CHAPTER TWO

ON THE ECONOMY OF WORDS

As we turn now for the remainder of our study to a demonstration of the Principle of Least Effort, we should keep in mind certain general considerations that will be helpful in guiding our steps. For example we should remember that if Least Effort is indeed fundamental in all human action, we may expect to find it in operation in any human action we might choose to study. In short, any human action will be a manifestation of the Principle of Least Effort in operation, if this Principle is true; therefore all human action is potentially grist for our mill.

In the interest of economy we shall select for our own demonstration first those particular kinds of human action which will most readily admit of the disclosure of the underlying Principle. That is, we shall strive constantly to approach and study our hypothetical Principle from what seems to us to be its most accessible side. For a scientific demonstration can be likened to mountain-climbing—a task in which the mountaineer may either select a path of easiest ascent if he is eager to reach the top, or where he may choose a path of pronounced obstacles if he desires primarily to impress others with his skill. In this study we shall select what seems to be the path of easiest ascent.

Our path is the one that begins with a study of human speech as a set of tools. More specifically, it begins with a study of a vocabulary of words as a set of tools. The reason for selecting this as a beginning is, as we shall see, that the study of words offers a key to an understanding of the entire speech process, while the study of the entire speech process offers a key to an understanding of the personality and of the entire field of biosocial dynamics. Hence the contents of the present chapter will be of crucial importance for our entire study because in this chapter we shall untie a knot that we shall find duplicated again and again in other biosocial phenomena. The care and completeness with which we untie this first knot will render all future knots so much the easier to untie.*

I. IN MEDIAS RES: VOCABULARY USAGE, AND THE FORCES OF UNIFICATION AND DIVERSIFICATION

Man talks in order to get something. Hence man’s speech may be likened to a set of tools that are engaged in achieving objectives. True, we do not yet know that whenever man talks, his speech is invariably directed to the

*For the sake of simplification we shall use the term least effort in the present chapter to apply not only to situations of least probable work, but also to situations in which the argument is restricted to immediate behavior, which is technically one of least work.
attainment of objectives. Nevertheless it is thus directed sufficiently often to justify our viewing speech as a likely example of a set of tools, which we shall assume to be the case.

Human speech is traditionally viewed as a succession of words to which "meanings" (or "usages") are attached. We have no quarrel with this traditional view which, in fact, we here adopt. Nevertheless in adopting this view of "words with meanings" we might profitably combine it with our previous view of speech as a set of tools, and state: *words are tools that are used to convey meanings in order to achieve objectives.*

Yet once we say that words are tools, we broach thereby the question of the possible economies of speech; and as soon as we inquire into the possible economies of speech we remember that the sheer ability to speak at all represents an enormous convenience in present-day human social activity, whereas the inability to speak is a signal handicap. Since both the conveniences of being able to speak, and the handicap of being unable to do so, refer admittedly to the saving of effort, we may say that there is a *potential general economy in the sheer existence of speech,* in the sense that some human objectives are more easily obtained with speech than without it. The case is similar to that of a set of carpenter's tools whose sheer existence may be said to have a potential general economy for the carpenter.

But beyond this potential general economy of speech there are further possibilities for economy in the manner in which speech is used. For if speech consists of words that are tools which convey meanings, there is the possibility both of a more economical way, and of a less economical way, to use word-tools for the purpose of conveying meanings. Hence in addition to the general economy of speech *there exists also the possibility of an internal economy of speech.*

Now if we concentrate our attention upon the possible internal economies of speech, we may hope to catch a glimpse of their inherent nature. Since it is usually felt that words are "combined with meanings" we may suspect that there is latent in speech both a more and a less economical way of "combining words with meanings," both from the viewpoint of the speaker and from that of the auditor. *

From the viewpoint of the speaker (*the speaker’s economy*) who has the job of selecting not only the meanings to be conveyed but also the words that will convey them, there would doubtless exist an important latent economy in a vocabulary that consisted exclusively of one single word -a single word that would mean whatever the speaker wanted it to mean. Thus if there were $m$ different meanings to be verbalized, this word would have $m$ different meanings. For by having a single-word vocabulary the speaker would be spared the effort that is necessary to acquire and maintain a large vocabulary and to select particular words with particular meanings from this vocabulary. The single-word vocabulary, which reflects the *speaker’s economy*, may be likened to an imaginary carpentry kit that consists of a single

---

*Later we shall define a *meaning* of a word as a *kind of response* that is invoked by the word.* 2
tool of such art that it can be used exclusively for all the \( m \) different tasks of sawing, hammering, drilling, and the like, thereby saving the labor of otherwise devising, maintaining, and using a more elaborate toolage.

But from the viewpoint of the auditor (the auditor's economy), a single-word vocabulary would represent the acme of verbal labor, since he would be faced by the impossible task of determining the particular meaning to which the single word in a given situation might refer. Indeed from the viewpoint of the auditor, who has the job of deciphering the speaker's meanings, the important internal economy of speech would be found rather in a vocabulary of such size that it possessed a distinctly different word for each different meaning to be verbalized. Thus if there were \( m \) different meanings, there would be \( m \) different words, with one meaning per word. This one-to-one correspondence between different words and different meanings, which represents the auditor's economy, would save effort for the auditor in his attempt to determine the particular meaning to which a given spoken word referred.*

As far as the problem of words and meanings is concerned, we note the presence of two farreaching contradictory economies that relate in each case to the number of different meanings that a word may have. Thus if there are an \( m \) number of different distinctive meanings to be verbalized, there will be (1) a speaker's economy in possessing a vocabulary of one word which will refer to all the \( m \) distinctive meanings; and there will also be (2) an opposing auditor's economy in possessing a vocabulary of \( m \) different words with one distinctive meaning for each word. Obviously the two opposing economies are in extreme conflict.

We may even visualize a given stream of speech as being subject to two "opposing forces." The one "force" (the speaker's economy) will tend to reduce the size of the vocabulary to a single word by unifying all meanings behind a single word; for that reason we may appropriately call it the Force of Unification. Opposed to this Force of Unification is a second "force" (the auditor's economy) that will tend to increase the size of a vocabulary to a point where there will be a distinctly different word for each different meaning. Since this second "force" will tend to increase the diversity of a vocabulary, we shall henceforth call it the Force of Diversification. In the language of these two terms we may say that the vocabulary of a given stream of speech is constantly subject to the opposing Forces of Unification and Diversification which will determine both the \( n \) number of actual words in the vocabulary, and also the meanings of those words.

In adopting the term force to describe the two opposite economies that

---

*Nor does the word need to be spoken; it may also be written. The situation of the writer-reader is analogous to that of the speaker-auditor in respect of internal economies of usage of words, even though a reader is not so immediately present to a writer as an auditor is to a speaker and even though the word usage of written speech may differ somewhat from that of spoken speech for reasons that we shall scout in a later chapter. If we continue for the time being to discuss words without dichotomizing between written and spoken verbalizations, we do so in the interest of a legitimate simplification which seems to be justified at the beginning of our analysis of words and their usage, as we think the reader will agree upon reflection.
are hypothetically latent in speech, we must remember that the term refers to what people will in fact do and not to what they are at liberty to do if they wish. For we are arguing that people do in fact always act with a maximum economy of effort, and that therefore in the process of speaking-listening they will automatically minimize the expenditure of effort. Our Forces of Unification and Diversification merely describe two opposite courses of action which from one point of view or the other are alike economical and permissible and which therefore from the combined view-points will alike be adopted in compromise. From this it follows that whenever a person uses words to convey meanings he will automatically try to get his ideas across most efficiently by seeking a balance between the economy of a small wieldy vocabulary of more general reference on the one hand, and the economy of a larger one of more precise reference on the other, with the result that the vocabulary of \( n \) different words in his resulting flow of speech will represent a vocabulary balance between our theoretical Forces of Unification and Diversification.*

II. THE QUESTION OF VOCABULARY BALANCE

We obviously do not yet know that there is in fact such a thing as vocabulary balance between our hypothetical Forces of Unification and Diversification, since we do not yet know that man invariably economizes with the expenditure of his effort; for that, after all, is what we are trying to prove. Nevertheless-and we shall enumerate for the sake of clarity-if (1) we assume explicitly that man does invariably economize with his effort, and if (2) the logic of our preceding analysis of a vocabulary balance between the two Forces is sound, then (3) we can test the validity of our explicit assumption of an economy of effort by appealing directly to the objective facts of some samples of actual speech that have served satisfactorily in communication. Insofar as (4) we may find therein evidence of a vocabulary balance of some sort in respect of our two Forces, then (5) we shall find ipso facto a confirmation of our assumption of (1) an economy of effort. Therefore much depends upon our ability to disclose some demonstrable cases of vocabulary balance in some actual samples of speech that have served satisfactorily in communication.

Fortunately, if a condition of vocabulary balance does exist in a given sample of speech, we shall have little difficulty in detecting it because of the very nature and direction of the two Forces involved. On the one hand, the Force of Unification will act in the direction of decreasing the number of different words to 1, while increasing the frequency of that 1 word to 100%. Conversely, the Force of Diversification will act in the opposite direction of increasing the number of different words, while decreasing their

*We shall consistently capitalize the terms, Force of Unification and Force of Diversification, in order to remind ourselves that these Forces do not represent forces as physicists traditionally understand the term, but only the natural consequences of our assumed underlying economy of effort. Moreover our term balance will include what are technically known as steady states and the equilibria of the physicist and of the economist.
average frequency of occurrence towards 1. Therefore number and frequency will be the parameters of vocabulary balance.

Since the number of different words in a sample of speech together with their respective frequencies of occurrences can be determined empirically, it is clear that our next step is to seek relevant empiric information about the number and frequency of occurrences of words in some actual samples of speech.

A. Empiric Evidence of Vocabulary Balance

James Joyce's novel Ulysses, with its 260,430 running words, represents a sizable sample of running speech that may fairly be said to have served successfully in the communication of ideas. An index to the number of different words therein, together with the actual frequencies of their respective occurrences, has already been made with exemplary methods by Dr. Miles L. Hanley and associates who have quite properly argued that all words are different which differ in any way "phonetically" in the fully inflected form in which they occur (thus the forms, give, gives, gave, given, giving, giver, gift represent seven different words and not one word in seven different forms). 3

To the above published index has been added an appendix from the careful hands of Dr. M. Joos, in which is set forth all the quantitative information that is necessary for our present purposes. For Dr. Joos not only tells us that there are 29,899 different words in the 260,430 running words; he also ranks those words in the decreasing order of their frequency of occurrence and tells us the actual frequency, $f$, with which the different ranks, $r$, occur. By consulting this appendix we find, for example, that the 10th most frequent word ($r = 10$) occurs 2,653 times ($f = 2,653$); or that the 100th word ($r = 100$) occurs 265 times ($f = 265$). In fact, the appendix tells us the actual frequency of occurrence, $f$, of any rank, $r$, from $r = 1$ to $r = 29,899$, which is the terminal rank of the list, since the Ulysses contains only that number of different words.

