The Journal of General Psychology, 1945, **32**, 127-148.

THE REPETITION OF WORDS, TIME-PERSPECTIVE, AND SEMANTIC BALANCE*

Harvard University

GEORGE KINGSLEY ZIPF

In the present study we shall attempt to show in preliminary outline how the rate of repetition of words in the stream of speech may be useful not only in indicating what we shall presently define as "time-perspective" but also in elucidating what we shall presently refer to as "semantic balance"—two terms of potential significance in the understanding of personality variants.

As far as the general frequency of occurrence of words is concerned, it has perhaps always been known by students of speech that a few words occur frequently while many occur rarely—a relationship that has become ever more striking as a result of the accumulation of detailed frequency lists of words for many languages as compiled by students of spelling, stenography, linguistics, and psychology.[1] Moreover the data of these lists have been subject in recent years to mathematical treatment of growing rigor.

As to the actual mathematical treatment, perhaps one of the earliest essays was made by J. B. Estoup (8) who published the observation of the approximate hyperbolic relationship between the number of new and different words encountered in successive samples of 1,000 words and the cumulative diversity of vocabulary.[2] Equally interesting were the implicit or explicit formulations of Godfrey Dewey (6) and L. P. Ayers (1). In 1928 E. V. Condon presented the graph of a set of data (5) which he described mathematically, and which we here present in the form of the equation:

$$(1) \qquad r \times f = C$$

in which r refers to the ranks of the different words of a sample of speech when ordered according to decreasing frequency of occurrence, f. In order to illustrate this rank-frequency relationship I present in Figure 1 two excellent sets of data, (*a*) for the 29,899 different words in the 260,430 running words in James Joyce's *Ulysses,* as determined by M. L. Hanley and associates in their excellent index thereto (10); and (*b*) for the 6,002 dif-

*Received in the Editorial Office on August 20, 1943.

[1] For extensive bibliographical data cf. (3), (9, p. 95 f.), (11, p. 10), (12, p. 53 f.) and (16).

[2] Essentially this type of investigation was continued independently by Carroll (3).

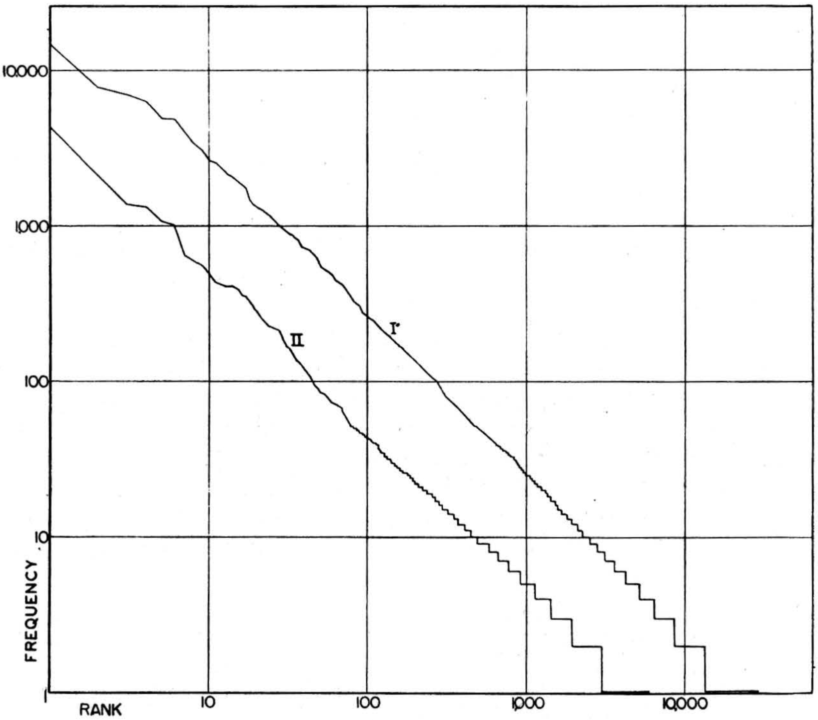

FIGURE 1
THE RANK-FREQUENCY DISTRIBUTION OF WORDS IN (I) JOYCE'S *Ulysses,* AND (II) AMERICAN NEWSPAPERS (ELDRIDGE ANALYSIS)

ferent words in samples of American newspapers aggregating 43,989 running words, as analyzed by R. C. Eldridge (7). In Figure 1 *rank, r,* is plotted logarithmically on the abscissa, and *frequency, f,* on the ordinate. The line connecting the successive points descends from left to right at an angle of 45° (i.e., with a slope of —1) as is to be expected from the above equation. The closeness of the fit in both sets of data is startling.

In 1932 Zipf (27) published his observations (for Plautine Latin, Peipingese Chinese, and the Eldridge newspaper data above) of close approximations to a linear relationship which may now be given in the form of the following equation.[3]

[3] For the development of this equation from Equation 1, cf. (21).

$$N(f^2 - \tfrac{1}{4}) = C \tag{2}$$

in which N refers to the number of different words of like frequency of occurrence, f, in a sample. This Equation 2 has been shown to be corollary to Equation 1, and to be dependent (as is Equation 1) upon the size of the sample examined.[4]

In 1935 J. C. Whitehorn tendered anonymously (25, p. 44) the observation that the Eldridge data above could be felicitously expressed as a harmonic distribution whose equation we shall now venture to present in the following generalized form:

$$F \cdot Sn = \frac{F}{1^p} + \frac{F}{2^p} + \frac{F}{3^p} + \ldots\ldots + \frac{F}{n^p} \tag{3}$$

where n represents the number of different words in the sample when ranked in the decreasing order of frequency, where F represents the frequency of the most frequent word (with the arbitrary assumption that $\frac{F}{n^p} = 1$) and where Sn represents the sum of the n harmonically seriated fractions of the right hand member of the equation (in Eldridge's count and in the *Ulysses* Sn is approximately 10), and finally where $p = 1$ in the case of the "standard distribution of the true harmonic series," which seems to be quite general in American and English.[5]

In Equation 3 to which Equations 1 and 2 are corollary when $p = 1$ we have a very serviceable mathematical description of a rank-frequency distribution of words.[6] Nevertheless—and this point cannot be stressed too vigorously—the generalized harmonic Equation 3, as well as the other two equations tell us absolutely nothing about the intervals between the repetitions of the different words of the sample. Thus for example in the *Ulysses* the word, *say,* whose rank is 100 ($r = 100$) and which occurs 265 times ($f = 265$) might occur once in every 1,000 running words; or it might occur in 265 immediate repetitions with no intervening words and then never occur

[4] The above references will serve as an adequate acknowledgment to Kosambi (14).

