Zipf (1949) Human Behavior + the Principle of Least Effort 1-7

CHAPTER ONE

INTRODUCTION AND ORIENTATION

Everyone in the course of his daily living must to some extent move about in his environment. And in so moving he may be said to take paths. Yet these paths that he takes in his environment do not constitute his entire activity. For even when a person is comparatively at rest, there is still a continual movement of matter-energy into his system, through his system, and out of his system if only in the accomplishment of his metabolistic processes. This movement of matter-energy also proceeds over paths. Indeed, a person's entire body may be viewed as an aggregate of matter that is in transit at differing speeds over different paths within his system. His system in turn moves about as a unit whole over paths in his external environment.

We stress this concept of movement over paths because we shall attempt to demonstrate in the course of our following chapters that every individual's movement, of whatever sort, will always be over paths and will always tend to be governed by one single primary principle which, for the want of a better term, we shall call the *Principle of Least Effort*. Moreover, we shall attempt to demonstrate that the structure and organization of an individual's entire being will tend always to be such that his entire behavior will be governed by this Principle.

And yet what is this Principle? In simple terms, the Principle of Least Effort means, for example, that a person in solving his immediate problems will view these against the background of his probable future problems, *as estimated by himself*. Moreover he will strive to solve his problems in such a way as to minimize the *total work* that he must expend in solving *both* his immediate problems *and* his probable future problems. That in turn means that the person will strive to minimize the *probable average rate of his work-expenditure* (over time). And in so doing he will be minimizing his *effort*, by our definition of effort. Least effort, therefore, is a variant of least work.

In the interest of defining and of elucidating the Principle of Least Effort, and of orienting ourselves in the problem of its demonstration, we can profitably devote this opening chapter to a preliminary disclosure of the Principle, if only on the basis of commonplace cases of human behavior that are admittedly oversimplified for the sake of a more convenient initial exposition.

I. THE SELECTION OF A PATH

Sometimes it is not difficult to select a path to one's objective. Thus if there are two cities, *A* and *B*, that are connected by a straight level highway

with a surface of little friction, then this highway represents simultaneously the *shortest,* the *quickest,* and the *easiest* path between the two cities—or, as we might say, the highway is at once a path of *least distance* and of *least time* and of *least work.* A traveller from one city to the other would take the same path regardless of whether he was *minimizing* distance, time, or work.

On the other hand, if the two cities happen to be separated by an intervening mountain range, then the respective paths of least distance, and of least time, and of least work will by no means necessarily be the same. Thus if a person wanted to go by foot from one city to another by least distance, he would be obliged to tunnel through the base of the mountain chain at a very great expense of work. His quickest course might be over the tops of the mountains at a great cost of labor and at great risk. His easiest path, however, might be a tortuous winding back and forth through the mountain range over a very considerable distance and during a quite long interval of time.

These three paths are obviously not the same. The foot-traveller between the two cities cannot, therefore, simultaneously minimize distance, time, and work in a single path between the two cities as the problem now stands. Which path, therefore, will he take? Or, since the above case is fairly typical of life's daily problems, in which impediments of various sorts obstruct our way, which path do we actually take? Clearly our selection of a path will be determined by the particular *dynamic minimum* in operation.

II. THE "SINGLENESS OF THE SUPERLATIVE"

The preceding discussion of the selection of paths not only illustrates the meaning of a *minimum* in a problem in dynamics but also prepares the ground for a consideration of the concept of the *"singleness of the superlative"* [1] which, incidentally, will provide an intellectual tool of considerable value for our entire inquiry.

The concept of the "singleness of the superlative" is simple: no problem in dynamics can be properly formulated in terms of more than one superlative, whether the superlative in question is stated as a *minimum* or as a *maximum* (e.g., a *minimum* expenditure of work can also be stated as a *maximum* economy of work). If the problem has more than one superlative, the problem itself becomes completely meaningless and indeterminate.