It is evident that the relationship between the various ranks, $r$, of these words and their respective frequencies, $f$, is potentially quite instructive about the entire matter of vocabulary balance, not only because it involves the frequencies with which the different words occur but also because the terminal rank of the list tells us the number of different words in the sample. And we remember that both the frequencies of occurrence and the number of different words will be important factors in the counterbalancing of the Forces of Unification and Diversification in the hypothetical vocabulary balance of any sample of speech.

Turning to the quantitative data of the Hanley Index we can see from the arbitrarily selected ranks and frequencies in the adjoining Table 2-1 that the relationship between $r$ and $f$ in Joyce's Ulysses is by no means haphazard. For if we multiply each rank, $r$, in Column I of Table 2-1 by its corresponding frequency, $f$, in Column II, we obtain a product, $C$, in Column III, which is approximately the same size for all the different ranks and which, as we see in Column IV, represents approximately 1/10 of the
260,430 running words which constitute the total length of James Joyce's *Ulysses*. Indeed, as far as Table 2-1 is concerned, we have found a clearcut correlation between the number of different words in the *Ulysses* and the frequency of their usage, in the sense that they approximate the simple equation of an equilateral hyperbola:

\[ r \times f = C \]

in which \( r \) refers to the word's rank in the *Ulysses* and \( f \) to its frequency of occurrence (as we ignore for the present the size of \( C \)).

**TABLE 2-1**

<table>
<thead>
<tr>
<th>Rank (( r ))</th>
<th>Frequency (( f ))</th>
<th>Product of ( r ) and ( f ) (( r \times f = C ))</th>
<th>Theoretical Length of <em>Ulysses</em> (( C \times 10 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>2,653</td>
<td>26,530</td>
<td>265,500</td>
</tr>
<tr>
<td>20</td>
<td>1,311</td>
<td>26,220</td>
<td>262,200</td>
</tr>
<tr>
<td>30</td>
<td>926</td>
<td>27,760</td>
<td>277,800</td>
</tr>
<tr>
<td>40</td>
<td>717</td>
<td>28,680</td>
<td>286,800</td>
</tr>
<tr>
<td>50</td>
<td>556</td>
<td>27,800</td>
<td>278,800</td>
</tr>
<tr>
<td>100</td>
<td>265</td>
<td>26,500</td>
<td>265,000</td>
</tr>
<tr>
<td>200</td>
<td>133</td>
<td>26,600</td>
<td>266,000</td>
</tr>
<tr>
<td>300</td>
<td>84</td>
<td>25,200</td>
<td>252,000</td>
</tr>
<tr>
<td>400</td>
<td>62</td>
<td>24,800</td>
<td>248,000</td>
</tr>
<tr>
<td>500</td>
<td>50</td>
<td>25,000</td>
<td>250,000</td>
</tr>
<tr>
<td>1,000</td>
<td>26</td>
<td>26,000</td>
<td>260,000</td>
</tr>
<tr>
<td>2,000</td>
<td>12</td>
<td>24,000</td>
<td>240,000</td>
</tr>
<tr>
<td>3,000</td>
<td>8</td>
<td>24,000</td>
<td>240,000</td>
</tr>
<tr>
<td>4,000</td>
<td>6</td>
<td>24,000</td>
<td>240,000</td>
</tr>
<tr>
<td>5,000</td>
<td>5</td>
<td>25,000</td>
<td>250,000</td>
</tr>
<tr>
<td>10,000</td>
<td>2</td>
<td>20,000</td>
<td>200,000</td>
</tr>
<tr>
<td>20,000</td>
<td>1</td>
<td>20,000</td>
<td>200,000</td>
</tr>
<tr>
<td>29,899</td>
<td>1</td>
<td>29,899</td>
<td>298,990</td>
</tr>
</tbody>
</table>

The data of this table give clear evidence of the existence of a vocabulary balance. We must not forget that Table 2-1 contains only a few selected items out of a possible 29,899; hence the question is legitimate as to the possible rank-frequency relationship between the rest of the 29,899 different words.

Although we cannot easily present in tabular form the rank-frequency relationships of all these different words, we nevertheless can present them quite conveniently on a graph, because we know that the equation, \( r \times f = C \), will appear on doubly logarithmic chart paper as a succession of points descending in a straight line from left to right at an angle of 45°. And if we plot the ranks and frequencies of the 29,899 different words on doubly
logarithmic chart paper, and if the points fall on a straight line descending from left to right at an angle of 45° we may argue that the rank-frequency distribution of the entire vocabulary of the *Ulysses* follows the equation, \( r \times f = C \), and suggests the presence of a vocabulary balance throughout.  

As to the details of the graphical plotting of this particular equation (which will be repeated again and again throughout our study) we shall plot successive ranks from 1 through 29,899 horizontally on the \( X \)-axis, or abscissa. Then, in measuring frequency on the \( Y \)-axis, or ordinate, we shall give for each rank a dot which corresponds to the actual frequency of occurrence of the word of that rank. After we have completed our graphing of the actual frequencies of our 29,899 ranked words, we shall connect the dots with a continuous line in order to note whether the line is straight, and whether it descends from left to right at the expected angle of 45°.

In Fig. 2-1 we present in Curve A the data of the entire *Ulysses* thus plotted, and the reader can assess for himself the closeness with which this curve descends from left to right in a straight line at an angle of 45°. In order to suggest that the *Ulysses* is not unique in respect of a *hyperbolic rank-frequency word distribution*, we include gratuitously in Curve B of Fig. 2-1 the rank-frequency distribution of the 6,002 different words in fully inflected form as they appear in a total of 43,989 running words of
combined samples from American newspapers as analyzed by R. C. Eldridge. Curve $C$ is an ideal curve of 45° slope that has been added to aid the reader’s eye.

Clearly the curves of Fig. 2-1 conform with considerable closeness to a straight line with the expected slope of 45°, except for the emergence of "steps" of progressively increasing size as the line approaches the bottom. Although we shall shortly see that these "steps" result from integral frequencies and are governed by the equation, $r \times f = C$, we may now only say that the data confirm our equation merely down to where the "steps" begin. However, we note that an extension of the straight line through the "steps" would in most cases cut them fairly squarely through the middle (for reasons to be explained later), and that therefore the "steps" are by no means capricious in occurrence but have an orderliness of their own that is clearly not unrelated to the orderliness of the straight line above.

B. The Significance of $r \times f = C$

Before discussing the reasons for the emergence of the "steps" in Fig. 2-1, let us dwell briefly upon the significance of the curves themselves which clearly show that the selection and usage of words is a matter of fundamental regularity of some sort of an underlying governing principle that is not inconsistent with our theoretical expectations of a vocabulary balance as a result of the Forces of Unification and Diversification.

Perhaps the easiest way to appreciate the fundamental regularity exhibited by our curves is to ignore for the moment how they do appear and to inquire instead how they might appear if no underlying governing principle were involved. In short, let us inquire into the various ways that a rank-frequency distribution both could, and could not, appear from the particular manner in which we are plotting the data so that we may see how remote the probabilities are of their conforming to the rectilinear distribution we have observed.

In the first place, since we are ranking the words from left to right in the decreasing order frequency, it is evident that the line that connects the succession of dots can at no point bend upwards, since an upward bend at any point would indicate an incorrect ranking of the data according to decreasing frequencies. On the other hand, the line can and, in fact, will proceed horizontally whenever adjacent ranks have precisely the same frequencies (as happens to be the case with the horizontal lines of the "steps" at the bottom of the curves of Fig. 2-1, as we shall presently see). Hence we may predict in advance that any rank-frequency distribution may never slope upward from left to right although it may be horizontal. But that is not all. We may also predict that a rank-frequency curve will never bend downwards in a true vertical, since the line must pass from left to right in order to connect the dots of adjacent ranks. The apparently vertical lines of the "steps" of Fig. 2-1 are not truly vertical, since they do in fact connect adjacent dots. On the other hand, as long as the line never becomes a true vertical, it can bend downwards with any slope at any point.

As far as our method of plotting our data is concerned, we may say in
advance that the line proceeding from left to right in a rank-frequency distribution may twist and turn at any point on the graph paper as long as it never bends upwards and never bends downwards in a true vertical. In this connection the reader might take a pencil and paper and draw lines of various configurations and contortions that connect the upper left-hand corner with the lower right-hand corner-linelines that avoid upward bends and true verticals-in order to assure himself of the vast number of possibilities that lie within the restrictions of our method of plotting. After completing his "random lines" the reader will appreciate the orderliness of the lines of Fig. 2-1; and he will see how this orderliness points to the existence of a fundamental governing principle that determines the number and frequency of usage of the words in the stream of speech, regardless of whether or not the speakers and auditors are aware of the existence of the principle, and regardless of whether or not our Forces of Unification and Diversification in vocabulary balance provide a necessary explanation of it. Since all the words of Fig. 2-1 had "meanings" in their respective samples, the reader may infer from the orderliness of the distribution of words that there may well be a corresponding orderliness in the distribution of meanings because, in general, speakers utter words in order to convey meanings.

III. THE ORDERLY DISTRIBUTION OF MEANINGS

Taking a temporary leave of the distribution of words in Fig. 2-1, let us now turn our attention to the question of the distribution of the meanings of words. We have previously argued that under the conflicting Forces of Unification and Diversification the number of different meanings to be verbalized will be distributed in such a way that on the one hand no single word will have all different meanings and that on the other hand there will be fewer than different words. As a consequence, we may expect that at least some words must have multiple meanings. There remains then the problem of determining, first, which words will have multiple meanings and, second, how many different meanings these words of multiple meaning will have. In the solution of this problem, the Forces of Unification and Diversification will stand us in good stead.

Let us begin by turning our attention to the most frequently used word in the stream of speech, with special reference to the actual samples of Fig. 2-1. We shall arbitrarily designate the frequency of this most frequent word with the letter, \( F_1 \). The question now remains as to the number of different meanings which are represented by \( F_1 \). And here we may say that, regardless of the size of \( m_1 \), if we multiply \( m_1 \) by \( f_1 \), which represents the average frequency of occurrence of the \( m_1 \) meanings, we shall obtain \( F_1 \) since \( F_1 \) is made up of the total frequencies of its different meanings. Therefore we may write:

\[
m_1 \times f_1 = F_1
\]

With this simple equation in mind, let us recall our previously discussed Forces of Unification and Diversification and inquire into their respective
influences upon the sizes of \( m_1 \) and \( f_1 \). Obviously, the Force of Unification which theoretically acts in the direction of putting all different meanings behind a single word will tend to increase the size of \( m_1 \) at the expense of the size of \( f_1 \). On the other hand, the Force of Diversification which theoretically acts in the direction of reducing the number of different meanings per word will tend to increase \( f_1 \) at the expense of \( m_1 \). Therefore the respective sizes of \( m_1 \) and \( f_1 \) of our previous equation will again represent the action of the opposing Forces of Unification and Diversification.

Of course, we do not know a priori what the comparative strength of these two Forces may be. Yet we have observed from the data of Fig. 2-1 that there is a hyperbolic relationship between the \( n \) number of different words in the samples and their respective frequencies of occurrence. Therefore we may suspect that our two Forces of Unification and Diversification stand, in general, in a hyperbolic relationship to one another, with the result that \( m_1 \) and \( f_1 \) will also stand in a hyperbolic relationship with one another, with the further result that \( m_1 \) will tend to equal \( f_1 \).