[5] For a general discussion of the mathematical properties of the generalized harmonic series and its dependence upon the size of the sample examined, cf. (22, Chaps. 1-4, and 6). I here report the observation of differences in the size of Sn from 10 to 2 in samples of children's speech (to be discussed in Chap. 4 of my forthcoming book, *The Principle of Least Effort*). I also report the observation of cases of negative slopes that are less than 1 (where $p = -$slope, in equation 3) in Nootka and another American Indian language (to be discussed *ibid.*); and a negative slope greater than 1 in the letters of a female paranoid schizophrenic as reported elsewhere (19).

[6] The equations are also corollary when p is not equal to 1, cp. (13).

again; or it might occur according to many other conceivable schemes. Yet no matter in what fashion the word, *say,* is repeated in the *Ulysses,* its point on the doubly logarithmic chart will be the same as long as its rank is 100 and its frequency is 265. For our above equations simply ignore the entire matter of *the rate of repetition of words* even though, as we shall now attempt to suggest empirically and theoretically, it may be precisely this *rate of repetition* of a person's words that may reveal much about the balance of his personality.

Empiric Approach to the *Rate of Repetition of Words*

In 1937 my then student, Alexander Murray Fowler, as previously reported elsewhere (21), undertook as a seminar topic the preliminary exploration of the number of pages that intervened between the repetitions of all the different words that occurred 5, 10, 15, 20, and 24 times in Joyce's Ulysses as determined by Hanley's *Index.*

Fowler's procedure, though inescapably onerous, was simple and essentially as follows. Thus each word that occurred five times was considered to have four intervals, *I,* between its occurrences. And the length of each of these four intervals in terms of intervening pages was determined by subtracting the respective page references from one another, as given in the *Index.* More explicitly the 1st interval was established by subtracting the number of the page on which the word first occurred from that of its second occurrence; the second interval was obtained by subtracting the page of its second occurrence from that of its third; the n — 1 interval by subtracting the page of its n — 1 occurrence from that of the nth. Naturally if the word is repeated on the same page, the interval between the two occurrences would be zero pages. In order to avoid operating mathematically with zero in the calculations below, one page was subsequently added (on Zipf's responsibility) to all intervals, so that, for example, if 20 pages resulted from the subtraction of two successive page-references for a word, the interval was said to be 21. Although this procedure will tend to distort the intervals in the direction of a greater length, the consequences will not be unduly serious.

To resume, after having determined the sizes of the intervals between each of the five occurrences of all the 906 words that occur five times (3,624 intervals in all), Fowler next tabulated the number of occurrences of the various interval-sizes, *I,* for all, the 1st, 2nd, 3rd, and 4th intervals both separately and combined. And in all cases he found not only that short intervals were much more abundant than longer ones, but also that the number, N, of intervals of a given size stood in an approximately inverse linear relationship to the size of the interval, I. This relationship can be described by the equation:

$$N^p \cdot I_f = \text{a constant.} \qquad (4)$$

in which f refers to the frequency of occurrence of the different words whose

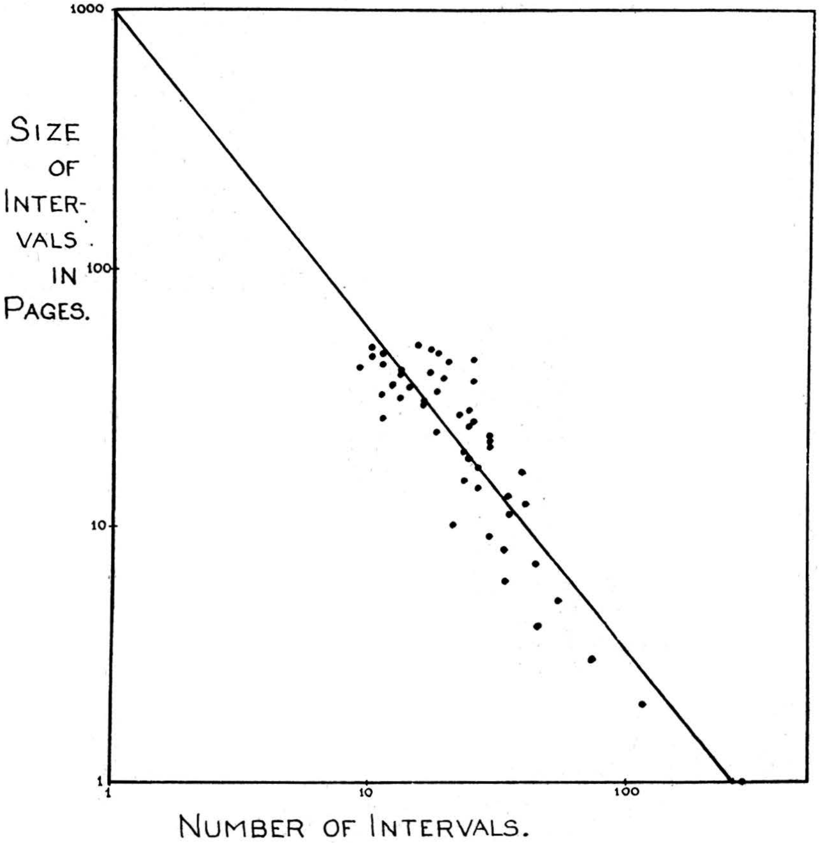

FIGURE 2

THE NUMBER OF INTERVALS OF LIKE SIZES (IN TERMS OF PAGES) BETWEEN THE REPETITIONS OF WORDS OCCURRING FIVE TIMES IN JAMES JOYCE'S *Ulysses* WITH INTERVAL-SIZES TAKING ON INTEGRAL VALUES FROM 1 THROUGH 50 PAGES INCLUSIVE intervals are being measured (in the present case, $f = 5$ because we are treating all words that occur five times in the *Ulysses;* hence the intervals are I_5) and where p is the absolute slope of the line fitted to the points when the data are plotted (as above in Figure 2) on doubly logarithmic paper with N on the abscissa and I_f on the ordinate.[7]

The reason for using the more general form, I_f, instead of I_5, in the above equation is that Fowler found the same inverse relationship for the number and sizes of intervals in each of the classes of words that occurred 10, 15, 20, and 24 times respectively in the *Ulysses.*

Subsequently I checked significant portions of Fowler's analyses and found them accurate to a very high degree. I also extended the analysis to words occurring 6, 12, 16, 17, 18, 19, 21, 22, 23 and 24 times in the *Ulysses* (page-references are not given by Hanley for words occurring more than 24 times) and found the same inverse relationship, although, in studying the data mathematically, I found significant differences in the size of the constant with differences in the size of f (see below).

