The stability of the printed frequencies of occurrence of 420 English conceptual nouns*

JOHN H. BOWEN
State University of New York at Albany, Albany, N.Y. 12203

The report presents evidence that a lognormal distribution provides a satisfactory model for the relative frequencies of occurrence of 420 conceptual nouns. The nouns had been randomly selected from the Thorndike-Lorge (T-L) source and had been used as the words in a familiarity scale. The model also provides a satisfactory description of the words from the scale which were common to the T-L count and the later Kučera-Francis (K-F) count and of scaled words in the T-L count which were missing from the K-F count. For common words, there has been a very small shift toward lower frequencies of occurrence in the time between the counts. The shift originated in frequency changes in words from the middle and upper frequency categories of the familiarity scale. The report suggests that differences in types of words are responsible for differences in distributions obtained by Carroll for all words in the K-F count and the distributions presented here. It is suggested, further, that the familiarity scale continues to be useful.

Two major counts of the relative frequencies of occurrence of printed words are available: the Thorndike-Lorge (1944) count (T-L) and the Kučera-Francis (1967) count (K-F). Psychologists have utilized the T-L count more often than the K-F count. As a result, a number of normative scales, including a familiarity scale prepared by the writer (Bowen, 1969), are based upon the T-L count. It is desirable to compare the counts and to be able to judge the effects of shifting usage to the later K-F count. The counts can be compared by way of a distribution model which fits both. The lognormal distribution shows promise as a model for word frequency counts, since it provides a good fit to the Lorge magazine portion of the T-L count and to the entire K-F count (Carroll, 1967, 1969).

The present report applies the lognormal model to the nouns in the writer's familiarity scale for (1) all 420 nouns in the scale, using T-L frequencies, (2) 267 nouns which were common to both frequency counts, using both T-L and K-F frequencies, and (3) 153 nouns which were contained only in the T-L count, using T-L frequencies. Also presented is the average frequency difference between T-L and K-F frequencies for words which were common to both counts, as well as the average frequency difference for common words in each frequency class of the familiarity scale. The report proffers observations concerning certain distributional distortions which were attributed by Carroll to sample size and comments upon the prospects for continued use of the familiarity scale.

*The writer gratefully acknowledges the support of a research grant and faculty research fellowship from the Research Foundation of the State University of New York. Also acknowledged with gratitude are: (1) permission from Dr. John B. Carroll to refer to his research bulletin and (2) the assistance of John Strauch.

METHOD

The words consist of 420 nouns which were selected with the aid of random numbers from pools of words in three frequency-of-occurrence categories and four conceptual categories in the T-L source. The frequency-of-occurrence categories were: $.22 < 1$ words per million (wpm), 1-4 wpm, and 5-100+ wpm. The conceptual categories were names of persons, occupations, animals, and articles of dress. The words and their familiarity scale values are reported elsewhere (Bowen, 1969).

The principal method of the study was to determine the goodness of fit of a normal distribution to the common logarithms of the relative frequencies of occurrence (ϕ) of the words (Carroll, 1967). Plots of normal deviates for cumulative proportions of words vs ϕ were made for the four sets of words which were described above. Nouns which were classified in the A category of the T-L count, with relative frequencies between 50 and 99 wpm, were assigned a ϕ value for the midpoint of that frequency class interval. For 50 wpm, $\phi = 1.69897 - 6.00000 = -4.30103$. For 99 wpm, $\phi = 1.99564 - 6.00000 = -4.00436$. And for the midpoint of the interval, $\phi = -4.15270$. Words which were classified in the AA category of the T-L count, with relative frequencies of occurrence of 100+ wpm, were assigned the ϕ value for 100 wpm ($\phi = 2.00000 - 6.00000 = -4.00000$).

Shifts in the relative frequencies of occurrence of 267 words which were common to both counts were computed by subtracting the K-F value of ϕ for each word from the T-L value of ϕ, averaging the differences and converting the average difference to an approximate wpm value. Thus, negative ϕ differences indicate that the words have *lower* frequencies of occurrence in T-L than in K-F, while positive ϕ differences indicate that the words have *higher* frequencies of occurrence in T-L than in K-F.

RESULTS

Figures 1-4 present plots of ϕ against the unit-normal deviates for cumulative proportions of words and the equations of the functions for each plot.[1] Inspection of the figures suggests that linear functions provide satisfactory fits to all the plots. Since the correlation was .98 in each case, the ϕ variable accounts for 96% of the linear variance of the relative deviates in Figs. 1, 3, and 4. In Fig. 2, the correlation is .97, and 94% of the linear variance of the deviates is accounted for by ϕ. The functions will be used to make comparisons of the distributions.