However if \( m_1 \) equals \( f_1 \) and since \( m_1 \times f_1 = F_1 \), then clearly \( m_1 \) will equal the square root of \( F_1 \), or \( \sqrt{F_1} \).

But now let us note that the above argument will apply mutatis mutandis to the \( m_r \) number of different meanings of the word whose comparative frequency of occurrence is \( F_r \), with the result that the following simple equation may be expected:

\[
m_r = \sqrt{F_r}
\]

This simple equation is of interest, for it means that if (1) we make a rank-frequency distribution of the words of a sample of speech, as was done for the Ulysses and Eldridge data of Fig. 2-1, and if (2) we find that this distribution yields the straight line of an equilateral hyperbola as found in Fig. 2-1, then (3) we may conclude from the nature of the above argument and equation that a rank-frequency distribution of the different meanings of those words on doubly logarithmic paper would yield a straight line descending from left to right to the point, \( X = n \), yet intercepting only \( \frac{1}{2} \) as much on the Y-axis as on the X-axis (that is, it will have what is technically called a negative slope of \( \frac{1}{2} \), or of .5). The reason for this is that the \( m_r \) number of different meanings for each of the \( r \)-ranked words will be represented on doubly logarithmic paper by a point that is in each case \( \frac{1}{2} \) of the \( F_r \) of the respective ranked words. We shall call this the theoretical law-of-meaning distribution.

To determine empirically whether this theoretical law-of-meaning distribution exists, we could take the data of Fig. 2-1 and, after consulting a suitable dictionary, we could graph the \( m_r \) number of different meanings for each \( r \) different word, and note the resulting meaning-frequency distribution. The resulting meaning-frequency distribution would refer only to the particular Ulysses and Eldridge word-frequency distributions, and therefore would lack a more general applicability.

It would be of more general applicability and equally valid for our purposes if we selected the more comprehensive word-frequency distribution of
English as made and published by E. L. Thorndike on the basis of a count of 10 million running words. Although Dr. Thorndike has published only the 20,000 most frequent words of his count, nevertheless these 20,000 words will represent the average frequencies of standard English better than the particularized vocabularies of the data of Fig. 2-1. It is true that Dr. Thorndike has for the most part ignored the inflectional endings of words; instead he has subsumed the frequencies of occurrence of practically all different inflectional forms of a given word under the dictionary form of that word (i.e., he used what is technically known as a *lexical unit*); however we have no reason to suppose that any "law of meanings" would be seriously distorted if we concentrated our attention upon *lexical units* and simply ignored variations in number, case, or tense. Nor need we be disturbed by the fact that Dr. Thorndike did not list the actual frequencies of the different words but merely noted the 1st thousand most frequent, the 2nd thousand most frequent, and so on down through the 20th thousand most frequent, with a further notation of whether a given word of the first 5000 words was among the first or second 500 words of its respective thousand. This lack of a precise numerical notation—far from invalidating his count—offers a genuine challenge to our thesis. For (1) if we are correct in generalizing upon the data of Fig. 2-2 by stating that the distributions are representative of English, and (2) if our theoretical *law-of-meaning distribution* be correct, then we may suspect, both (3) that Thorndike’s 20,000 words would follow a hyperbolic rank-frequency distribution of words and (4) that the distribution of meanings of the 20,000 words when plotted on doubly logarithmic graph paper will yield a negative slope of .5 as previously explained. Therefore we may test our theoretical law-of-meaning distribution by turning directly to an analysis of the average $m$ number of different meanings per word in each of the 20 successive sets of one thousand words.

Fortunately for the analysis of the meanings of the 20,000 words, we have available the *Thorndike-Century Dictionary* which selected the $m$ different meanings to be presented for each word (*except for the 500 most frequent*) on the objective basis of Dr. Irving Lorge’s *The English Semantic Count*. Hence the $m$ number of actually used different meanings for each word in the dictionary has been determined empirically, with the result that in making our meaning-frequency analysis we need not fear including archaic or obsolescent meanings which might well distort our distribution.

Thanks to the help of some of my students, who undertook the task of noting the number of different meanings in Thorndike’s dictionary for each of the 20,000 words of the list, we present in Fig. 2-2 the average number of meanings per word (on the ordinate) for each successive set of 1000 words on the abscissa. Since the *average number of meanings per word* in each thousand refers in fact to the 500th word (or class-middle) of each thousand, the points on the abscissa represent these class-middlees in all cases; that is, they represent the values of the 500th, 1500th, 2500th, … 19,500th words respectively.

A glance at the data of Fig. 2-2 suffices to show that the points descend
in a strikingly straight line which is not far off from our theoretically expected negative slope of .5 (i.e., -.5). If we calculate by least squares the slope of the best straight line through the points, we arrive at the value -.4605 (± .0083) with the Y-intercept at 18.05 (antilog). This calculated value is not far off from our expected -.5 slope.

The approximation may be even closer than that if we remember that The English Semantic Count was not used for the 500 most frequent words (whose differentiation of meanings is truly difficult for reasons that will be apparent in our following chapter). Because of this consideration the first point at the left of our chart is suspect. If we ignore it and recalculate the slope for the remaining 19 points we have a slope of -.4656 (± .0027), which is slightly nearer to the expected -.5 slope.

Fig. 2-2. The meaning-frequency distribution of words.

If we turn our attention now to the 10 successive sets of 500 words which constitute the 5000 most frequent words in the list, and if we again ignore the suspected first 500 words for reasons already presented, we have a slope of -.4899 (± .0030), with which we may scarcely quarrel as an approximation of a -.5 slope.*

It is of course regrettable that additional sets of data on this important point are not available. Nevertheless the results of even this one study are so striking that pending the future findings of empiric analysis we are not rash in concluding that a law-of-meaning distribution exists according to which the m average number of meanings per word of a thousand words (when ranked in the order of decreasing frequency) will equal the square

*This slope is probably the most reliable, since it refers to the most frequent 10,000 words that are likely to be found in an optimum sample of 100,000 running words. For a discussion of an optimum sample see below.
root of the average frequency of the words' occurrence (or will decrease according to the square root of the rank).

Although later we shall again return to the entire question of the "meanings of words" with the problem of defining the term meaning, we may even now feel that our theoretical Forces of Unification and Diversification have led us to the empiric disclosure not only of a simple equation for the distribution of words (in the form, \( r \times f = C \), with \( r \) an integer) but also of a simple equation for the distribution of the meanings of those words which may be put down in the form of the equation, \( m_r = \sqrt{F_r} \), in which \( F_r = C_r \).

Incidentally, the fact that we have no actual rank-frequency distribution for the 20,000 words of Dr. Thorndike's frequency list does not invalidate our above conclusion; on the contrary, we shall present so many word-rank-frequency distributions in our following pages that the reader will be more than ready to believe that, if we had a rank-frequency distribution of the 20,000 most frequent words of the Thorndike analysis, it would probably be rectilinear like those of Fig. 2-1, at least for the first 10 or 12 thousand most frequent words.

With the assurance of the law-of-meaning distribution of Fig. 2-2, let us return now to a study of the significance of the rectilinear distributions of Fig. 2-1, which descend with a negative slope of 1 except for the "steps" of increasing magnitude at the bottom, and which indicate the existence of a vocabulary balance.

**IV. THE INTEGRALITY OF FREQUENCIES**

Some pages back we implied that the "steps" at the bottom of the two curves of Fig. 2-1 emerge as a natural consequence of the fact that words cannot occur with fractional frequencies. Since "steps" of this sort will be found in the graphs of following chapters, we may profitably digress at this point to show the more precise relationship between the sizes of these "steps" and the equation, \( r \times f = C \).

As far as the "steps" themselves are concerned, the upright lines (or "risers") are of no significance, since they merely connect the last dot of the horizontal line above with the first dot of the horizontal line below, and have been added to aid the reader's eye. On the other hand, the horizontal lines (or "treads") of the "steps" are quite significant, since they represent the number of different words, or ranks, that occur with the same frequency. Thus the horizontal line of Curve A that hugs the abscissa represents the 16,432 different words that occur one time in the *Ulysses*; the horizontal line of the next "step" above represents the 4,776 words that occur twice; the third horizontal represents the 2,194 words that occur three times, and so on up until the horizontal lines become so short that they no longer appear.

The reason for these horizontal lines becomes obvious if we remember that the frequency with which a word occurs can only be integral (or a whole number) in any actual sample of speech. Thus, in any actual sample
of speech, a word can occur 1, 2, 3 or some other whole number of times but not \(1\frac{1}{2}, 2\frac{1}{3}, 3\frac{1}{4}\) or any other fractional number of times. Yet since our equation, \(r \times f = C\), as stated, will theoretically necessitate the fractional occurrences of words which are impossible in practice, it is quite evident that there will be no actual rank-frequency dots on those portions of the descending diagonal line which lie between integral frequencies (e.g., between 2 and 3). Instead it is quite conceivable that those frequencies which theoretically should occur with fractional frequencies will actually occur with the nearest integral frequency. That is, for example, those words which theoretically should occur from 2\(\frac{1}{2}\) to 3\(\frac{1}{2}\) times, according to the equation \(r \times f = C\), will actually occur 3 times, since 3 is the nearest integral frequency. And that would also mean that the straight diagonal line above, if extended to the bottom, would cut approximately through the center of the horizontal lines of these steps, as an inspection of Fig. 2-1 shows is more or less the case. Hence the emergence of these "steps" is merely a natural consequence of the equation, \(r \times f = C\), in terms of integral frequencies.

Indeed the equation, \(r \times f = C\), which is responsible for the emergence of the "steps," will also determine both the positions and the sizes of the various "steps," as we shall now illustrate.

For it can be shown (and has been shown elsewhere\(^{10}\)) that the \(N\) number of different words of the same \(f\)-integral frequency of occurrence (under the conditions of the equation, \(r \times f = C\)) will be inversely proportionate to the square of their frequency (approximately) -or, stated somewhat more precisely in equation form, that:

\[N \left( f^2 - \frac{1}{4} \right) = C\]

However instead of interrupting our demonstration in order to derive this new equation mathematically from the hyperbolic equation, \(r \times f = C\), let us simply argue that if this new equation is true, we may expect to find in Joyce's Ulysses that when we multiply the \(N\) number of actual words of like frequency of occurrence by the square of that frequency minus the constant \(\frac{1}{4}\) [that is, when we multiply \(N\) by \((f^2 - \frac{1}{4})\)], the resulting products will approximate the constant, \(C\). And since the necessary data for this multiplication are available, let us proceed to it forthwith.