In order to illustrate graphically the nature of the above-mentioned data, I present in Figure 2 on doubly logarithmic chart paper (with N on the abscissa and I_5 on the ordinate) the number and sizes of all intervals between repetitions from $I_5 = 1$ page through $I_5 = 50$ pages, for all the 906 words occurring five times in the *Ulysses.* The slope of the line of best Y's ($Y = \log I_f$) for the data of Figure 2, as calculated by least squares, is —1.25 (the root-mean-square deviation being .168). Hence we may describe these points mathematically by the equation

$$N^{1.25} I_5 = \text{a constant}$$

if we remember that the curve is discontinuous and that I_5 has only integral values from 1 through 50.

As to the slopes and errors of the remaining 13 frequency-classes (*viz.,* the words occurring 6, 10, 12, 15, 16, 17, 18, 19, 20, 21, 22, 23, and 24 times respectively) I must resort to tabular presentation because of exigencies of limited space. Hence in Table 1 adjoining are presented in Col. IV the negative slopes of the best line of Y's as calculated by least squares; in Col. V is the root-mean-square deviation of these best lines of Y's; and in Col. VI is the Y-intercept of the best line of Y's (actually the antilog of the

[7] The measurement of N on the abscissa instead of on the ordinate, as is to be preferred traditionally, was deliberately decided upon in order to bring the data of Figure 2 into conformity with those of Figure 3. A similar plotting of N was adopted by V. Pareto for his income-curve. The relationship is not altered if the coördinates are reversed.

TABLE 1

Calculated values of negatives slopes, errors, and intercepts of the number, N, of interval-sizes, I_f, between the repetition of words in 14 frequency-classes, f, as fitted to the equation, $aX + Y = C$, and where $X = \log N$ and $Y = \log I_f$, and where I_f has integral values from 1 through 21 inclusive.

I	II	III	IV	V	VI
No. of analysis	Frequency of Occur. (f)	No. of different words of like f	Slope of best Line of Y's (negative)	Error (root-mean square)	Y-Intercept (antilog thereof)
1	5	906	1.21	.151	716
2	6	637	1.20	.169	666
3	10	222	1.27	.106	677
4	12	155	1.24	.111	491
5	15	96	1.15	.096	328
6	16	86	.96	.124	153
7	17	79	1.22	.174	422
8	18	62	1.20	.120	264
9	19	63	1.21	.148	350
10	20	69	1.29	.124	944
11	21	52	1.05	.138	212
12	22	50	1.10	.117	264
13	23	44	1.24	.113	352
14*F*	24	34	1.01	.158	136
15*Z*	24	34	1.05	.147	153

Y-intercept). The calculations are based upon interval-sizes, I_f, from 1 through 21 pages—that is, for the 21 smallest interval-sizes—for all the frequency-classes enumerated in Col. II (the number of different words in each frequency-class is added gratuitously in Col. III). Analysis No. 14*F* of Col. I is Fowler's analysis, and No. 15*Z* is Zipf's independent analysis of the intervals between words occurring 24 times (these are included to suggest the probable closeness of the two separate investigations).

Before turning to an inspection of Table 1, two points should be mentioned in advance. *First,* I restricted the calculation to the 21 smallest interval-sizes because some of the larger interval-sizes between 21 and 50 pages were lacking to some of the higher frequency-classes (cf. Figure 3 below), and hence made impossible a calculation of slopes for a series of points which included instances of $N = 0$; for comparative purposes the 21 smallest interval-sizes were selected as being common to all. *Second,* the *Y*-intercept was added in Col. VI to give the reader an indication of the sizes of the respective constants of the equations.

The negative slopes of Col. IV which range from .96 to 1.29 (with the median at 1.20) clearly reveal a degree of correspondence that is too high to be ascribed to the purely haphazard or random in the entire matter of

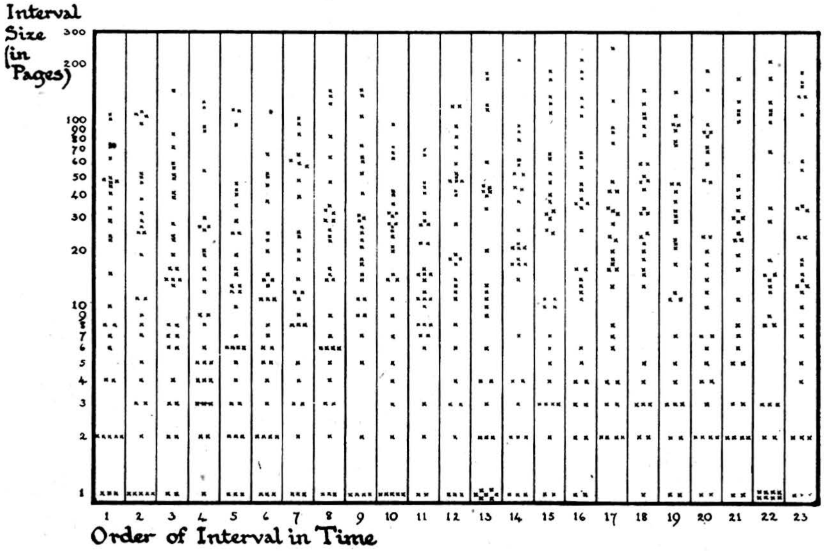

FIGURE 3

THE NUMBER OF INTERVALS OF LIKE SIZE IN THE 23 SUCCESSIVE INTERVALS BETWEEN THE 24 OCCURRENCES OF ALL WORDS OCCURRING 24 TIMES IN JAMES JOYCE'S *Ulysses*
(Each cross denotes an observed interval.)