The first comparison involves Figs. 1 and 2, the distributions for 267 words which were common to both frequency counts, with K-F frequencies, and the distribution for all 420 words, with T-L frequencies. Despite their similarities of slope and intercept, the distribution of common words is attenuated at the low-frequency (high negative log) end of the distribution in comparison to the distribution for all words. The nature of the attenuation is demonstrated by two additional results. First, as shown in Fig. 4, the distribution for 153 words which appeared in T-L but not in K-F is displaced toward the low-frequency end of the scale when compared with all other distributions. Second, of these 153 words, 87 were in the $.22 < 1$ wpm category, 49 were in the 1-4 wpm category, and 17 were in the 5-100+ wpm category of the familiarity scale.

The next comparison is made between the distributions in Figs. 2 and 3, which are based upon 267 words which were common to both counts. The K-F frequencies were assigned to the words to generate Fig. 2; T-L frequencies were assigned to the words for Fig. 3. The function in Fig. 2 is displaced to the left of the function in Fig. 3. The average shift is $\phi = .06745$, or approximately .47 wpm. The average shift for 53 common words in the $.22 < 1$ wpm category is $\phi = -.48291$, for 91

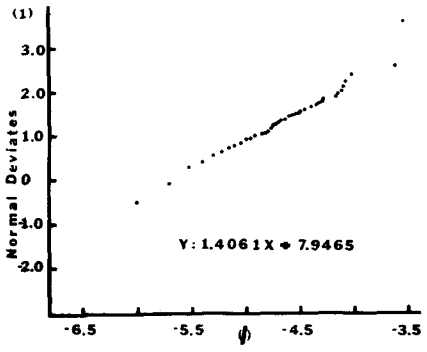

Fig. 1. Relationships between the common logs of the relative frequencies of occurrence (ϕ) for the words and normal deviates for cumulative proportions of words for 267 words common to both frequency counts, using K-F frequencies.

common words in the 1-4 wpm category it was $\phi = .01293$, and for 123 common words in the 5-100+ wpm category it was $\phi = .34493$.

DISCUSSION

The results support a number of general conclusions. First, a lognormal distribution provides a satisfactory description of the relative frequency-of-occurrence distribution for a random selection of conceptual nouns from the T-L source and for the words from that selection which appeared in the K-F count. This conclusion is in accord with Carroll's (1967, 1969) results. The conclusion must be qualified in several ways. First, as Carroll (1967) has noted, there are unresolved conceptual and computational difficulties in using sample data to estimate the parameters of the *asymptotic* lognormal function. It is because of such difficulties that the results of this report have not been used to estimate parameters of the more general model. Second, a comparison of our results with Carroll's suggests that there may be families of lognormal distributions for subsets of words. One basis for the suggestion is that our distributions are shifted toward the low-frequency end of the ϕ scale in comparison with all of Carroll's distributions. Further, Carroll (1967) has shown that with decreasing sample size there is increasing curvature in the low-frequency region of the lognormal plot. Our plots do not show the marked degree of curvature which Carroll's results would have predicted for our sample size. The frequency discrepancies between our results and Carroll's may be due to the fact that we have dealt with nouns, while Carroll used grammatically unselected words. It would be expected that the nonnoun portions of written language would have generally higher frequencies of usage than the nouns because nonnouns have a greater involvement in the production of written structure. The frequency counts support that expectation. Thus, in using only nouns, we have greatly attenuated the high-frequency portion of the distribution for all words and have generated one of the subfamilies of that distribution. What is surprising, however, is the fact that so great a reduction in the size of the population of words seems to have been associated with a reduction in the curvilinearity which Carroll attributed to sample size.

The second general conclusion from the study is that the lower the T-L frequency of occurrence of a noun, the greater is the probability that it will not appear in the K-F count. A related conclusion is that there appears to have been a shift in the time between the frequency counts to slightly lower frequencies of occurrence for the nouns. The shift toward the lower frequencies arises from the words in the middle (1-4 wpm) and upper (5-100+ wpm) frequency groupings. However, all shifts were very small, and none was sufficient to cause a word to fall in a different T-L frequency category.