In Table 2-2 we present the actual products of our multiplication of the above factors for 15 arbitrarily selected frequency classes in the Ulysses (Column II); and also gratuitously for four Latin plays by Plautus (Column III), which we add for good measure to suggest that this number-frequency relationship is not uniquely characteristic of English.\(^{11}\)

Inspecting in turn each of the two columns, II and III, of Table 2-2, we find that the calculated values of each column are approximately the same for the entire column, with the exception of the lowest two frequencies of the Ulysses. This means that both the Ulysses and the four plays of Plautus have approximately the number of different words of like frequency for the selected frequency-classes that we should expect theoretically on the basis of our equation. And that means in turn, more specifically for the Ulysses, that the horizontal lines of the "steps" of Fig. 2-1 are approximately
ON THE ECONOMY OF WORDS

of the right length to satisfy the equation \( r \times f = C \) in the sense that the words which theoretically would have a fractional frequency do in fact have the nearest integral frequency.*

Since a more complete presentation of the above relationship is possible on a graph, we present in Fig. 2-3 the data for the frequency-classes from 1 through 50 for (A) the *Ulysses*, with each circle indicating the \( N \) number of different words (measured logarithmically on the abscissa) of

### TABLE 2–3

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

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

like frequency, \( f \) (measured logarithmically on the ordinate). For full measure, we add gratuitously a set of data (B) for the same frequency-classes for a non-English language this time for the Homeric Greek of the *Iliad* (plotted in black dots) as determined from Prendergast's *Concordance* thereto by the patient and careful hands of my former student,

*The size of \( C \) of Column II of Table 2-2 should be the same as the size of \( C \) in Table 2-1 which was about 26,000. The slight difference between the two is ascribable to the fact that there is a very slight bend at the top and bottom of the *Ulysses* curve.
ON THE ECONOMY OF WORDS

Dr. Harold D. Rose. (The Plautine material has already been published in full elsewhere.)

Inspecting the two curves of Fig. 2-3, we note that in each case they descend from left to right, on the whole, in such a fashion that the X-intercept is twice that of the Y-intercept, as is to be expected from the exponent, 2, in \( N(f^2 - \frac{1}{4}) = C \).* Dr. Joos informed me privately (September 16, 1937) that the calculated exponent of the Ulysses data falls between 1.99 and 2.01! According to his calculation (Language XII [1936] 199) the least-square exponent for the Plautus data of Table 2-2 is 1.98. These are remark-

![Graph showing number-frequency relationship of words.](image)

**Fig. 2-3. The number-frequency relationship of words. (A) Homer's Iliad; (B) James Joyce's Ulysses.**

ably close approximations to our theoretically expected 2. My least-square calculation of the Iliad data of Fig. 2-3 is 2.15 ± 0.01.**

In view of the orderly nature of the "steps" that emerge in a rank-frequency distribution of words, we may say that the two curves, A and B, of Fig. 2-1 closely approximate the equation of an equilateral hyperbola. Moreover, the gratuitously added information about the number, \( N \), of words of like frequency in the lower frequency-range of other languages suffices to assure us that "vocabulary balance," in the sense of a general orderliness of word distributions, is not peculiar to English. In our following chapter we shall present further rank-frequency distributions from samples of many other languages, as we return again and again to a consideration of the Forces of Unification and Diversification. For the time being, however, we shall study the special case of English, which will also tell us much about the nature of vocabulary balance in general.

*We remember that the difference between \( f^2 \) and \( f^2 - \frac{1}{4} \), is important only for very small values of \( f \). Thus when \( f \) is even as large as 4, the difference between \( f^2 \) and \( f^2 - \frac{1}{4} \) is only a difference between 16 and 15 \( \frac{3}{4} \). With \( f = 10 \), the difference is between 100 and 99 \( \frac{3}{4} \).

**Since Prendergast does not give the frequencies of the most frequent words in the Iliad we cannot plot its rank-frequency distribution.
V. THE INTEGRALITY OF RANK

The rank that a word has must also be an integral number. Thus there may be a 1st, 2nd, or 3rd rank, but not one that is $\frac{1}{2}$, $\frac{1}{3}$, or $\frac{2}{5}$ or any other non-whole number. This fact leads us to a very curious consideration, since our equation, $r \times f = C$, makes no provision for the integrality of $r$. Indeed, as far as this equation is concerned, both $r$ and $f$ may take on all values from positive to negative infinity which, in terms of speech, would be unthinkable. Hence the equation of the equilateral hyperbola is inadequate as a final description of what we have found. Let us now find a more adequate equation.

A. The Equation of the Harmonic Series

If we turn back to Table 2-1 and inspect the items of Column IV under the heading of "The Theoretical Length of the Ulysses" we note that the various items in this column approximate 260,430, which is the total number of running words in the novel. That is, when we multiply by 10 the constant, $C$ (of $r \times f = C$), we arrive at the approximate total number of running words. That fact leads us now to an obvious and instructive consideration about ranks and frequencies which we have hitherto ignored.

In ranking the words of Joyce's *Ulysses* in the decreasing order of their frequencies, we have ranked them according to the simple integral series, $1, 2, 3, \ldots n$, in which $n$ represents the terminal rank of the 29,899th and last different word in the *Ulysses* sample.

In observing that the product of a word's rank, $r$, when multiplied by its frequency, $f$, is a constant, $C$, (according to the equation $r \times f = C$), we may conclude at once that the different frequencies, $f$, of these ranked words will decrease in the order of the proportions of the following simple harmonic series:

$$1, \frac{1}{2}, \frac{1}{3}, \ldots \frac{1}{n}$$

since every frequency, $f$, when multiplied by its rank, $r$, will yield a constant.

Now this harmonic series is interesting. For if we say that $F$ represents the actual frequency of the most frequent word (i.e., the word whose $r = 1$), then it follows that $F/r$ will be the actual frequency of any word of rank, $r$. Thus, for example, the 10th most frequent word (i.e., $r = 10$) will have the frequency $F/10$, and the $n$th most frequent word (i.e., $r = n$) will have the frequency, $F/n$.

Indeed, we may view the entire *Ulysses* as consisting of the following approximate sum of the $n$ number of different words in terms of $F$ and ranked in the order of decreasing frequency:

$$Ulysses = \frac{F}{1} + \frac{F}{2} + \frac{F}{3} + \ldots + \frac{F}{n}$$

an equation in which the denominators of the above fractions refer to the successive ranks of the respective words, and where $n = 29,899$.

But at this point a curious consideration arises. Since we may infer from Table 2-1 that the total number of running words in Joyce's novel
is approximately equal to 10 times the frequency, \( F \), of the word whose rank is one (i.e., \( 1 \times F \times 10 = \text{about} \ 260,000 \)), we may say that \( 10F \) is approximately the length of *Ulysses*. Indeed we may even make the following approximate equation about all the words in the *Ulysses*:

\[
10F = \frac{F}{1} + \frac{F}{2} + \frac{F}{3} + \ldots + \frac{F}{n}
\]

When we inspect this equation, we note that in the last analysis we have merely multiplied by \( F \) the underlying simple equation:

\[
10 = 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n}
\]

and that, in turn, means that the constant, \( 10 \), of Column IV of Table 2-1 represents nothing more than the sum of the \( n \) harmonically seriated fractions of the right-hand member of the above equation.

Now the \( n \) fractions of a harmonic series, when added together, will yield a sum. This sum of the \( n \) members of a harmonic series can be represented technically by \( S_n \) (which means the sum, \( S \), of the \( n \) members of the series). The size of \( S_n \) will obviously depend upon the size of \( n \) (i.e., the number of fractions added), with the result that we can calculate the approximate size of any \( S_n \) if we know the size of \( n \) (and vice versa).

But at this point we must be careful. Though the actual value, \( S_n = 10 \), may be significant for our understanding of *Ulysses*-and the same constant is found to be approximately descriptive of the Eldridge newspaper material of Fig. 2-1-nevertheless the particular value, \( S_n = 10 \), is of no vital significance for our understanding either of the harmonic series or of the equation, \( r \times f = C \), or ultimately of the Forces of Unification and Diversification or the Principle of Least Effort that assumedly lies behind them. The size of \( S_n \) merely tells us the size of \( n \) (the number of different words that are used). Values other than 10 would do equally well. Indeed in Chapter Four we shall find that the limited vocabularies of children have an \( S_n \) that is characteristically smaller than 10. Therefore, in order to avoid relating our Harmonic Equation to any particular value of \( S_n \), we shall present it in the following general form, which will suffice to describe many of our future sets of data:

\[
F \cdot S_n = \frac{F}{1} + \frac{F}{2} + \frac{F}{3} + \ldots + \frac{F}{n}
\]

Graphically, this preceding *Equation of the Harmonic Series* will appear as a succession of dots which, if connected by a line as in Fig. 2-1, will descend from left to right at an angle of \( 45^\circ \) (technically a negative slope of 1) like the curves of Fig. 2-1. The size of \( S_n \) merely helps to locate the point on the \( X \)-axis which the rectilinear curve will intercept: the larger the size of \( S_n \), the farther to the right the straight line will be located on the graph. Moreover, if we postulate that the above Harmonic Equation is *saturated* when the line intercepts equal distances on the \( X \)-axis and on the \( Y \)-axis (and this is important) then we may say algebraically that the above Harmonic Equation in saturated form refers to a condition in which \( F = n \).
or, if one will, in which \( \frac{F}{n} = 1 \). And that means, in turn, that the \( n \)th word occurs only a single time in the sample.

Now this Harmonic Equation in saturated form (hereinafter abbreviated to \( F \cdot Sn \)) offers a more faithful description of the approximate data of Fig. 2-1 than was the case with the hyperbolic equation, \( r \times f = C \). For the Harmonic Equation, \( F \cdot Sn \), tells us that \( r \) must be integral; and from it we can easily anticipate mathematically the "steps" that will emerge naturally from the fact that frequencies, \( f \), can only be whole numbers.

And so let us remember that whenever in the future we find a rank-frequency distribution which, when plotted as in Fig. 2-1, reveals a slope of -1 (i.e., with the \( Y \)-intercept equal to the \( X \)-intercept) then we may infer that the distribution is that of the saturated harmonic series with \( F = n \).

B. The Question of Sample Size and the Closure of Speech

Although we have already proceeded quite far in our analysis of the number and frequencies of occurrences of words, there remains one basic problem to which we shall now turn—the problem of the *optimum-size* of the sample of speech to be examined—that is, when the size of the sample is "just right" and not "too long" nor "too short." The nature of this problem can perhaps be most easily elucidated if we express it in terms of a simple mechanical tuning-fork analogy in which \( n \) different tuning-forks will represent our \( n \) different words, and where the frequencies, \( f \), of the different forks will correspond to the frequencies, \( f \), of our words.

Let us imagine that we have a long board on which are attached \( n \) number of different tuning-forks of such sizes that during a \( T \) interval of time they will all vibrate respectively according to the saturated harmonic equation, with \( F = n \), viz.:

\[
F \cdot Sn = \frac{F}{1} + \frac{F}{2} + \frac{F}{3} + \ldots + \frac{F}{n}
\]

Thus far the group-behavior of our tuning-forks during a \( T \) interval will correspond to that of a vocabulary of words in the stream of speech, if we ignore momentarily as unessential the fact that the tuning-forks represent a synchronous phenomenon in the sense that they all vibrate at once, whereas the stream of speech is a diachronous phenomenon in the sense that only one word can be uttered at a time. The chief didactic value of this mechanical analogue will appear when we vibrate the board for a \( t \)-interval that is either shorter than \( T \) or longer than \( T \).