the spacing of the occurrences of repetitive words in the stream of speech. For there is nothing in our manner of handling the data which presents an a priori reason for expecting a variation of negative slope between .96 and 1.29 inclusive (with the arithmetic mean slope of all 15 analyses at 1.16) because theoretically many other slopes could occur quite as well as that represented by the equation, $N^{1.2} \times I_f = a\ constant,$ which seems to be approximated by our data. Nor can we cogently ascribe this high degree of correspondence in slope and error to the fact that we have restricted our interval-sizes to the 21 smallest page-intervals; for *first* of all we saw in Figure 2 where the 50 smallest page intervals were selected for words occurring five times that the negative slope was 1.25, with error of .168, which is a negligible difference from the slope of 1.21 and error of .151 for the 21 smallest page-intervals of the same frequency-class as given in Table 1; and *second* we shall find in Figure 3 above that the inverse relationship between N and I_f holds for the entire sample. Moreover the correspondence in slopes and errors cannot be ascribed to the fact that we selected a page as a unit of measurement of intervals (although a smaller unit would be indicated for words of very high frequency[8]); nor can the correspondence in slope and error be ascribed to the fact that we added one page as a constant to all interval-sizes in order to avoid operating with zero. Least of all may we believe that the marked agreement of slopes and errors results from the fact that we perforce selected frequency-classes in the lower ranges because the *Index* did not give page-references for words whose f is greater than 24; the most that can (and should) be said in this respect is that the correlation has been demonstrated only for the lower frequency-range (the words occurring 1 through 24 times represent 76,764 occurrences out of the total of 260,430 running words in the *Ulysses*).

In fact by selecting our samples from the lower frequency-classes with correspondingly few intervals, we have concentrated upon that portion of the entire frequency-distribution where we should perhaps least expect to find that any governing principle would be effectively operative upon the spacing of repetitions. For although we might conceivably be prepared to find some sort of principle governing the spacing of repetitions of highly frequent words that occur on the average once, twice, or thrice *in every 100 words,* that is a far cry from a principle that governs the repetition of words of low frequencies that occur once, twice, or thrice *in every 100 pages.*

Indeed that the "$N \cdot I_f$ relationships" of Equation 4 does in fact extend to the few very long intervals of 100 or more pages as well as to those of 21 and fewer pages can be shown graphically in Figure 3 where, for the sake of illustration, is presented the arbitrarily selected case of all the intervals between all the repetitions of all words occurring 24 times in the *Ulysses.* In Figure 3 the 23 successive intervals between the 24 occurrences are plotted arithmetically from left to right on the abscissa, and the sizes of intervals in pages from 1 to 300 are plotted logarithmically on the ordinate. Each cross on the scatter-diagram represents the occurrence of an interval whose "order in time" is indicated on the abscissa and whose size is measured logarithmically on the ordinate.

[8] Without pretending to a complete mathematical treatment of the question of the most suitable unit of interval-size (the treatment should await an empiric analysis), I suggest that the unit selected for the interval might decrease as f increases. The reader is reminded that we are not discussing the *average length of intervals* between the repetitions of words (which are of course inversely proportionate to the word's frequency, f) but to variations in the length of intervals in the $f - 1$ intervals of words occurring f times. In simplest form, as shown by Figure 3 *supra,* the variation is $N = log\ I_f$ for each interval.

According to my inspection of Figure 3 the crosses are distributed quite evenly over the scatter-diagram. That is, there is no systematic bunching of crosses at any particular interval or at any particular interval-size, although minor variations of negligible significance are present, as is not surprising. And this even distribution of the crosses means that the very long intervals of many scores of pages follow the same principle of scatter as the very short ones. The present scatter-diagram which was selected for presentation because it represents the highest frequency-class analyzed, as opposed to the lowest frequency-class of Figure 2, is typical of the scatter-diagrams of all the other frequency-classes of Table 1 (*mutatis mutandis*) whose presentation here is precluded by limitations of space.

In a certain respect the data of Figure 3 constitute the most important of those presented, since they not only illustrate the inverse relationship between N and I_f, but they also emphasize the enormously important point that *the various interval-sizes are distributed among the repetitions without favoritism to the order of that interval in time.* Thus for example we cannot argue from the sheer presence of a short interval in the stream of speech that it has occurred either early or late in the total occurrences of the respective word. Thus 100 pages may elapse before a word is used a second time; or it may be used a second time in the same sentence. For if the data of Figure 3 mean anything at all, they mean that *the different interval-sizes tend to be evenly distributed over time* (and not according to some quasi-scheme of perseveration or of "recency").

Theoretical Discussion

In observing that interval-sizes tend to be evenly distributed over time we may have reached a point of considerable theoretical importance for our understanding of linguistic process. Although this theoretical point together with its implications for the dynamics of living process in general is being treated bookwise[9] in considerable detail, it is perhaps not inappropriate to close the present writing with an outline of what the foregoing empirical observations may mean theoretically in terms of the dynamics involved. This we shall do by resorting to two related mechanical analogues which will be useful in suggesting that "time perspective" and a kind of "semantic balance" are fundamental in the type of speech-behavior which we find illustrated in Figure 3.

[9] *The Principle of Least Effort* now being prepared for publication.

1. *The Bell-Analogy and "Time Perspective"*

Let us take n bells that are equivalent in size and equally difficult to ring, and then let us attach them to a long straight board in such a manner that the bells are equally spaced along the board. At one end of the board we shall place a blackboard ruled with n-columns for the respective bells; and we shall also station a demon there to act as bell-ringer. The demon must ring one bell once each second of time, and after he has finished ringing a bell once he must return to the blackboard to record that fact in the bell's column. Thus in order to ring one bell 10 times, or 10 bells once each, he will make 10 round trips down the board and back in the space of 10 seconds, and will have 10 marks therefor on the blackboard. (And we shall ask the demon to make his round trips over shortest distances).

This analogue is interesting for many reasons. First of all the demon's work, w, *in terms of making a round trip to ring a given bell,* will increase in direct proportion to the bell's distance, d, from the blackboard (or $w = d$). And since the distance of the respective bells increases integrally from the blackboard (i.e., $1d, 2d, 3d, \ldots\ldots, nd$), it follows that the bells are arranged in respect of the the demon's work, w, in getting to and from them according to the simple series, $1w, 2w, 3w, \ldots\ldots, nw$.