We do not mean that a particular situation will never arise in which the minimizing of one factor will not *incidentally* entail the minimizing of another or other factors. Indeed, in our preceding section we noted a situation in which the easiest path between two cities might be a straight level highway that also represented the shortest and quickest path. Instead we mean that a general statement in dynamics cannot contain more than one superlative if it is to be sensible and determinate, since a situation may arise in which the plural superlatives are in conflict.

Perhaps the simplest way to emphasize the singleness of the superlative is to present as an example a statement with a single superlative that is meaningful and determinate. Then we shall note how meaningless and indeterminate the statement immediately becomes when a second superlative is added.

As a suitable example we might take the imaginary case of a prize offered to the submarine commander who sinks the *greatest number* of ships in a given interval of time; in this case, *maximum number* is the single superlative in the problem. Or we might alter the terms of the problem and offer a prize to the submarine commander who sinks a given number of ships in the *shortest possible* time; in this second case, *time* is the *minimum;* and, since it is the only superlative in the statement, the problem is quite meaningful and determinate. In either of the above examples the submarine commander can understand what the precise terms of the prize are.

Yet when we offer a prize to the submarine commander who sinks the *greatest number* of ships in the *shortest possible time,* we have a double superlative—a *maximum* number and a *minimum* time—which renders the problem completely meaningless and indeterminate, as becomes apparent upon reflection.

Double superlatives of this sort, which are by no means uncommon in present-day statements, can lead to a mental confusion with disastrous results.*

In the present study we are contending that the entire behavior of an individual is at all times motivated by the urge to minimize effort.

The sheer idea that there may be only one dynamic minimum in the entire behavior of all living individuals need not by itself dismay us. The physicists are certainly not dismayed at the thought that all physical process throughout the entire time-space continuum is governed by the one single superlative, *least action.*† Indeed, the presence of only one single superlative for all physical process throughout the entire time-space continuum can even be derived logically from the basic postulate of science that there is a unity of nature and a continuity of natural law (in the sense that the same laws of nature govern all events in time-space). For, according to this postulate, the entirety of time-space, with all its happenings, may be viewed as constituting a single problem in dynamics which in turn can have only one single superlative—a superlative which in the opinion of physicists is that of *least action.*[4]

By the same token, the sheer idea of there being one single superlative for all living process is not in and for itself an *a priori* incredibility.

On the other hand, there is also admittedly no *a priori* necessity for our believing that all living process does in fact behave at all times according

* As pointed out years ago, the frequent statement, "in a democracy we believe in the *greatest* good for the *greatest* number," contains a double superlative and therefore is meaningless and indeterminate. (In Part Two we shall see that the distribution of goods and services are in fact governed by a single superlative.) Intimately connected with the "singleness of the superlative" is what might be called the *singleness of the objective* whose implications are often overlooked (i.e., the pursuit of one objective may preclude or frustrate the pursuit of the second objective). These two concepts apply to all studies in ecology.[2]

† The principle of least action was first propounded by Maupertuis in the eighteenth century, and has been subsequently conceptually sharpened by others.[3]

to one single invariable superlative, such as that of least effort. That, after all, must first be established empirically, as was done with the principle of least action. We can even now note how bizarre the effect would be if a person behaved at one moment according to one dynamic minimum, and at the next moment according to an entirely different dynamic minimum. Nor would the effect be any less bizarre if one person's life were governed throughout by one superlative while his neighbor's life followed a totally different superlative.

In order to emphasize the ludicrousness of a variety of different superlatives, let us assume that each person consists of two parts, and that each part has a different dynamic superlative of its own. For example, let us assume that one part of the person is governed by least work while the other is governed by least time. In that case the person will represent two distinct problems in dynamics, with the result that he will be, effectively, two distinctly different individuals with two distinct sets of dynamical principles. One part of him, in its eagerness to save work, might conceivably even "get lost" from the other part of him, in its eagerness to save time.