If we vibrate it for a \( t \)-interval that is appreciably shorter than \( T \), then the resulting frequency-distribution of the group-vibrations will be far removed from that of the saturated harmonic equation, \( F \cdot Sn \), if only because the forks of lower frequencies will not have had time to round out their allotted few vibrations. Turning to the phenomena of words, we may remark that the same relationship will obtain also for an excessively short sample of speech—say 100 words—in which even some of the most fre-
quent words might conceivably have no opportunity to repeat, not to mention the more rarely occurring ones which might not be able to occur at all. From so short a sample, in which most words would occur only once at best, we could scarcely detect the fundamental harmonic seriation which would only emerge as the size of the sample of running words approached that of $F \cdot Sn$.

On the other hand, if we extended the interval of vibration of the forks beyond $T$ to $T + t$, we should overshoot the saturated series, $F \cdot Sn$. For although the summated vibrations during the $T$ portion of the $T + t$ interval would represent the proportions of our equation, $F \cdot Sn$, nevertheless the summated vibrations during the entire $T + t$ interval might well yield graphically a very bent and distorted rank-frequency distribution (although much will depend upon the size of $t$). And similarly with an $F \cdot Sn + a$ sample of running words.*

Hence the $F \cdot Sn$ sample-length of words, like the $T$ period of vibration of the tuning-forks, represents an optimum length in which the saturated harmonic equation will precisely reveal itself. Although we shall not tarry at this point to ask how a rank-frequency distribution might appear graphically if the sample deviates markedly from optimum size, we admit that it is extremely fortunate that early empiricists happened to examine samples of speech whose length approached optimum size. For if they had selected samples of very small or of very large sizes, the rectilinear rank-frequency relationship might have been disclosed only with great difficulty.**

Now that we have stressed the obvious—namely, that the harmonic equation that involves $F \cdot Sn$ running words will appear with close approximation only in samples whose lengths closely approximate $F \cdot Sn$ running words—let us proceed a step further. If the number and frequencies of words in the stream of speech are the resultants of the opposing Forces of Unification and Diversification (or of any other "forces" for that matter), then the *interval of $F \cdot Sn$ running words may well have dynamic meaning as a unit, in reference to which a given speaker may be said to talk "too little" or "too much." That is, if we explicitly assume that the harmonic seriation, $F \cdot Sn$, represents a fundamental principle that governs the number and frequency of usages of words in speech, then we can only conclude that a given speaker "naturally" selects both the topics of his conversations and the words with which he verbalizes them in such a way that the resulting frequency-distribution of his continuing stream of speech will meet the exigencies of our equation, $F \cdot Sn$, without "too little" or "too much" talk. And this, in turn, means that inherent in the stream of speech is a

*As the interval of vibration takes on multiple values of $T$, the distribution becomes rectilinear, with the curve rising and with the $n$th rank having a frequency equal to the multiple value of $T$. The same is true of the increasing lengths of samples of speech; the chance of finding one of optimal length, or of multiples thereof, is very small.

**The relationship between the $n$ number of different words in samples of different sizes with the same or different fundamental values of $Sn$ for optimum sizes has been studied theoretically and empirically with great ingenuity by Dr. J. B. Carrol (The Psychological Record, II [1938], no. 16. 379-386). However, in this connection much may depend upon the rate of repetition in a sample. See below.
dynamic unit which we might call a closure (or a cycle, or a rhythm) which may be defined roughly as a length of speech during which a particular group of verbal tools has completed its collective behavior once. What else this closure may signify we do not yet know; in fact we happened upon its presence only by noting the particular fashion in which words (and the tuning-forks of our analogy) are used. In Chapter Seven we shall find that the stream of speech can be organized in such a way that no closure is inherent in it, and that schizophrenic speech may almost be characterized by the absence of closure.

But the concept of a closure leads to an even more important consideration which comes to light once we ask the following question, whose answer we have already partly foreshadowed: Suppose you could procure only sub-optimum samples for analysis; would you then never be able to discern the underlying regularity in the number and frequency of usage of words?

The nature of this question, which ultimately refers to a rate of usage, can be elucidated by stating it briefly in the terms of our erstwhile board of tuning-forks, which we shall now cause to vibrate for a t-interval that is so short that the harmonic seriation will not become manifest. This done, we face the problem of detecting the underlying harmonic seriation of the entire board by studying the few vibrations of the several individual forks of high frequency that have had the opportunity to complete several vibrations, and which thereby provide us with our only information about the organization of the total board. From an inspection of these few but frequent vibrations we should find that the higher the rate of vibration of a given fork is, the more nearly its cumulative vibrations, \( f \), during the \( t \)-interval will approximate the proportions of the harmonic series:

\[
\frac{f}{1}, \frac{f}{2}, \frac{f}{3}, \frac{f}{4}, \ldots
\]

Whether we are able to deduce from these few harmonic vibrations that the entire board of forks was harmonically seriated would depend somewhat upon the length of the \( t \)-interval.

Turning now to the phenomena of speech, may we infer from the above mechanical analogue that the rate at which the words repeat in a very small sample will also give us a corresponding clue to the fundamental harmonic seriation of vocabulary usage? Obviously, we are not yet wise enough about the facts of speech to draw any such inference, for we know absolutely nothing about the rate with which words are repeated in the stream of speech, and therefore we may not assume that our tuning-fork analogue is a true analogy in this respect.

As far as the sheer harmonic equation, \( F \cdot Sn \), for words is concerned, the sole matter of consequence is the total frequencies of occurrences of the respective words and not their rate of occurrence, or the length of intervals between their repetitions. For example, it makes no difference to the saturated harmonic equation, \( F \cdot Sn \), whether the \( F/1 \) occurrences of the most frequent word are bunched together one right after another in the
stream of speech, or whether they appear once in every $Sn$ running words, or whether they occur according to some other scheme. For as long as the most frequent word occurs $F$ times in $F \cdot Sn$ running words, it satisfies the harmonic equation, no matter how its $F$ occurrences may be distributed over time. And *mutatis mutandis* with the remaining words of the series.

And so it becomes clear that the harmonic equation, $F \cdot Sn$—although obviously of great descriptive value for our verbal phenomena as far as it goes—is nevertheless not of final validity by itself as a complete description, since it tells us nothing about the *length of intervals between the repetition of words* in the stream of speech. Let us turn to this topic.

**VI. THE LENGTH OF INTERVALS BETWEEN REPETITIONS**

Perhaps the easiest approach to an understanding of the problem of the length of intervals between the repetitions of words is to find out what is actually the case in an extensive sample of speech like Joyce's *Ulysses*, whose adoption for this particular purpose seems to be recommended by the existence of Hanley's excellent *Index* thereto in which page references are given for the occurrences of all words that are used 24 or fewer times.

In 1937 my then student, Dr. Alexander Murray Fowler, as previously reported elsewhere, undertook as a seminar topic the preliminary exploration of the number of pages that intervened between the repetition of all the different words that occurred 5, 10, 15, 20 and 24 times in Joyce's *Ulysses* (as determined from Hanley's *Index*). He found an interesting inverse relationship between the length of intervals on the one hand and the number of intervals having that length on the other.

Before discussing this inverse relationship, let us first discuss Fowler's methodological procedure which, though inescapably onerous, was analytically simple and essentially as follows. Each word that occurred, say, 5 times was considered to have 4 intervals, $I$, between its occurrences. And the length of each of these 4 intervals in terms of intervening pages was established by subtracting the respective page references from one another as given in the *Index*. More explicitly, the 1st interval was determined by subtracting the page on which the word first occurred from the page on which it occurred the second time; similarly the $n - 1$ interval was obtained by subtracting the page of the $n$-th occurrence from that of the $n$-th occurrence from that of the $n$th for any word in question.

Naturally, if the word is repeated on the same page, the subtraction yields zero as the size of the interval. To avoid operating with zero in the calculations below, I subsequently added 1 page to all intervals so that, for example, if 20 pages resulted from the actual subtraction of two successive page-references, the resulting interval was nevertheless considered to be 21.*

*It would have been a somewhat better statistical practice, to have added 1/2 page instead of 1. However, in the onerous task of calculating slopes and errors, as presented below, by keeping y integral instead of fractional, I was able to use my tables for integral values and thereby save an enormous amount of work. If the reader substituted 1/2 page for 1 page for the below discussed calculation of slopes, he would increase all slopes by
After having determined the sizes of the intervals between each of the 5 occurrences of all the 906 words that occur 5 times in *Ulysses* (3624 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. In all cases he found not only that short intervals were much more abundant than longer ones, but also that the *N* number of intervals of a like *I* size was inversely related to the size of *I*, or, in the form of a general equation:

\[ N^p \times I_f = \text{a constant (approximate)} \]

In this equation *f* refers to the frequency of occurrence of the different words of like frequency whose intervals are being measured (in the present case, where we are treating all words that occur 5 times in the *Ulysses*, or *f* = 5, we may speak of *I*5). Moreover, the exponent, *p*, represents graphically the absolute slope of the line fitted to the points when the data are plotted on doubly logarithmic paper with *N* on the abscissa and *I*5 on the ordinate as in Fig. 2-4.*

The reason for using the more general term, *I*5, in the above equation instead of *I*5, is that the same inverse relationship for the number and sizes of intervals was found in each of the classes of words that occurred 10, 15, 20, and 24 times respectively in the *Ulysses*. After submitting Fowler’s observations to the usual routine checking and finding them accurate to a high degree, I extended the analysis to include words occurring 6, 12, 16, 17, 18, 19, 21, 22, 23 and 24 times in the *Ulysses*, which latter is the upper frequency limit for which page references are given in Hanley’s *Index*. In all cases I noted the same inverse relationship although, in studying the data mathematically, I found significant differences in the size of the constant in the equation, as we shall presently see when we discuss the intercepts.

To illustrate graphically the nature of the above mentioned data, there are presented in Fig. 2-4, on doubly logarithmic paper with *N* on the abscissa and *I*5 on the ordinate, the actual numbers and sizes of all intervals between repetitions from *I*5 = 1 page through *I*5 = 50 pages for all the 906 words occurring 5 times in the *Ulysses*. The negative slope of the line of best *Y*’s (where *Y* = log *I*5) for the data of Fig. 2-4 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:

\[ N^{1.25} \times I_5 = \text{a constant} \]

if we remember that *I*5 has only integral values from 1 through 50.

*The measurement of *N* on the abscissa instead of on the ordinate, as is traditionally usual, was deliberately decided upon in order to bring the data of Fig. 2-4 into conformity with those of Fig. 2-5 which are plotted in the traditional manner. The relationship is not altered if the coordinates are reversed.*

---

*Thus specifically the negative slope of 1.15 reported below in Table 2-3 for words occurring 15 times (the arbitrarily selected 5th analysis) increases to 1.34 when the data are recalculated on the basis of an added 1/2 page instead of 1. The same would be found with the other sets of data. This slight difference in slope does not alter our argument.*
As to the remaining frequency classes (viz., the words that occur 6, 10, 12, 15, 16, 17, 18, 19, 20, 21, 22, 23, and 24 times respectively), we obviously cannot afford the space of presenting a graph for each of them. However, since we are interested only in the slopes and errors of our material, we can and shall give the complete information in tabular form for all the above frequency classes in Table 2-3. Referring to this table with its

![Graph](image)

**Fig. 2-4. The interval-frequency relationship.** The number of different intervals of like size (in pages) between the repetitions of words occurring five times in Joyce’s *Ulysses*.