Now if we ask our demon to ring each bell with a frequency, f, that is inversely proportionate to the round-trip work involved, or in equation form, $w \times f = C$, he will ring the closer (and easier) bells proportionately more often than the distant (and harder) bells. And since the ranked-frequency in decreasing order, r, with which each bell is rung will be equal to the bell's w above, we come upon the familiar equation:

$$(1) \qquad r \times f = C.$$

However if we now ask the demon to ring all bells according to Equation 1 but to stop after he has rung the nth and farthest bell once ($n = C$) and after he has rung all other bells their allotted times, then the n bells will have been rung approximately[10] according to the equation

$$(3) \qquad F \cdot Sn = \frac{F}{1} + \frac{F}{2} + \frac{F}{3} + \ldots\ldots + \frac{F}{n}$$

in which $F \cdot Sn$ represents the total of round trips made (as well as the total

[10] This equation is only approximate, since a bell can be rung only an integral number of times whereas the equation calls for fractional frequencies, hence the emergence of Equation 2 above, as discussed in (21).

number of running seconds of time) and where F represents the total number of times the nearest bell is rung, and where $\frac{F}{n} = 1$ (or, if you will, where $F = n$), wth p omitted above because it equals 1.

Of course the above equation puts no restriction upon the *order* in which the demon rings the bells. Thus he may ring the nearest bell its allotted F times before ringing the 2nd nearest bell its allotted $F/2$ times, and so on progressively down the board until he has rung the nth and farthest bell a single time. In short he might always ring "the easiest remaining bell first," while postponing as long as possible the more distant and hence more difficult bells. The chief drawback of ringing "the easiest first" is that the demon will be forced to run faster and faster, and therefore to work at an ever increasing rate, as he proceeds farther and farther down the board, if he is to complete each round-trip within the prescribed second. And in so doing he will be *unevenly distributing his work over time* with the risk of collapsing before he gets the nth bell rung.

In order to correct this uneven distribution of work over time, we may ask the demon to distribute his work as evenly as possible over time while still ringing his bells according to Equation 3. Yet as soon as he does distribute his work evenly over time, he will automatically ring the bells in such a way that the sizes of the interval, I_f, between the respective repetitions of the bells will approximate the equation:

$$(4) \qquad N^p \cdot I_f = \text{a constant}$$

with the exponent, p, equal to 1. The reason for this is that from second to second the demon will be counterbalancing the cumulative work, w, with the cumulative frequencies, f; that is, he will try to expend ½ his total work in each half of the $F \cdot Sn$ seconds, ¼ in each quartile, and $1/F$th in each Sn seconds. Furthermore this will mean that every time the demon rings a distant bell, whose w is large, he will have to ring a succession, or *cluster,* of nearer bells, whose w is small.

Indeed if we view the demon's entire activity as consisting of interspersing difficult bells with *clusters* of easier bells, we can perhaps most readily grasp why there will be proportionately more short intervals between repetitions than longer ones. For, to begin, we know that the larger the bell's w is, the rarer will its ringing be; by the same token, the longer the compensating *cluster* of smaller bells is (that is, the greater the number of pealings of easier bells, when multiplied by their work), the rarer that cluster's occurrence will be. And just as more distant bells and longer *clusters* will be proportionately rare, so too will easier bells and shorter *clusters* be proportionately more frequent.

Now since the *clusters* consist of the easier bells (that is, they consist of proportionately more easier bells), and since the easier a bell is, the proportionately more often it is rung, it follows that *within clusters* the bells will be rung with a high rate of repetition (that is, they will be rung with short intervals in between). Indeed there will be not only many short intervals between repetitions *within clusters,* but also proportionately so. Hence *within clusters* we may expect an approximation to the equation, $N \cdot I_f =$ *a constant.*

Of course the sizes of all intervals between repetitions are computed not only *within clusters* but also between *clusters.* However since the sizes of *clusters* tend to vary inversely in proportion to their number, it follows also that *between clusters* there will be proportionately more shorter intervals between repetitions than longer ones. Therefore in measuring the number, N, of interval-sizes, I (and therefore of I_f), between the ringings of the same bell (or any frequency-class of bells) we shall find an approximation to the equation:

(4*a*) $N \cdot I =$ *a constant,*

which is the more general statement of the equation, $N \cdot I_f =$ *a constant.*

And this will be true as aforesaid because our demon will be constantly counterbalancing the difficult but more rarely pealing bells at the further end of the board with the rapid repetition of the easier and more frequently pealing bells at the nearer end of the board. A statistical analysis could reduce the accumulation of marks on the blackboard to a scatter-diagram similar to that of Figure 3. If we gave each bell a distinguishing name and recorded each bell's name when rung, then the frequency-distribution of the succession of names would be approximately that of the succession of words in Joyce's *Ulysses.* And from the above equation, the other equations could be deduced (but not vice versa).

Of course other explanations of the working of the bell-analogy can be tendered. Thus the demon could ring the *n*th bell once and the 1st bell F times, and after that balance the rare but difficult pealings of the bells at the farther end with the frequent and easy pealing bells at the near end. No matter how the analogy is explained, however, the demon would be balancing the frequency of easy acts against the rarity of difficult acts so that during every Sn seconds he will expend as nearly as possible $1/F$th of his total work. Although upon first inspection the premium upon shorter rather than longer intervals between repetitions would seem to indicate a "law ol perseveration," such a conclusion does not seem to be necessary if we postulate an even distribution of work over time.