Nor would the situation be different if we assume that a person now minimizes one factor and now another without any single governing principle behind the total phenomenon. For if the person's entire metabolistic and procreational system is organized, say, for the purpose of minimizing work in all its action, then there would have to be a simply staggering alteration of structure and of operation if the person in question were suddenly to minimize time. Since sudden alterations of such proportions are unknown, we are perhaps not overbold in suspecting *a fortiori* that an individual's entire activity from birth to death is governed throughout by the same one single superlative which, in our opinion, is least effort.

But that is not all. If we remember the extent to which offspring inherit the forms and functions of their parents, we may suspect that this inheritance is possible only if the offspring also inherit the parental dynamic drive that governs the parental forms and functions that are inherited.

Furthermore, if we view the present-day variety of living process as the result of slow evolutionary changes from an initial similarity of form and function, then we can understand *a fortiori* how the one initial single common dynamic superlative might well remain unchanged from generation to generation, regardless of how enormous the changes in forms and functions might become; and that, in turn, will mean that all individuals, regardless of their differences in form and function, will still be governed by the same single superlative.

But though we may argue at length as to the plausibility of one single superlative for all living process, yet, even if this were the case, we should still need to disclose what, in fact, the particular superlative in question is.

An actual disclosure of the single hypothetical superlative in question may be difficult for quite obvious reasons. If we take our previous example of the two cities with an intervening mountain chain, in which the paths of least distance, least time, and least work are three different paths, we are obliged in all candor to admit that sometimes one of these paths is taken

and sometimes another. For that matter, a tunnel may be dug through the base of the mountain to save distance, while airplanes are flown over the same mountain to save time, while pack horses continue to take the easier and more leisurely winding route. Or, to take another case, sometimes the reader will dart through traffic at considerable risk in order to save time in crossing a street; and sometimes he will take the longer and safer path to the corner, where he will wait for the traffic light. Even if we assume that we are all governed by the same one single dynamic superlative, which superlative is it?

But although the superlatives in the foregoing examples seem to be different, are they nevertheless irreconcilable? Before answering this question, let us remember the physicists' claim that according to their law of falling bodies, all free-standing bodies will fall (by least action) to the earth. Yet, despite this claim, we have all observed how leaves sometimes rise in the air, or how birds take off from the ground and fly out of sight, much as if they were exceptions to the law of falling bodies. Of course we know from a more careful inspection of the problem that these leaves and birds are by no means exceptions to the law of falling bodies; on the contrary, if all the factors in the problem are taken into consideration, they are behaving quite in accordance to the physical law in question.

May not the same be true of the three different paths to the other side of the mountain? Even though each of these paths may be taken simultaneously by someone, and even though a given person may now take one path and now another, there remains the possibility that the adoption of one or another by an individual under varying sets of circumstances is governed by the operation of some further single dynamic minimum that forever remains invariant. In any event, we shall argue that such is the case.

More specifically, we shall argue that if we view the above types of situations against the broad background of the individual's present and future problems, we shall find that an extraordinary expenditure of work at one moment, or an extraordinary haste in one situation, may simply be temporary devices for reducing the probable rate of the individual's work expenditure over subsequent periods of his life.

In short, we shall argue that the invariable minimum that governs all varying conduct of an individual is least effort.

III. THE PRINCIPLE OF LEAST EFFORT

Perhaps the easiest way to comprehend the meaning and implications of the Principle of Least Effort is to show the inadequacies of sheer *least work*, to which *least effort* is closely related. This is all the more worth doing because some persons (see below) apparently believe that least work is the basic minimum of living process, as often seems to be the case in particular situations that are considered out of context.

If we remember, however, that an individual's life continues over a longer or shorter length of time, then we can readily understand how the least work solution to one of his problems may lead to results that will

inevitably increase the amount of work that he must expend in solving his subsequent problems. In other words, the minimizing of work in solving today's problems may lead to results that will increase tomorrow's work beyond what would have been necessary if today's work had not been completely minimized. Conversely, by expending more work than necessary today, one may thereby save still more work tomorrow.

And, as we have argued about the functional relatedness of today and tomorrow, we may argue about the functional relatedness of the entire succession of events throughout the individual's whole life, in which the rate of his expenditure of work at one moment may affect the minimizing of his work at subsequent moment(s).

