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.[1] 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 viewpoints 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 $\frac{1}{10}$ 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

Arbitrary Ranks with Frequencies in James Joyce's *Ulysses* (Hanley Index)

I Rank (r)	II Frequency (f)	III Product of I and II ($r \times f = C$)	IV Theoretical Length of Ulysses ($C \times 10$)
10	2,653	26,530	265,500
20	1,311	26,220	262,200
30	926	27,780	277,800
40	717	28,680	286,800
50	556	27,800	278,800
100	265	26,500	265,000
200	133	26,600	266,000
300	84	25,200	252,000
400	62	24,800	248,000
500	50	25,000	250,000
1,000	26	26,000	260,000
2,000	12	24,000	240,000
3,000	8	24,000	240,000
4,000	6	24,000	240,000
5,000	5	25,000	250,000
10,000	2	20,000	200,000
20,000	1	20,000	200,000
29,899	1	29,899	298,990

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.[4]

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

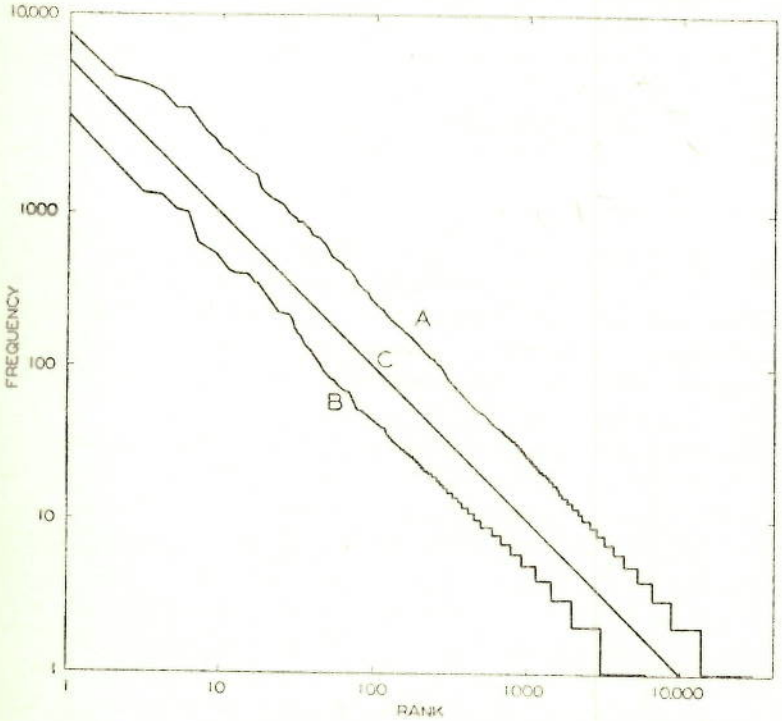

Fig. 2–1. The rank-frequency distribution of words. (A) The James Joyce data; (B) the Eldridge data; (C) ideal curve with slope of negative unity.

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.[5] 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—lines 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 m 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 m different meanings and that on the other hand there will be fewer than m 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 m_1 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