Upon the successful completion of his task the demon may be said to have revealed a 100 per cent time-perspective, or, as we shall say, a 1.00 time-perspective (referring to the equation, 4, or, $N^p \cdot I_f$ = *a constant* where $p = 1.00$). That will mean that the demon has understood and executed his problem as a *group problem,* seeing that every act influences every other act, while correctly assessing at all moments the influence of his past acts upon his present behavior, as well as that of his present behavior upon that of his future conduct in respect of the ordering of the ringing of the bells. *Time-perspective, then, in terms of our bell-analogy means not only the performance of acts with a frequency that is inversely proportionate to the work involved (with the expenditure of work minimized), but also the even distribution of all work over time.*

The difficulty of ringing the bells according to a 1.00 time-perspective can perhaps be best illustrated by briefly noting various types of imperfect time-perspective where the sole shortcoming is that of a faulty distribution of work over time. One such type would be that of the "easiest first" which we have already mentioned. In this instance, we remember, the demon would ring the 1st bell F successive times, then the 2nd bell $F/2$ successive times, and so on down the board until he had rung the nth bell once. Each successive bell will necessitate an increased rate of work in making the round trip, with the result that the demon will have to expend the same amount of work during the last, $F \cdot Sn$th second in getting to and from the nth bell a single time as he spent during the first F seconds in getting to and from the 1st bell F times. Mathematically the sizes of all intervals between the repetitions of bells will be 1 round-trip; hence the slope of his Number-Interval distribution will be 0; and we might even refer to this condition as one of .00 time-perspective since he distributed his work over time with minimal perspicuity. Of course, whether with .00 time-perspective or with 1.00 time-perspective, the demon will expend the same actual amount of total work in ringing the n bells according to the equation of the harmonic series. But in the case of .00 time-perspective the uneven distribution of work would lead to a cyclical rate of work-expenditure which would be absent in the 1.00 time-perspective. This cyclical rate of work-expenditure would have its minimum at the beginning and would rise to a maximum at the $F \cdot Sn$th second only to drop to a *minimum* at the $F \cdot Sn + 1$st second, and so on, as the demon rings the bells day in and day out. This cyclical rate of work-expenditure (which would not appear with 1.00 time-perspective, no matter how long the demon rang the bells) I should like to be permitted to designate as *cyclothymic unbalance* with a view to a future treatment of the same in greater detail.

Another type of abnormal time-perspective (with normal = 1.00) is perhaps that represented by the median 1.20 slope of Joyce's *Ulysses* which, according to our present theoretical analysis, suggests a slightly abnormal preference for longer intervals (if for the sake of argument we ignore the errors of Col. VI, Table 1, and also the fact that one page has been added to each interval). Thus having once "rung a bell," Joyce tends systematically to avoid its repetition abnormally. In other words, events of the past (as represented by words) seem to be systematically more remote from the present than is actually the case with 1.00 time-perspective. Although this general type of over-long time distortion is probably not infrequent among those personalities who focus their attention primarily upon the present moment, it is interesting to note that this paricular distortion of time is found in a novel that is characterized for just that attribute (if we may so interpret the words, "stream of consciousness" writing).

Other types of time-perspective—and not necessarily linear—can be defined in terms of the bell-analogy, yet there is one we mention cursorily lest it be ignored. We refer to the case in which the demon saves work and simplifies the problem of distributing his work evenly over time by simply bending the straight board into a quasi-arc. In this fashion the distant bells become nearer, and the demon can take short-cuts to them. This type of time-distortion we shall call *schizophrenic unbalance* and we shall treat it in greater detail in a future publication.

Time-perspective, in terms of the distribution of minimalized work over time (with all its endless ramifications) would seem to be an inviting topic for the study of the normal and abnormal of human mental behavior.

2. *The Tool-Analogy and "Semantic Balance"*

Although the bell-analogy has the virtue of illustrating mechanically our equations for the distribution of words, nevertheless its shortcomings should not escape us. One obvious shortcoming is its *rigidity* which becomes apparent when we remember that the repetitions of bells are supposed to be analogous to the repetition of words. And by *rigidity* we mean the fact that the spacing (and hence frequencies) of our bells cannot be altered, and that the demon cannot "change his job." If the board of bells were a perfect analogy to the usage of a vocabulary of words, then a given speaker would not only have to talk but would have to use his fixed vocabulary of words with fixed meanings with fixed frequencies.[11]. Yet we know that in practice a vocabulary shifts in size and content while its verbal entities are constantly subject to changes in form and in *form* and in *meaning* (or, as we may say, to *linguistic changes* and *semantic changes*).

In order to avoid the rigidity of the bell-analogy let us transpose the arrangement of bells into a corresponding arrangement of tools on a straight board that extends out in front of an artisan (our erstwhile demon). To this artisan we now give the following injunction: Perform jobs with tools with a maximum economy of work, with no restrictions placed upon the jobs performed or upon the tools used except that work must always be minimized. If the artisan complains about the one-dimensional board for his tools, he will remember that it is to correspond to the one dimensionality of speech which has no above or below, no right and no left (25, p. 256 f.); besides the problem would be essentially the same if cast in terms of two or three dimensions (24).

Now for the sake of getting our artisan started we shall give him an order for a quantity of like artifacts to be fabricated with tools with least work. And at once our artisan will find it economical to place the most frequently used tools nearest to him in order to minimize the work of reaching for them. In general [and for reasons set forth elsewhere (22, Chap. 3)] he will find it economical to arrange all tools along the board in such a manner that *the sum of the products.of the frequency of usage (f) of all tools, when multiplied by their mass (m) and by their distance from him (d) will be a minimum.*[12] In brief, the arrangement will be such that the sum of

[11] Some evidence in support of such a fixed rate (Chapple's "Interaction-Rate" essentially) of verbal usage can be found, I think, in the empiric observations of conversation-lengths made by Chapple (4, p. 10-16), as Chapple argues extensively in his excellent writings. In the *Principle of Least Effort* the attempt will be made to reconcile Chapple's findings with those of the present writing by referring both to the *inertia of jobs and tools* of the tool-analogy (*infra*), according to which it is economical to use all tools with frequencies and in combinations as determined by the arrangements pre-established in *semantic balance* as a result of past economical adjustments to past jobs.

[12] For the sake of simplicity we shall assume a constant friction, μ, that is directly proportional to distance, d, and that all work involved in using a tool once is equal to its $m \times d$ (thus ignoring for simplicity the tool's size, s, which however will also be subject to the "law of abbreviation" because the more voluminous the tools are, the less compactly they can be packed on the board, with the result that the farther—i.e., the greater the d—the artisan must in general reach, with a greater

the products of $f \times m \times d$ of all tools will be a minimum. Hence the comparative distance, d, of a tool from the artisan will depend upon the comparative smallness of its product $f \times m$.