In view of the implications of the above quite obvious considerations, we feel justified in taking the stand that it is the person's *average rate of work-expenditure over time* that is minimized in his behavior, and not just his work-expenditure at any moment or in any one isolated problem, without any reference to his future problems.

Yet a sheer *average rate of work-expenditure over time* is not an entirely meaningful concept, since no mortal can know for certain what his future problems are going to be. The most that any individual can do is to estimate what his future problems are *likely to be,* and then to govern his conduct accordingly. In other words, before an individual can minimize his average rate of work-expenditure over time, he must first estimate the probable eventualities of his future, and then select a path of least average rate of work through these.

Yet in so doing the individual is no longer minimizing an average rate of work, but *a probable average rate of work;* or he is governed by the principle of the *least average rate of probable work.**

For convenience, we shall use the term *least effort* to describe the preceding least average rate of probable work. We shall argue that an individual's entire behavior is subject to the minimizing of effort. Or, differently stated, every individual's entire behavior is governed by the Principle of Least Effort.

Now that we have described what the Principle of Least Effort is, let us briefly illustrate it.

At the risk of being tedious, let the first example be our previous case of the two towns, *A* and *B,* that are separated by an intervening mountain range. Here we can see the enormous amount of work that could be saved in travel and trade if the two towns were connected by a tunnel of least distance through the base of the mountain; we can also see the enormous amount of work that it would take to construct such a tunnel. We are simply arguing that when the probable cost in work of digging the tunnel is estimated to be less than the probable work of not having the tunnel, then, if the necessary work for construction is available, the tunnel will be dug. The problem relates, therefore, to the probable amounts of work involved, as

* To avoid a possible verbal confusion, let us note that we are not discussing *least probable* average rate of work, but a *probably least* average rate of work.

estimated by one or more persons. Naturally, these persons can have been mistaken in their estimates, with the result that the tunnel can either succeed beyond their wildest hopes, or dismally fail. For we do not deny that "a person's hindsight is generally better than his foresight." We merely claim that a person acts on the basis of his "foresight"—with all that that will later be found to imply—according to the Principle of Least Effort.

The above type of argument will also apply to a path of least time over the mountain. Thus the enormous cost of flying munitions over the mountain to save time in supplying an army in combat on the other side may be more than justified by the future probable work that is thereby saved.

These cases of the different paths to the other side of the mountain represent instances of collective action and of collective economies, since, for example, a tunnel through a mountain is obviously not constructed by a single person but by the collective effort of a great many persons.

And yet we are not restricted to examples of collective effort in illustrating our Principle of Least Effort, which we contend also applies to an individual's own behavior. We might take the case of a student whose particular path of least effort out of his classroom would seem offhand to be the path that leads from his seat to the nearest aisle, and thence out of the door, through the hall, to the nearest stairway. On the other hand, in the event of a fire, the student might conceivably prefer to run with least time to the nearest window and adopt a path that is simultaneously a path of least work and of least time and of least distance to the ground. This path will also be a path of least effort, as estimated by himself, even at the risk of months in the hospital with a broken back. Other students may prefer to take paths through the smoke-filled corridors. These paths are also paths of least effort, as estimated by the students in question. Afterwards, when, as, and if all the students foregather, they can decide which of them, in the light of subsequent events, actually were the shrewdest gamblers in the sense of having both most correctly comprehended the nature and estimated the probabilities of the problem in their lives that was caused by the unexpected fire.

From this second example we note that the operation of the Principle of Least Effort is contingent upon the *mentation* of the individual, which in turn includes the operations of *"comprehending"* the "relevant" elements of a problem, of *"assessing their probabilities,"* and of "solving the problem in terms of least effort." We mention this vital consideration of *mentation* right here and now, so that we may prepare ourselves for the task of defining mentation, and of showing that the structure and operation of mentation are also governed throughout by the Principle of Least Effort, since an individual's mentation is clearly a part of his total behavior, and hence subject to our Principle of Least Effort.

