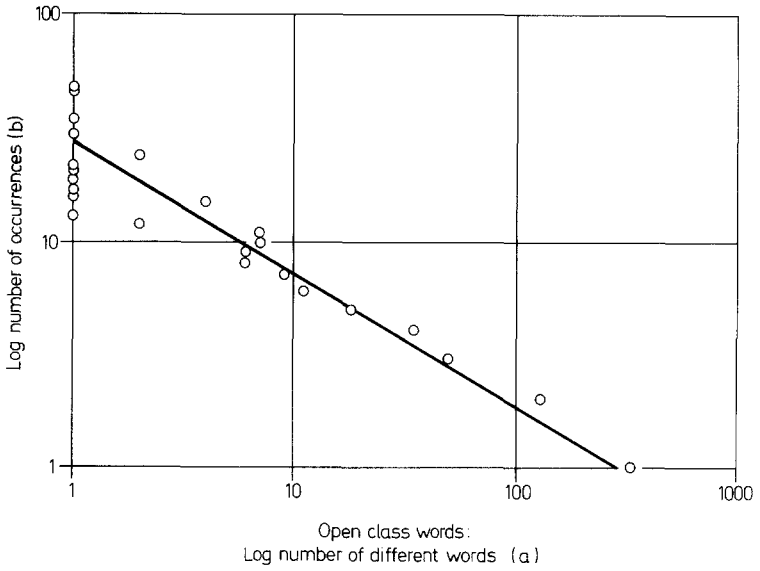

Fig. 1. Log number of occurrences as a function of log number of different words for open class words

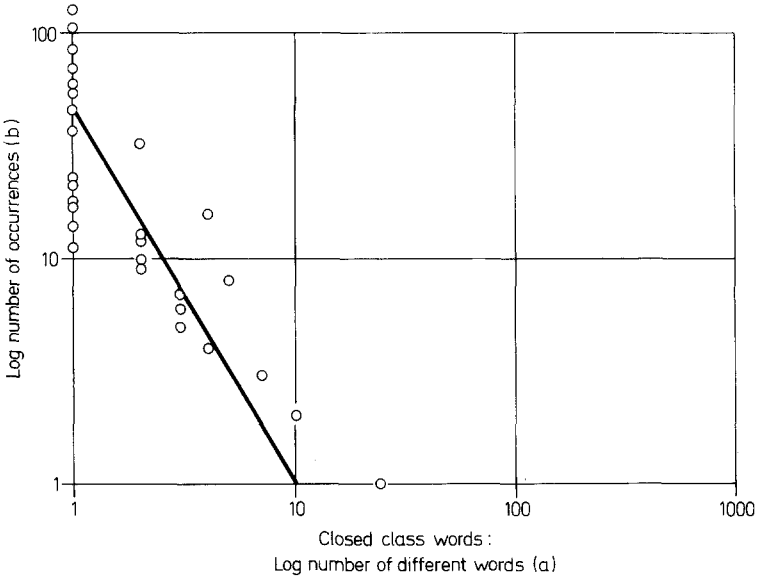

Fig. 2. Same as Fig. 1, but for closed class words

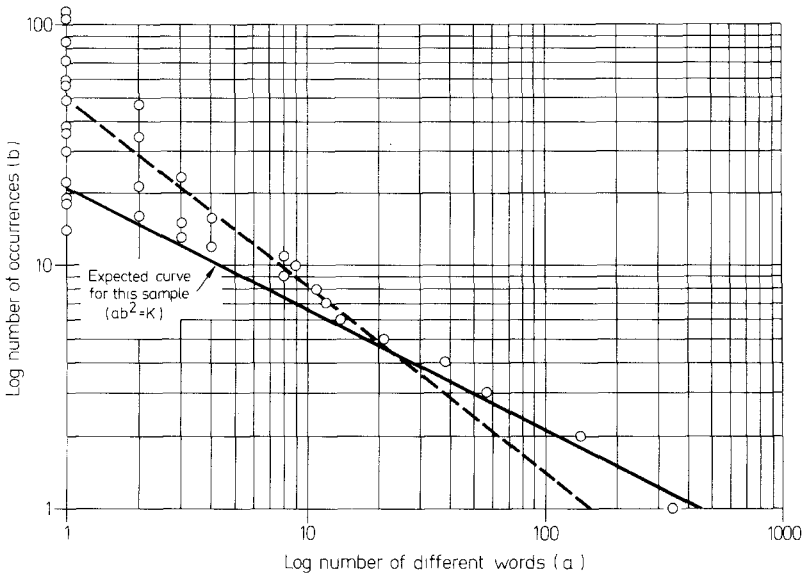

Fig. 3. Same as Figs. 1 and 2, but for the total sample of words. Solid line: Logarithmic form of Zipf's law; dashed line: regression equation as in Figs. 1 and 2

Pearson correlations were also obtained, and showed high negative correlations between a and b, as expected. The r's for open-class, closed-class, and total sample were, respectively: -0.95, -0.85, and -0.91.[5]

For comparison with the exponential equation representing Zipf's law ($ab^2 = k$), the above regression equations also may be expressed in exponential form as follows:

$$(1)'\text{: } ab^{1.685} = 278.0$$
$$(2)'\text{: } ab^{0.593} = 98.8$$
$$(3)'\text{: } ab^{1.289} = 144.5$$

However, since the constant k (which is equal to the value of b at the x-intercept) varies with the size of each sample, a comparison between the above obtained equations and Zipf's formula can be made only in terms of the exponents (i.e., the reciprocals of the slopes in the above regression equations). Alternatively, the slopes in Equations (1), (2), and (3) may be compared with the predicted slope of -0.5.