The above minimal equation becomes suggestive when the workshop is viewed in the light of passing time, or, as we shall say, *in dynamic process*. For, since we have placed no restrictions upon the kinds of tools used, the artisan will find it economical to invent *easier* tools which, by definition, will be tools of lesser mass (m). Yet he should not concentrate his inventive abilities indiscriminately upon the n different tools; on the contrary he will find it economical always to concentrate upon those particular tools whose products, $f \times m \times d$, are above average, since they are the ones that consume an above-average amount of work. Let us call this *the alpha drive towards simplification*. Its net effect in dynamic process will be towards making all tools equal in mass (m), with the result that the tools of our tool-analogy will become ever more like the bells of our bell-analogy with all the attendant equations.

But the drive towards simplification does not stop with the equivalence in mass of all tools, since it is always economical to invent easier tools (provided the total work of invention and replacement is less than that of maintaining the older tool). Nevertheless once the tools are all approximately equivalent in mass, then it becomes economical for the artisan to concentrate his inventive abilities upon the most frequently used tools (*the beta drive towards simplification*) because an ounce clipped from a tool whose frequency is F will be equal in work-saved to a pound that is clipped from a tool whose frequency is $F/16$. The net effect in dynamic process of this *beta drive towards simplification* will be an inverse relationship between the mass of a tool and the frequency of its occurrence. This inverse relationship we shall call the *law of abbreviation* of tools. And by substituting the *length of a word* for the *mass of a tool,* we come upon the law of abbreviation of words (viz., "the length of a word tends to be inversely related to its frequency of usage).)[13]

But that is not all. The artisan may invent a new gadget for a particular

expenditure of work). This minimal equation above will be recognized as a variant of Maupertuis's principle of least-action (15) which with further postulates and elaboration will be extended to "mental phenomena" in my forthcoming *Principle of Least Effort.*

[13] Cf. (18, p. 67): "We have found evidence that differences in frequency even among words occurring less than two times in a million are related to differences in number of syllables or of phonemes."

task which has hitherto been performed by the combination of several tools (e.g., a fountain pen for an ink-well and pen). If the gadget saves total work, it should be adopted (the equivalence of a *neologism* in speech); an economical place should be accorded to it on the board, and the combination of tools it displaces should be discarded (the equivalence of *archaic words*). In such a fashion the new in tools or words displaces the old, with concomitant rearrangements of one or more tools on the board.

However the reverse of the above is also possible. Thus the artisan may find that a given permutation (or *pattern*) of different nearby tools when used together can perform a specialized task more easily than a specialized tool at greater distance (and out the latter will go). We shall call this *the urge towards the economical permutation of easier tools.* And this urge, which will be constantly present, will have a very curious result in dynamic process. For it will result in making the more frequently used tools also the more diversely used tools; whereas the less frequently used tools will tend to be the more specialized tools. In short there will be a direct relationship between the frequency of a tool's usage and the diversity of its usage. Translated into terms of words and their meanings—with a *word* equivalent to a *tool,* and a word's *meaning* equivalent to a specific *usage* of a tool in terms of jobs, we may expect to find *a direct relationship between the number of different meanings of a word and its relative frequency of occurrence.*[14]

Of course by now our tool-analogy has become much more refined than our bell-analogy. Nevertheless let us remember that our fundamental *alpha* and *beta drives towards simplification,* and the fundamental equation of the minimal sum of all products of $f \times m \times d$ will still be operative with the result that the artisan need only to distribute his work evenly over time in order to produce approximations to the equations developed in reference to the bell-analogy. Nor would the condition be altered if we intruded here a discussion of the Forces of Unification and of Diversification that control the size of n, as already discussed elsewhere.[15]

But instead of continuing in the present vein with our tool analogy, as the artisan proceeds to fill for us the now forgotten order of like artifacts that we gave to him to get him started, let us have him finish the order and run out of our jobs for his tools. Since he is obliged to use tools on

[14] In my opinion some interesting empiric support of this "Principle of Diversity of Meanings" can be found in the charts of Fries (9, p. 83-86). The Lorge-Thorndike semantic count when available should provide further valuable information if the number of different meanings is given for the various frequency-classes.

[15] The Force of Unification is called the Force of Repetitiousness in (19).

jobs, the artisan must look for further jobs. And in order to save the work of re-tooling, it will be economical for him to seek a job like the one he has just despatched since that is the kind of job for which his tools have become economically arranged and designed (the *inertia of jobs*—including verbal jobs, since we all prefer our own repertoire of clinchés). But failing there he will seek those available jobs which will entail the least amount of work of re-tooling, *in seeking jobs for his tools.*[16]

Now when he undertakes the task of re-tooling he will have several alternatives. First he can alter the form of pre-existent tools in order to bring them into conformity with their new usage. This we shall call *linguistic-semantic change* (25). Or he can preserve the old form in a new usage (e.g., use a hair-pin for a key) in which case the tool undergoes a *semantic change.* · Or he may re-design an old tool to perform an old usage more economically (*linguistic change*).

However the terms *linguistic* and *semantic changes,* or ·a combination thereof, for changes in *form* and in *meaning* respectively are of interest to us only because they introduce at long last the concept of *semantic balance.* And by *semantic balance* we mean: *the alteration (including the accession of the new and the elimination of the old) of the forms and usages of tools, as well as the alteration of jobs, in order to match tools with jobs and jobs with tools for the sake of minimizing the total work of survival.* Although this definition is offered without any restriction within the field of individual and social behavior of organisms (for reasons to be explained bookwise in detail), nevertheless in the terms of our tool analogy with specific reference to the form and meanings of words, a condition of *semantic balance* can conceivably be inferred to exist (for reasons already presented) from the emergence of a Number-Interval distribution in which the sizes of intervals between the repetition of words stands in an inverse proportionality to their number (or where the number of intervals of a given size is equal to the logarithm of the size approximately). In short *semantic balance* includes 1.00 time-perspective in which the rate of work-expenditure is constant.

But now that we have defined *semantic balance* which in the case of the repetition of words will appear as a recti-linear distribution, let us remember that during the periods of re-adjustment to a new situation (i.e., "re-tooling"), we may not expect recti-linear distributions. Indeed non-linear dis-

[16] Fundamental to the Principle of Least Effort is the completely relativistic postulate of *the reciprocal economy of tools and objectives,* according to which both tools and jobs are altered in order to become reciprocally more matched. This postulate is corollary to a more primary postulate to be discussed in the *Principle of Least Effort.*

tributions[17] may be quite instructive in studying the personality, for they may reflect the conscious or unconscious struggles of the personality in altering the tools and objectives of his life in order to minimize the work of survival.