Our findings have some implications with regard to the continued use of the familiarity scale which provided the words. We have noted that 153 out of 420 words in the scale appeared in the T-L count but did not appear in the K-F count and that the greatest loss, 87 words, was in the lowest frequency grouping of the scale. It is possible to take two positions relative to the words in T-L which were not in K-F: (1) that the words were in current usage but were not counted in K-F because of the nature of the sampled genres or the sampling procedure, or (2) that the words were no longer in current usage. Supporting the first position is the fact that Webster's Third New International Distionary (1967) lists all of the words which were missing in K-F except those person's names which have not been related to real or literary objects or events. The dictionary and all materials which were used in the K-F count were selected in 1961. Further, it is difficult to see how any factors, except sampling or genre selection, could have caused the words, "billfold," "rear-admiral," "dressing-gown," "bluefish," "porker," "tapeworm," "cabby," "bouncer," "dairyman," and "optician," to name a few, to have failed to appear in K-F. These words were in the lowest frequency category of the familiarity scale. The issue of the frequencies of occurrence of the missing words must, of course, remain

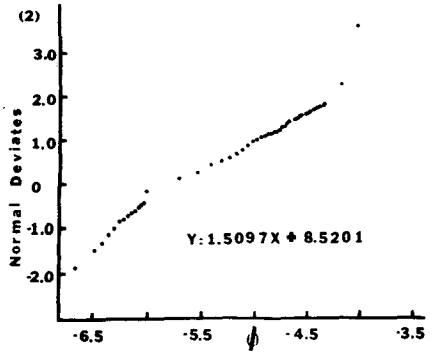

Fig. 2. Relationships between the common logs of the relative frequencies of occurrence (ϕ) for the words and normal deviates for cumulative proportions of words for 420 words, using T-L frequencies.

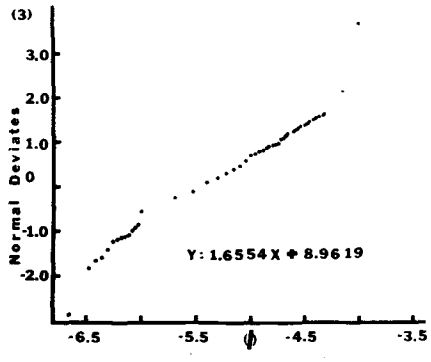

Fig. 3. Relationships between the common logs of the relative frequencies of occurrence (ϕ) for the words and normal deviates for cumulative proportions of words for 267 words, using T-L frequencies.

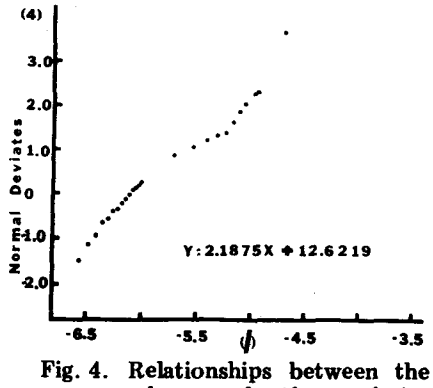

Fig. 4. Relationships between the common logs of the relative frequencies of occurrence (ϕ) for the words and normal deviates for cumulative proportions of words for 153 words in the T-L but not in the K-F count, using T-L frequencies.

unsettled. However, the data show that frequency shifts are very small for words which were common to both counts. If shifts in the frequencies of the missing words were the same as shifts in the frequencies of the common words, then it would be expected that little error would be introduced through the continued use of the scale.

REFERENCES

BOWEN, J. H. Familiarity scale values for 420 nouns in twelve combinations of frequency of occurrence and conceptual categorization. Psychological Reports, 1969, 25, Monograph Supplement 3.

CARROLL, J. B. On sampling from a lognormal model of word-frequency distribution. In H. Kučera and W. N. Francis (Eds.), *Computational analysis of present-day American English*. Providence: Brown University, 1967.

CARROLL, J. B. A rationale for an asymptotic lognormal form of word-frequency distributions. Research Bulletin RB-69-90, 1969, Educational Testing Service, Princeton, N.J.

KUČERA, H., & FRANCIS, W. N. *Computational analysis of present-day American English*. Providence: Brown University, 1967.

THORNDIKE, E. L., & LORGE, I. *The teacher's word book of 30,000 words*. New York: Bureau of Publications, Teachers College, Columbia University, 1944.

WEBSTER'S Third New International Dictionary. (4th ed., unabridged) Springfield, Mass: Merriam, 1967.