15 different analyses as listed in Column I for the respective frequency classes of Column II whose number of different words are in Column III, we find the desired information about slopes, errors, and the like in Columns IV, V, and VI. In Column IV are the negative slopes of the line of best Y’s as calculated by least squares; in Column V are the root-mean-square deviations of the lines of the slope; and in Column VI are the Y-intercepts of the best lines of Y’s (actually the antilog of the Y-intercept). Analysis No. 14F is Dr. Fowler’s analysis of *I_2A*, while No. 15Z is mine; both are included to suggest the close correspondence between his and my analyses.

These calculations are based upon interval-sizes from 1 through 21 inclu-
sive for all the 15 different analyses instead of for the intervals from 1 through 50, as was the case with the data of Fig. 2-4. The reason for restricting the calculations of the present material to the 21 smallest interval sizes is that some of the larger interval sizes in several of the sets of data had no occurrences (i.e., $N = 0$) - a fact that made calculation and comparison more difficult. And so in order to be able to compare all 15 sets of data the upper limit of $I_f$ was reduced to an interval of 21 pages, for which all 15 sets of data had instances of intervals for all the units from $I_f = 1$ through $I_f = 21$. These 21 lowest units will suffice to reveal the presence of any tendency towards systematization in the entire tabular material.

### Table 2-3

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

Turning now to the entries of Column IV of Table 2-3 we find that the negative slopes range from .96 to 1.29 with the median at 1.20. The sizes of the errors in Column V show no marked correlation in their variations with the lower, middle, or upper frequency classes, but seem to be distributed without favoritism throughout the 15 analyses.

As to the slopes of Column IV with their median at 1.20, we may say without fear of contradiction that the high degree of correspondence between these values precludes the completely haphazard or random. In fact they point unmistakably to the existence of a fundamental correlation between $N$ and $I_f$. And since this correlation will be found of importance for
the balance of our study, we shall now briefly marshal conceivable objections to our acceptance of this correlation as of fundamental importance.

First of all, it cannot be objected that this correlation results merely from the fact that we have selected only the 21 smallest interval sizes, since in Fig. 2-4 we found for $I_5$ a negative slope of 1.25 with error of ± 0.168 for the 50 smallest intervals, which is not appreciably different from the slope of 1.21 with error of ± 0.151 for the 21 smallest interval sizes of $I_5$, as given in Analysis No. 1, Table 2-3. Hence by more than doubling the range of interval sizes we do not appreciably alter the slope or error. Moreover, when in Fig. 2-5 we present the complete data for $I_{24}$ we shall note by inspection that the correlation is valid for all of the interval sizes, including those that are far larger than 50 pages.

Second, the correlation between $N$ and $I_f$ cannot be ascribed to the fact that we decided to select a page as a unit of measurement, since the relationship would still hold if we took 2 pages, or 1/10 of a page, or if we multiplied $I_f$ by some other constant, or used words instead of pages as a unit (a course that would be advisable in the case of the repetition of words of highest frequency). Nor does the correlation result from the fact we have arbitrarily added one page to each interval-size in order to avoid operating with zero; such an addition might slightly modify the numerical value of each slope but not the essential correspondence between all slopes.

Third, the correlation cannot be dismissed on the ground that we have considered no frequency-class larger than $I_{24}$ which, as we remember, is the highest frequency class for which the necessary page references are given in Hanley’s Index. On the contrary, it is precisely the presence of this correlation among words of very low frequencies that is one of its most startling features. For we might conceivably expect to find some sort of principle governing the spacing of repetitions of the highly frequent words that occur 1, 2, or 3 times in every 100 words but not for those occurring 1, 2, or 3 times in every 100 pages.

And finally and most important of all, we may not overlook the obvious fact that the correlation is by no means an a priori necessity. For as a matter of fact an artificial sample of speech could be fabricated in which (out of many different possibilities) every interval was 1. Indeed the lack of any a priori reason for the necessary existence of the correlation between $N$ and $I_f$ alone suffices to justify our acceptance of it.

And so we shall continue on the basis of the belief that the relationship represented by the equation,

\[ N^p \times I_f = \text{a constant} \]

is fundamental, and not a statistical artifact. *

In stating that $N \times I_f = \text{a constant}$ for the intervals between the repeti-

*As far as the constant in the above equation is concerned, its respective values for the 15 analyses of Table 2-3 are indicated by the antilogs of the Y-intercepts of Column VI, where we note a slight tendency towards an inverse relationship between the size of the intercepts and the sizes of the frequency classes of Column II. But clarification of this point should await the results of further quantitative investigations, particularly of the very highest frequency classes.
tions of all words of like frequency of occurrence, we must beware imputing to the equation more than it actually describes. For although it describes a hyperbolic preference for short intervals, it is noncommittal about whether (1) on the one hand, the shortest intervals will occur either early or late in the total repetition of words, or whether (2) on the other hand, the intervals of different sizes will tend to be distributed evenly over the entire sample of speech. Indeed, as we shall now see from the data of Table 2-4 and of Fig. 2-5, it is curiously enough this second eventuality (2) that is unmistakably the case.

To illustrate the above point (2) we present in Table 2-4 the dispersion of all the 1-page intervals (i.e., \( f = 1 \)) among the \( f - 1 \) repetitions of all the words in each of 10 arbitrarily selected frequency classes which we have

### TABLE 2-4

The dispersion of single-page intervals between the \( f - 1 \) repetitions of all words that occur with ten arbitrarily selected frequencies of occurrence, \( f \), in Joyce's *Ulysses* (Hanley's Index).

#### A

**The First 12 Intervals between Repetitions**

<table>
<thead>
<tr>
<th>No. of Sample</th>
<th>( f )</th>
<th>( f - 1 )</th>
<th>Intervals between Repetitions in Order of Appearance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>1 2 3 4 5 6 7 8 9 10 11 12</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>5</td>
<td>62 55 62 58 52</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>11</td>
<td>7 19 15 16 9 12 18 16 12 15 14</td>
</tr>
<tr>
<td>3</td>
<td>16</td>
<td>15</td>
<td>10 10 13 18 11 16 11 11 9 11 9 9</td>
</tr>
<tr>
<td>4</td>
<td>17</td>
<td>16</td>
<td>4 3 5 6 4 8 5 10 11 9 14 5</td>
</tr>
<tr>
<td>5</td>
<td>18</td>
<td>17</td>
<td>9 11 6 5 6 7 7 6 9 6 2 6</td>
</tr>
<tr>
<td>6</td>
<td>19</td>
<td>18</td>
<td>3 8 5 11 5 6 13 9 6 5 6 8</td>
</tr>
<tr>
<td>7</td>
<td>21</td>
<td>20</td>
<td>3 4 10 5 8 9 3 10 8 11 7 7</td>
</tr>
<tr>
<td>8</td>
<td>22</td>
<td>21</td>
<td>7 5 8 12 5 9 5 9 6 7 5 8</td>
</tr>
<tr>
<td>9</td>
<td>23</td>
<td>22</td>
<td>3 5 6 4 8 4 3 2 7 3 4 4</td>
</tr>
<tr>
<td>10</td>
<td>24</td>
<td>23</td>
<td>3 5 2 1 5 3 3 3 4 5 2 3</td>
</tr>
</tbody>
</table>

#### B

**The Intervals from 13 through 23**

<table>
<thead>
<tr>
<th>No. of Sample</th>
<th>( f )</th>
<th>( f - 1 )</th>
<th>Intervals between Repetitions in Order of Appearance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>13 14 15 16 17 18 19 20 21 22 23</td>
</tr>
<tr>
<td>3</td>
<td>16</td>
<td>15</td>
<td>6 8 12 8 7 8</td>
</tr>
<tr>
<td>4</td>
<td>17</td>
<td>16</td>
<td>8 6 7 8 4</td>
</tr>
<tr>
<td>5</td>
<td>18</td>
<td>17</td>
<td>5 6 4 5 7 4</td>
</tr>
<tr>
<td>6</td>
<td>19</td>
<td>18</td>
<td>2 7 10 5 7 4</td>
</tr>
<tr>
<td>7</td>
<td>21</td>
<td>20</td>
<td>6 6 2 1 7 8 4 2</td>
</tr>
<tr>
<td>8</td>
<td>22</td>
<td>21</td>
<td>6 6 7 10 7 10 9 5 2</td>
</tr>
<tr>
<td>9</td>
<td>23</td>
<td>22</td>
<td>5 7 3 6 2 7 2 3 1 3</td>
</tr>
<tr>
<td>10</td>
<td>24</td>
<td>23</td>
<td>7 3 2 0 1 2 2 2 8 3</td>
</tr>
</tbody>
</table>
analyzed. From this table it is evident that the shortest intervals show no preponderant favor either for the beginning, or for the end, or for any other part of the total sample. Indeed, they seem to be scattered more or less evenly among their respective succession of intervals. And the same is also the case for the interval sizes from $I_f = 2$ through $I_f = 21$, as well as for all other interval sizes for all our analyses (although limitations of space prevent our including the data here).

Moreover in Fig. 2-5, which is completely typical of all the 15 different samples of Table 2-3, we find a complete scatter diagram of the dispersion of all the interval sizes among all the words that occur 24 times in the

![Fig. 2-5. The distribution of intervals between repetitions among the words occurring twenty-four times in James Joyce's *Ulysses*.](image)

*Ulysses*. The 23 successive intervals between repetitions are indicated arithmetically on the abscissa while the actual sizes, $I_{24}$ of the intervals are measured logarithmically on the ordinate. The small crosses locate all intervals in respect of their size and of their position of occurrence. The data are from analysis No. 15Z.

As far as we can see, the crosses on Fig. 2-5 are distributed quite evenly over the scatter diagram, with no further systematization whatsoever. The same is also the case in the scatter diagram of the remaining 14 samples, although we cannot afford the space to present them here. Hence the equation, $N^r \times I_f = a constant$ (approximate), applies not only to all the combined $f - 1$ intervals of each given frequency class, $f$, but also individually to each of the $f - 1$ intervals of each frequency class. In other words on the basis of our analyses of the distribution of interval sizes we may say, in general: the inverse relationship between $N$ and $I_f$ is found to apply not
ON THE ECONOMY OF WORDS

only to the intervals of an entire frequency class, \( f \), but also to each of the successive \( f-1 \) intervals of that class. \(^{14}\)

Now that we have established the presence of this inverse relationship between the number and sizes of intervals between the repetition of words, the task remains of probing into the dynamic reasons for its existence.