At this point, however, a word of caution is very much in order out of fairness both to the reader and to the author. The author knows that "working models" and "mechanical analogues" for natural phenomena come and go in the world of science, and he appreciates that the present mechanical analogues need further extensive theoretical elaboration and empiric support before they can be accepted as being anywhere near correct. Yet in his defense the author replies that, in the light of the present great accumulation of unambiguous empiric correlations of high degree in the matter of the frequency-distribution of words, a serious beginning must now be made towards providing an interpretation of these correlations in terms of dynamic equilibria.[18] This article admittedly offers only a preliminary outline of such an interpretation; nevertheless it is being followed by a more extensive theoretical elaboration and empiric support in a separate publication, in which the analogues will be extended to the general problems of the form, function, and organization of behavior, both individual and collective. In this separate publication will be presented numerous sets of data for the evolution of children's speech, and for the speech-variants of psychotics,[19] two studies that are germane to the present article but which limitations of space preclude including.

C. Summary

1. In the present paper we have shown that within the restrictions stated and as specifically defined, the following equations which refer to the distribution of words in the stream of speech are mathematically related:

$$r \times f = C. \tag{1}$$

$$N(f^2 - \tfrac{1}{4}) = C. \tag{2}$$

$$F \cdot Sn = \frac{F}{1^p} + \frac{F}{2^p} + \frac{F}{3^p} + \ldots\ldots + \frac{F}{n^p} \tag{3}$$

(in the special case where $p = 1$, and where $F = n$).

$$N \cdot I_f = \text{a constant.} \tag{4}$$

We have presented empiric data here (or elsewhere) in illustration of these equations, the last of which refers to the *rate of repetition of words*. We

[17] The topic of non-linear distributions is broached (19).

[18] In this connection cf. the important contribution by E. G. Boring (2).

[19] Cf. (20) also (19); for bio-social dynamics cf. (22).

have shown that Equations 1, 2, and 3 can be derived from Equation 4, but not *vice versa*.

2. In the terms of the mechanical conditions of a bell-analogy we have tried to explain Equation 4 dynamically as representative of the even distribution of minimalized work over time. In this connection we have defined *time--perspective* in the terms of Equation 4, and have suggested certain types of pathological distortions of what we have defined as 1.00 *time-perspective* (e.g., *cyclothymic unbalance* is one such pathological distortion suggested).

3. By altering the rigid conditions of the bell-analogy to the more relativistic conditions of the tool-analogy in which "jobs seek tools and tools seek jobs in the reciprocal matching of all tools and all objectives for the sake of a most economical survival," we have attained approximations to the same equations as those of the bell-analogy. In terms of the tool-analogy we have also defined *semantic balance* together with the mechanisms of *linguistic changes* and *semantic changes* as devices for maintaining and for restoring *semantic balance*. Inherent in the tool-analogy is a law of *abbreviation of size*.

References

1. Ayres, L. P. Measuring Scale for Ability in Spelling. New York: Russell Sage Foundation, 1915.
2. Boring, E. G. Statistical frequencies as dynamic equilibria. *Psychol. Rev.*, 1941, **48**, 279-301.
3. Carroll, J. B. Diversity of vocabulary and the harmonic-series law of word-frequency distribution. *Psychol. Rec.*, 1938, **2**, 379-386.
4. Chapple, E. D. Personality differences as described by invariant properties of individuals in interaction. *Proc. Nat. Acad. Sci.*, 1940, **26**, 10-16.
5. Condon, E. V. Statistics of vocabulary. *Science*, 1928, **67**, 300.
6. Dewey, G. Relative Frequency of English Speech Sounds. Cambridge: Harvard Univ. Press, 1923.
7. Eldridge, R. C. Six Thousand Common English Words. Buffalo: Clement Press, 1911.
8. Estoup, J. B. Gammes Sténographiques. (5 ed.) Paris: 148 Ave. du Maine, 1917.
9. Fries, C. C. English Word Lists. Washington, D. C.: American Council on Education, 1940.
10. Hanley, M. L. Word Index to James Joyce's Ulysses. Madison, Wis., 1937.
11. Haugen, E. Norwegian Word Studies. (Vol. I.) Madison, Wis.: 1942.
12. Johnson, W. Language and Speech Hygiene. General Semantics Monograph, No. 1. Chicago: Institute of General Semantics, 1939.
13. Joos, M. Review. *Language*, 1936, **12**, 197.
14. Kosambi, D. D. On valid tests of linguistic hypotheses. *New Indian Antiq.*, 1942, **5**, 21-24.

15. DE MAUPERTUIS, M. Essai de Cosmologie. Paris: 1751.

16. SKINNER, B. F. The verbal summator and a method for the study of latent speech. *Psychol. Rec.,* 1937, **1**, 71-76.

17. THORNDIKE, E. L. On the number of words of any frequency of use. *Psychol. Rec.,* 1937, **1**, 397-406.

18. ———. Studies in the psychology of language. *Arch. of Psychol.,* 1938, No. 231, 58-67.

19. WHITEHORN, J. C., & ZIPF, G. K. Schizophrenic Language. *Arch. Neurol. & Psychiat.,* 1943, **49**, 831-851.

20. ZIPF, G. K. Children's speech. *Science,* 1942, **96**, 344-345.

21. ———. Homogeneity and heterogeneity in language. *Psychol. Rec.,* 1938, **2**, 347-367.

22. ———. National Unity and Disunity. Bloomington, Ind.: Principia Press, 1941.

23. ———. Observations of the possible effect of mental age upon the frequency-distribution of words. *J. of Psychol.,* 1937, **4**, 239-244.

24. ———. On the economical arrangement of tools; the harmonic series and the properties of space. *Psychol. Rec.,* 1940, **4**, 147-159.

25. ———. Psycho-Biology of Language. (2nd ed.) Boston: Houghton Mifflin, 1939.

26. ———. Reply to M. Joos. Language, 1937, **13**, 60-70.

27. ———. Selected Studies of the Principle of Relative Frequency in Language. Cambridge: Harvard Univ. Press, 1932.

28. ———. Unity of nature, least-action, and natural social science. *Sociometry,* 1942, **5**, 48-62.

Millbrook P. O.
Duxbury, Massachusetts