Although it is possible to estimate the value of k for the present sample by drawing a straight line through the points plotted on double logarithmic paper for this sample, this method of estimation is somewhat subjective. Moreover, using the data to ob-

5 Tests of significance for these correlations are not appropriate since the correlated pairs were taken from a single subject's data. Thus, no statistical generalization can be made to a population value of r.

tain estimates of where the data points should be is essentially circular. Therefore a value of k for this sample was estimated by using Zipf's results in the following way. For each of Zipf's three samples a line through his data points on the double logarithmic paper had been drawn by Zipf. From these lines it was possible by inspection to read off the approximate value of k for each of these samples. The ratio of k to the size of the sample was found to vary between approximately 0.11 and 0.17. For this sample, the estimate of k was found by multiplying the sample size by the mean of the three ratios (0.144). This value, rounded to the nearest whole number, was 405. This estimate is close to what one could reasonably expect from drawing a line with a slope of -0.5 through the data points, yet it was less arbitrary as it was based on prior data rather than the data in hand.

Thus, the form of Zipf's law found for this sample is:

$$ab^2 = 405$$

The linear, or logarithmic, equation corresponding to the above may be compared to equation (1), (2), and (3) above. This equation is:

$$y' = -0.5x' + 2.61$$

Discussion

The results for the total sample suggest that Zipf's law provides a reasonable summary of the relationship between a and b *for the lowest frequency words*. However, the law predicts fewer different words among the most frequently used words than were actually found. This finding is revealed by the steeper slope of the regression line for the total sample as compared with the logarithmic form of Zipf's law.

This result is actually what Zipf found for written samples. Indeed, Zipf commented in some detail on the distortion of the relationship between a and b among the most common words (1935 pp. 41–43).

These results do not support the implicit assumption in Zipf's work that large samples are necessary to demonstrate Zipf's law. Nor do the results support his assumption that the regularities he found in written text would be significantly distorted in transcribed speech. However, the present data do not permit any conclusion about whether a larger sample, or a written sample (from the same subject), would have significantly altered the association between a and b.

The comparison between open- and closed-class words is revealing. It is obvious that open-class words provide the closest approximation to Zipf's law. However, even these words are distorted in the same direction as the total sample; i.e., toward a greater variety of the most frequently used words.

The sample of closed-class words, comprising a relatively larger proportion of common words, was so distorted in this direction that Zipf's law cannot be said to provide an accurate summary of this subset of the data. Thus, a crucial grammatical distinction does affect the usefulness and accuracy of Zipf's law in describing speech behavior. Zipf, however, assumed that his law applies to speech behavior in general, without regard to syntax. This assumption, apparently, is false.

One implicaton of the above is that Zipf's law is systematically distorted for high frequency words partially because closed-class words comprise a relatively high proportion of the most common words. This is true despite the fact that the total number of closed-class words is low in relation to the number of open-class words. The distortion of Zipf's law at higher frequencies may be partially referred to the grammatical composition of those frequencies. In contrast, Zipf's account of this distortion has reference only to the dynamics of communication as affected by the needs of the auditor (1935, Chap. 6).

As regards the possible psychological differences between writing and speaking which should affect frequencies of the words used, these results do not shed much light on such differences. Conceivably, the differences may relate to the different proportions of rare and common words in oral and written samples. For example, if writers write to a richer recognition vocabulary than speakers speak to, they will use more rare words. This effect would perhaps contribute to a relatively higher incidence of frequently used words in oral samples, which would distort the relationship between a and b at high frequencies, as was observed. Thus, if deviation from Zipf's law occurs more in oral than in written samples, the grammatical composition perhaps is a more likely source of the deviation than gestures as Zipf believed.

The psychological significance of confirming Zipf's law in speech behavior is open to debate. Several authors have suggested that Zipf's law, as demonstrated in written samples, is a consequence of statistical constraints within the speech system, and has no relationship to the motives behind the speaker.s behavior, as Zipf assumed (Cherry 1966; Miller 1965; Zipf 1935).

In a broad sense, however, observed relationships have psychological significance whenever they help to describe and predict behavior with more precision than was possible before. The demonstration that Zipf's law holds approximately for a sample of speech has, in this sense, more psychological significance than Zipf's findings with written samples, which are subject to numerous trials and revisions before they take final form. Although these findings may pertain more to the language than to speakers' motives, they do point to the improved description and prediction of speech behavior. Indeed, they suggest that at this level of analysis speech behavior is even more lawful than Zipf supposed.

References

Cherry C (1966) *On human communication.* MIT Press, Cambridge, Mass.

Mandelbrot B (1952) An informational theory of the structure of language based upon the theory of the statistical matching of messages and coding. Proceedings of the London Symposium

Miller GA (1965) Introduction to First MIT Press Paperback Edition of Zipf (1935)

Zipf GK (1935) *The psychobiology of language.* Houghton Mifflin, New York

Received January 27, 1981/September 24, 1981