**VII. THE PROBLEM OF SPREADING WORK OVER TIME (THE EVEN DISTRIBUTION OF WORK OVER TIME)**

For some pages now we have been so actively engaged in our quantitative study of the phenomena of words that we have allowed our interest in the Principle of Least Effort to slip into the background of our discussions. Now, however, as we turn our attention to inquiring into the dynamic reasons for the quantitative distributions, \( N^p \times I_f = \text{a constant} \) (approximate), we should be well advised to begin by asking what this distribution could mean in terms of Least Effort; and if we remember from the data of Table 2-4 and of Fig. 2-5 that the various interval sizes, \( I_f \), are distributed evenly over the entire time of the sample, we may suspect that the Effort involved in the spacing of words and their repetitions in the stream of speech may also be evenly distributed over time. This suspicion brings us to a consideration of the hitherto largely neglected concept of time during which all work is expended.

By the term the even distribution of work over time, we can mean only one thing: viz., that the total work, \( W \), to be expended during the interval of time, \( T \), will be expended at such a rate that during any unit interval of time, \( t \), the same amount of work, \( w \), will be expended, or \( \frac{w}{t} = K \).

In terms of a sample of speech which consists of \( F \cdot Sn \) running words, the even distribution of work over time will mean that the total work expended during the sample will be so distributed that the same amount will be spent during each of the two halves; or the same between each of the four quartiles, or, if one will, the same between each of the \( F \) number of \( Sn \) running words. That is, the rate at which work is expended must be as near constant as possible, regardless of the time unit selected.

Now that we have defined what we mean by the concept of the even distribution of work over time, the very immediate task remains of explaining theoretically the equation for words, \( N^p \times I_f = \text{a constant} \) (approximate), in terms of the even distribution of work over time. To this end our first step (A) will be to construct a mechanical analogue (the Bell Analogy) in which minimized work will be distributed evenly over time and in which we shall find frequency distributions that can be described by the same equations as we have observed to be descriptive of the distribution of words in the stream of speech; this analogue will introduce the concept of time-perspective. Then (B) we shall tie the contents of this chapter together in a simple summary which will serve to orient us for the remainder of the book, in which the Principle of Least Effort will appear ever more as a fundamental governing principle in all biosocial phenomena.
A. The Bell Analogy and the Meaning of "Time Perspective"

In order to begin with our theoretical explanation of our various equations, let us take \( n \) bells that are equivalent in size and equally difficult to ring,\(^*\) and let us attach them to a long straight board in such a manner that the bells are equally spaced along the board which, incidentally, will offer a constant resistance to all movement. 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 the bell ringer. The demon must ring one bell 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 ten times, or ten bells once each, he will in either case make ten round trips down the board and back in the space of ten seconds, and will have 10 marks therefor on the blackboard; and since we shall ask the demon to make his round trips over shortest distances, all his work will be thereby minimized.

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 from the blackboard increases integrally according to the series \( 1d, 2d, 3d, \ldots, nd \), it follows that the demon's work in getting to and from the respective bells will increase according to the series \( 1w, 2w, 3w, \ldots, nw \).

Now if we ask our demon to ring each bell with a frequency, \( f \), that is inversely proportionate to the work, \( w \), of running to and from it -or in equation form, \( w \times f = C \)- he will ring the closer (and hence easier) bells proportionately more often than the more 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:

\[
    r \times f = C
\]

However, if we now ask the demon to ring all bells according to this equation, but to stop after he has rung the nth and farthest bell once, and after he has rung all other bells with their allotted frequencies, then the \( n \) bells will have been rung approximately according to the equation:

\[
    F \cdot Sn = \frac{F}{1} + \frac{F}{2} + \frac{F}{3} + \ldots + \frac{F}{n}
\]

in which \( F \cdot Sn \) represents the total number of round trips made (as well as the total number of seconds of time), and where \( F \) represents the total frequency of the nearest bell, with \( F = n \). The above equation is only approximate since, for reasons presented some pages back, there can be no fractional frequencies.

The above equation, \( F \cdot Sn \), puts no restriction upon the order in which the demon rings the bells. Thus he may ring the nearest bell its

\(^*\)More specifically, we shall assume that it takes work to get to and from a bell, but that it takes no work to ring the bell itself.
allotted $F$ times before ringing the second nearest bell its allotted $F/2$ times, and so on progressively down the board until he has rung the $n$th and farthest bell a single time. In short he may always ring "the easiest remaining bell first" while postponing as long as possible the more distant and hence more difficult bells. The chief drawback to 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 $n$th bell rung.

To correct this uneven distribution of work over time, we shall ask the demon to distribute his work as evenly as possible over time while still ringing his bells according to the equation $F \cdot Sn$. 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 intervals, $l_i$ between the respective repetitions of the bells will approximate the general equation:

$$N \times l_i = a \text{ constant}$$

The reason for the approximation to this equation is that from second to second the demon will be counterbalancing the cumulative work, $w$, with the cumulative frequencies, $f_i$; that is, he will try to expend half his total work in each half of the $F \cdot Sn$ seconds, $\frac{1}{4}$ th in each quartile, and $1/F$ th in each $Sn$ seconds, and so on. Or, differently expressed, 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; so too, the longer the compensating cluster of easier bells is (that is, the greater the number of pealings of easier bells, when multiplied by their work), the rarer will be that cluster's occurrence. 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 will be rung, it follows that within clusters the bells will be rung with an above average high rate of repetition (that is, they will be rung at intervals that are much shorter than average). Indeed, there will be not only many above average short intervals between repetitions within clusters, but also proportionately so. Hence within clusters we may expect to find an approximation to the equation $N \times l_i = a \text{ constant}$ (approximate).

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 larger ones. Therefore in measuring the number, $N$, of interval-sizes, $I_f$, between the ringings of the same bell (or of any frequency class of bells) we shall find an approximation to the equation

$$N \cdot I_f = a \text{ constant} \quad \text{(approximate)}$$

and this will be true as aforesaid because our demon will be constantly counterbalancing the difficult but more rarely pealing bells at the farther 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 Fig. 2-5. 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*. That is, the succession of names rung out by the bells would constitute an artificial sample of running speech in which we should find approximately all the equations hitherto discussed in connection with Joyce’s *Ulysses*.

Now from the equation, $N \times I_f = a \text{ constant} \quad \text{(approximate)}$, the other equations can be deduced, although the reverse is not the case. Hence in this sense the equation, $N \times I_f = a \text{ constant} \quad \text{(approximate)}$, may be considered to be primary to the others - a consideration to which we shall presently turn.

But before we turn to the primarity of the above Number-Interval relationship of our Bell Analogy, we should stress the point that other explanations of the workings of the Bell Analogy are quite conceivable. Thus the demon could ring the $n$th bell once and the 1st bell $F$ times, and after that balance the rare but more difficult pealings of the bells at the farther end of the board with the frequent and easier pealing bells at the nearer end. Yet no matter how the analogy is explained, the demon would still be balancing the frequency of easy acts with the rarity of difficult acts so that during every $S_n$ seconds he will expend as nearly as possible $1/F$ th of his total work. For every $S_n$ seconds during which an above-average amount of work is expended (i.e., one in which more than $F/W$ has been expended - which can happen only by pealing an above-average number of harder bells) there must follow $S_n$ seconds during which a corresponding below-average amount of work is expended (i.e., one in which less than $F/W$ is expended - which can happen only by pealing an above-average number of easier bells), and *vice versa*. Hence there will emerge proportionately more shorter intervals between repetitions than longer ones.

Our demon, in facing the problem of our bell analogy, is working upon a true *group problem* in which every act influences every other act in bringing forth a population of successive acts in which minimized work is distributed at as nearly a constant rate as possible. In order to solve this *group problem* successfully, the demon must perceive it as a *group problem* which is cast in terms of *time*; or, as we shall say, the demon must have by definition a 100% *time perspective* which means, in terms of our bell analogy, not only the effective performance of acts with a frequency that is inversely
proportionate to the work involved, but also the even distribution of that work over time.

This concept of time perspective, which refers to the demon's ability both to grasp, and to react to, a group problem successfully is of interest not only because it permits of future objective definitions of various kinds of psychotically unbalanced time perspective but also because it shows that the relationship, \( N \times I_f = \text{a constant} \) (approximate), is not at all necessary to the stream of speech, since other distributions are quite possible.

Perhaps the easiest way to comprehend the meaning of a 100% time perspective is to reconnoiter a few possible theoretical cases of faulty time perspective in which by definition all work, though minimized, will be unevenly distributed, and where we shall not find an inversely proportionate relationship between \( N \) and \( I_f \). One such case of faulty time perspective is that of the "easiest first," which we have already mentioned and which, as we remember, represents nothing more than ringing each easiest remaining bell its allotted \( F/r \) times in uninterrupted succession before proceeding to the next easiest bell. In this case all intervals between repetitions would be 1, with the result that the slope of the distribution on doubly logarithmic graph paper would be zero. Therefore we might describe the case of "easiest first" as one of zero time perspective.

Whether the time perspective, thus described, is .00 or 1.00 the total amount of expended work, \( W \), remains the same. The difference is the rate at which the work is expended. With a 1.00 time perspective, the rate of work expenditure is as nearly constant as possible, but with a .00 time perspective, the demon works at an ever-increasing rate - running ever faster and faster- until he reaches a maximum velocity at the \( F \cdot Sn \) th second, only to drop precipitously again to minimum velocity at the \( F \cdot Sn + 1 \) st second. Inherent in the case of the "easiest first" is an automatic emergence of cycles (as the result of our previously discussed closure) in the rate of work which may be of interest for neurological theory.

Thus, for neurological purposes, any time perspective that is less than 1.00 might be construed as being indicative of what might be called a cyclothymic unbalance, which would be characterized by an abnormal succession of "ups and downs" of activity.

Turning now to those hypothetical cases of time perspective that may be larger than 1.00, we recall that the negative slopes of the samples of Joyce's Ulysses of Table 2-4 approximate the median value of 1.20. If for the sake of our more general exposition we ignore the errors of Column VI of that table, as well as the fact that we arbitrarily added 1 page to each interval, we may say that Joyce systematically tended to avoid repeating words. That is, in the terms of our Bell Analogy, in the Ulysses the acts of the past (whether of bells or of words) tend to be systematically treated as if they were more remote from the present than would actually be the case with a 1.00 time perspective (i.e., the events of yesterday were treated, say, as if they occurred day before yesterday). In short, in the Ulysses the present moment seems to engross and preoccupy Joyce systematically at the expense of past moments— a statement with which students of Joyce's "stream of consciousness" school of writing would have little to quarrel. In any
event, we may feel that something is "wrong" with Joyce's time perspective if we define 1.00 as being "right."

Other hypothetical cases of faulty time perspective are thinkable, such as those in which the curve becomes markedly concave upwards or downwards as if the single group problem were either solved as several split problems or as an appendage to other and larger problems. But upon these hypothetical cases we need not dwell, since for our present purposes we are interested solely in illustrating the possible meanings of time perspective in terms of our equation, \( N^a \times I_f = a \text{ constant} \) (approximate) which we shall hereafter designate the number-interval equation.