NOTE

1. The high positive Y-intercepts are the result of the fact that the functions are plotted in the third and fourth quadrants of a geometric space and have positive slopes.

Learning mixed social structures

NANCY M. HENLEY
University of Maryland, Baltimore County, Md. 21228
and
ROBERT B. HORSFALL*
The Johns Hopkins University, Baltimore, Md. 21218

The goodness of figure of social structures formed of mixed symmetric and assymmetric relations was predicted by a formula taking into account both linearity and balance. The formula was tested in a learning experiment in which 71 Ss learned one of eight six-man hypothetical structures, varying in predicted goodness of figure. Mean errors were 36.63 among the "good" figures and 54.98 among the "poor" figures; in an analysis of variance, the F ratio for the dichotomous goodness of figure variable was 33.42 (df = 1,273, $p < .001$). The product-moment correlation between predicted goodness and learning errors was −.67.

It would be a fine world if social psychologists were able to predict the amount of strain inherent in any social situation, given the relationships among the persons involved. Such predictability would rest largely on a knowledge of the amount of "structural strain" characteristic of different social situations, i.e., strain resulting from imbalances or inconsistencies in the network of relations. Heider's (1946, 1958) balance theory, especially as generalized by Cartwright & Harary (1956), has attempted just this sort of prediction and has been largly supported by experimental verification (Jordan, 1953; Burdick & Burnes, 1958; Kogan & Tagiuri, 1958; Morrissette, 1958; Davol, 1959; Zajonc & Burnstein, 1965; De Soto, Henley, & London, 1968). A second approach to the prediction of structural strain is that taken by those (generally sociologists) concerned with status congruency or crystallization (Benoit-Smullyan, 1944). Numerous studies have shown strain to be associated with incongruency among the various statuses of an individual in a group (e.g., Adams, 1953; Homans, 1953; Lenski, 1954; Exline & Ziller, 1959; Jackson, 1962; Burnstein & Zajonc, 1965; Kasl & Cobb, 1967). Still another approach has been taken by De Soto (1960, 1961), who has brought forward evidence for the existence of ordering, grouping, and symmetry schemas which affect people's cognitions of social groups, strain being produced when the schemas are violated (Wunderlich, Youniss, & De Soto, 1962; Kuethe & De Soto, 1964; De Soto, London, & Handel, 1965; Mosher, 1967; De Soto, Henley, & London, 1968). De Soto has suggested the "predilection for single orderings [1961]" to account for the effects of status incongruency, and "conceptual good figure [De Soto & Albrecht, 1968]" as a concept unifying his approach and that of Heider.

Continuing in the latter tradition, Henley, Horsfall, & De Soto (1969) suggest a graph theoretical approach to assessing conceptual good figure in tournaments (structures formed completely of asymmetric relations).[1] This approach is extended to structures formed of mixed symmetric and asymmetric relations, with good predictive power. Because the present paper builds on the ideas reported in that study, it will help to summarize their predictions and findings. Those authors first used the Kendall & Smith (1940) coefficient of consistency, 1 − (actual cyclic triples/maximum possible cyclic triples), as an index of "linearity" in tournaments, predicting an increase in errors as linearity decreases. This measure was found to correlate well ($r = -.90$) with errors made in learning hypothetical social structures. They next applied a modification of the measure to a reanalysis of earlier results reported by De Soto (1960); the modification was made so that linearity could be measured in structures which are incomplete and/or formed of mixed symmetric and asymmetric relations. Again, the predictive power of the index was high ($r = -.95$). Horsfall & Henley (1969) later applied the modified measure in a study of strain and probability ratings of triadic structures formed of mixed relations, but found high strain and low probability more closely associated with the presence of a single negative relation in a structure than with the presence of a single cycle. They concluded that the failure of the measure to account for rated strain and probability might be due either to differences in the experimental tasks (i.e., learning vs rating) or to the fact that "people might be able to discern and recall well interpersonal relationships in structures which they would find unpleasant or which are unlikely to occur [p. 187]."

The present study was undertaken in order to resolve certain questions resulting from the foregoing research on mixed structures. First, De Soto's structures were composed of four men and were largely incomplete, with few actual relations to learn; the Horsfall and Henley structures were complete but triadic, allowing only a division into "good" and "bad" figure. Does the modified coefficient of consistency predict well for large, complete, and more complex mixed structures? Second, what effect will negative relations have in learning mixed structures; specifically, will the mixed triads with a single negative line

*Now at Simon Fraser University, Burnaby, B.C., Canada.