The above illustrative cases of faulty time perspective show that different samples of speech, on the one hand, may reveal marked variations in the values of the number-interval equation which refers to the rate of distributing work over time; on the other hand, the same samples may closely approximate the harmonic equation, \( F \cdot Sn \), which after all refers to the fact that work is being minimized regardless of the rate at which the work is distributed.

However, lest the reader doubt the general existence of a fundamental inverse relationship between the length and number of intervals between the repetition of words, we present in Table 2-5 the number-interval relationship of the words occurring 5, 10, 15, 20, and 25 times in the Old English epic Beowulf, and in Homer's Iliad. The Beowulf study was made by my former student, Mr. Allen Sorensen, who used the glossary of the Klaeber edition of Beowulf for that purpose; the Iliad study was done by my former student Dr. Harold D. Rose, who used Prendergast's Concordance and whose study of the Iliad we have previously mentioned. The methods of analysis were essentially those adopted by Dr. Fowler for the Ulysses. The data in both cases are in terms of intervening lines with the unit interval of 100 lines. We see from Table 2-5 that by and large there is an inverse relationship between the number and size of intervals in these two ancient texts which in time, style, and content are far removed from Joyce's Ulysses. The graphs of these several sets of points (which space limitations preclude presenting) do not yield a very satisfying rectilinearity throughout, although the scatter diagrams show no preference for any order of the interval. As to the Iliad, we remember that in Fig. 2-3 we also saw a remarkable rectilinearity in the number-frequency relationship of words. As to the Beowulf, we shall study its rank-frequency distribution in our next chapter.*

*In our next chapter, we shall also present linear rank-frequency data for the American Indian language, Nootka; we mention this now because in spite of the fact that the Nootka data represent only 10,000 running words, we nevertheless find a similar number-interval equation in terms of the number of intervening words. Thus the most frequent word which occurs 107 times in the sample has 31, 12, 15, 3 and 3 repetitions in the respective 5 classes of 1-10, 11-20, 21-30, 31-40, 41-50 intervening words. The 17 different words that occur 10 times have 35, 14, 9, 5 and 0 repetitions for the same 5 classes; and the 583 words that occur only twice have 131, 39, 34, 20 and 19 repetitions. The remainder of the data is similar. Therefore the Ulysses, Iliad, or Beowulf are similar in this respect to the American Indian language, Nootka.
### Table 2-5

The Distribution of the *Number* of the Repetitions of words occurring 5, 10, 15, 20, and 25 times respectively according to the *Length of Intervals* in terms of intervening lines in (A) the Old English of *Beowulf* (Sørensen count), and in (B) Homer's *Iliad* (Rose count).

<table>
<thead>
<tr>
<th>Length of Interval in Lines of Verse</th>
<th>A. Beowulf</th>
<th>B. Iliad</th>
</tr>
</thead>
<tbody>
<tr>
<td>1–100</td>
<td>91</td>
<td>284</td>
</tr>
<tr>
<td>101–200</td>
<td>51</td>
<td>115</td>
</tr>
<tr>
<td>201–300</td>
<td>40</td>
<td>67</td>
</tr>
<tr>
<td>301–400</td>
<td>36</td>
<td>72</td>
</tr>
<tr>
<td>401–500</td>
<td>32</td>
<td>54</td>
</tr>
<tr>
<td>501–600</td>
<td>31</td>
<td>55</td>
</tr>
<tr>
<td>601–700</td>
<td>11</td>
<td>53</td>
</tr>
<tr>
<td>701–800</td>
<td>11</td>
<td>47</td>
</tr>
<tr>
<td>801–900</td>
<td>13</td>
<td>57</td>
</tr>
<tr>
<td>901–1000</td>
<td>16</td>
<td>48</td>
</tr>
<tr>
<td>1001–1100</td>
<td>15</td>
<td>47</td>
</tr>
<tr>
<td>1101–1200</td>
<td>7</td>
<td>43</td>
</tr>
<tr>
<td>1201–1300</td>
<td>4</td>
<td>46</td>
</tr>
<tr>
<td>1301–1400</td>
<td>3</td>
<td>42</td>
</tr>
<tr>
<td>1401–1500</td>
<td>3</td>
<td>43</td>
</tr>
<tr>
<td>1501–1600</td>
<td>4</td>
<td>41</td>
</tr>
<tr>
<td>1601–1700</td>
<td>4</td>
<td>45</td>
</tr>
<tr>
<td>1701–1800</td>
<td>2</td>
<td>38</td>
</tr>
<tr>
<td>1801–1900</td>
<td>2</td>
<td>36</td>
</tr>
<tr>
<td>1901–2000</td>
<td>3</td>
<td>36</td>
</tr>
<tr>
<td>2001–3000</td>
<td>7</td>
<td>271</td>
</tr>
<tr>
<td>3001–4000</td>
<td>2</td>
<td>207</td>
</tr>
<tr>
<td>4001–5000</td>
<td>164</td>
<td>164</td>
</tr>
<tr>
<td>5001–6000</td>
<td>94</td>
<td>94</td>
</tr>
<tr>
<td>6001–7000</td>
<td>73</td>
<td>73</td>
</tr>
<tr>
<td>7001–8000</td>
<td>41</td>
<td>41</td>
</tr>
<tr>
<td>8001–9000</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>9001–10,000</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>over 10,000</td>
<td>29</td>
<td>29</td>
</tr>
</tbody>
</table>
None of these sets of data represents so extensive a sample as Joyce's *Ulysses*, and none is as clearcut in its rectilinearity. Just what the various deviations in slope may mean for the "personalities" of the individual writers or speakers in question is a problem that awaits future study. In the meantime, we shall continue in the belief that the extensive sets of *Ulysses* data, with their rectilinear correlations of negative slopes of nearly 1, were sufficient to justify an exploratory construction of our Bell Analogy.

Returning now to the Bell Analogy, let us remember that we have shown no further connection between the Bell Analogy on the one hand and the *number-interval equation* of word repetitions on the other, except for a similarity of mathematical description. Hence we have no right to conclude on the basis of our mechanical analogue alone that the *number-interval equation* of word repetitions does indeed represent the even distribution of minimized work over time.

On the other hand, the Bell Analogy is undeniably an objective picture of some of the salient features of the distribution of words. By using this analogy, and by refining upon it in our following chapter, we shall be in an ever better position to study the dynamics of speech without losing sight of the fact that words are a living phenomenon.

**B. Summary and Prospectus**

In the present chapter we have done two things.

In the *first* place we have presented quantitative data on the frequency of occurrence of words and of meanings in the stream of speech for the sake of showing the orderliness of the phenomenon. There was no *a priori* necessity for the particular orderliness observed, and the further corroborative data of subsequent chapters will ever more decrease the likelihood that the particular types of observed distributions are merely the result of chance.

In the *second* place, we have attempted to rationalize our empiric data on the basis of economy.* This attempt at rationalization is only a beginning, and is by no means offered in the belief that it is anything more than a preliminary orientation.

The chief step in the rationalization was the Bell Analogy in which we faced the question not only of minimizing work but also of the rate of distributing that minimized work over time. This question of distributing minimized work over time is fundamental to the thesis of this study, since it will lead in a later chapter to the concept of minimizing the average rate of work expenditure, and then to the concept of minimizing the *probable* (average) rate of work expenditure which, by definition, is *least effort*.

The Bell Analogy was deliberately constructed in such a way that the rate of work would be minimized over time with as little variation as possible from second to second. It was also constructed so that the distribution of the sounds of the bells would be similar to that of the distribution of

*Hyperbolic distributions of the type observed quite often suggest a governing consideration of economy, even as bell-shaped curves suggest a certain general kind of rationalization.
words in the stream of speech. Thanks to the objectivity of the Bell Analogy, we could see that the *number-interval equation* which referred to the inverse proportionality between the frequency and length of intervals between repetitions was primary to the *equation of the harmonic series* with which the present chapter started. That is, the latter equation can be deduced from the former, but not the reverse.

So much, then, in summary for what we have done in the present chapter.

Looking to the future, we can either continue the theoretical rationalization as we refine ever more upon our somewhat rigid Bell Analogy. Or we can present further empiric data. Or we can do both, as we alternate between the inductive and the deductive methods in the manner of the natural sciences.

In our following chapters we shall do both, as we present further sets of observation which we shall attempt to rationalize theoretically with an ever-increasing degree of refinement. By a consistent use of this inductive-deductive manner of analysis, we shall learn of the forms and functions of the large and small entities of the stream of speech, both from the view-point of the speaker and from that of the auditor, and we shall simultaneously learn more about their underlying principles of organization until we reach a point in Chapter Five where we can see that the entire phenomenon of speech is presumably subject to the Principle of Least Effort.
ON THE ECONOMY OF WORDS

References in Chapter Two (pp. 546-547 in printed book)


3. M. L. Hanley, *Word Index to James Joyce's Ulysses*, Madison, Wis., 1937 (statistical tabulation by M. Joos). Historically considered, Joyce's *Ulysses* was selected for analysis in the attempt to show that the harmonic distribution would not be found in samples of this size [cf. M. Joos, review of Zipf, *Psycho-Biology of Language*, in *Language*, Vol. 12 (1936), 196-210; cf. my reply, with quotations from M. H. Stone, "Statistical methods and dynamic philology," *Language*, Vol. 13 (1937), 60-70]. Dr. Joos's *Ulysses* data provided the most clear-cut example of a harmonic distribution in speech of which I know.


8. Originally published, G. K. Zipf, "The meaning-frequency relationship of words," *Journal of General Psychology*, Vol. 33 (1945), 251-256. My students who helped with the tabulation were, the Misses E. L. Goucher, L. Hill, R. Hubbard, E. Kleinschmidt, V. T. Spang, E.
9. In the present case, as will be discussed again in detail in our following chapter, we have adopted an operational definition of meaning (i.e., we have counted the different meanings in a dictionary). Cf. P. W. Bridgman, "Operational Analysis," *Philosophy of Science*, Vol. 5 (1938), 114-131; also, his *Logic of Modern Physics*, New York: Macmillan, 1927. In Chaps. 5 and 7 we shall approach the problem of meanings from a different angle.

10. Argument presented in G. K. Zipf, "Homogeneity and heterogeneity in language," *Psychological Record*, Vol. 2 (1938), 347-367. The argument was essentially as follows: Let \( N_f \) equal the number of different words of integral frequency, \( F \). Assume that this includes all words, the \( r x f = C \) distribution, that lie between \( F + 1/2 \) and \( F - 1/2 \). Thence

\[
N_f = \frac{C}{F^{1/2}} - \frac{C}{F^{1/2}}
\]

(i.e., the subtraction of the rank, \( R_{f+1/2} \) from the rank \( R_{f-1/2} \). Simplifying we have

\[
N_f = \frac{C}{F^2/4}
\]

For the more general case in which the rank-frequency distribution has other slopes than -1.00, cf. the ingenious analysis of Mr. Joos, *Language*, Vol. 12 (1936), 196-210, in which Joos shows that the exponent of \( f \) of a rank-frequency distribution is approximately larger by 1.00 than the exponent of \( F \) of the number-frequency distribution.


14. The argument of the bell analogy and its related tool analogy will be found, *ibid*.