The foregoing examples suffice to illustrate what the Principle of Least Effort is, and what its implications may be for everyday problems. By and large, our explanations of the above commonplace examples are pretty much in line with the way the reader himself would have explained them. We mention this consideration in order to suggest that our chief task may not be that of persuading the reader to adopt a totally new way of thinking, but rather of formally describing + of scientifically establishing the basic principle of our habitual way of thinking/

Zipf (1949) Human Behavior + The Principle of Least Effort (1949)

19-22

(plus 22-31 on Zipf's Law)
don't read yet

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

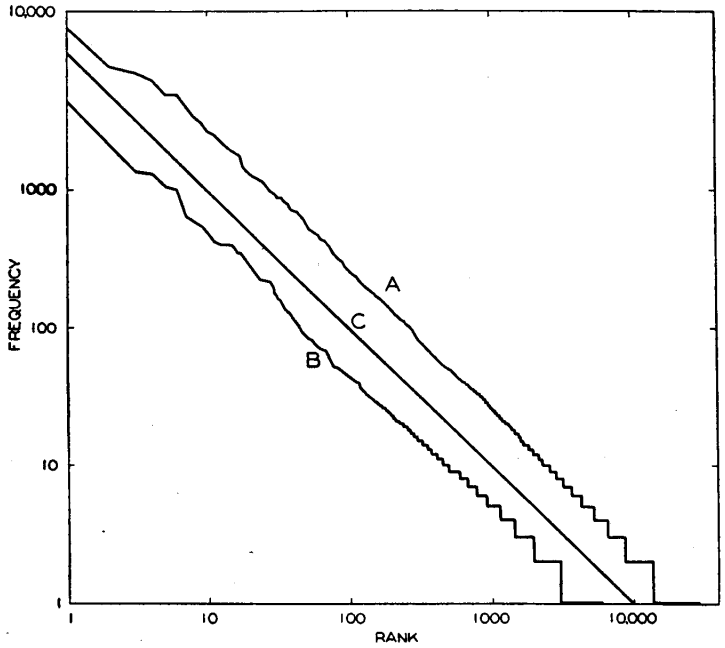

Fig. 2–1. The rank-frequency distribution of words. (A) The James Joyce data; (B) the Eldridge data; (C) ideal curve with slope of negative unity.

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

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 ½ as much on the Y-axis as on the X-axis (that is, it will have what is technically called a negative slope of ½,* 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 ½ 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.[6] 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.*[7] 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-middles 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.[8]

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.

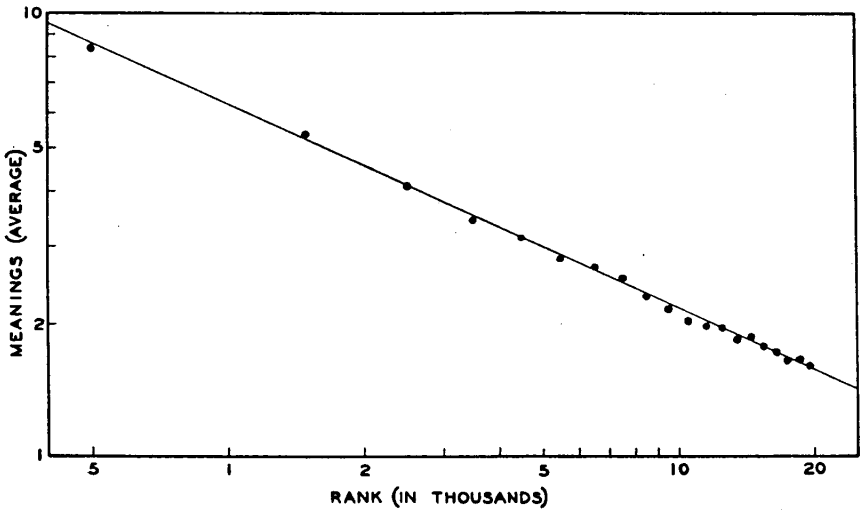

Fig. 2–2. The meaning-frequency distribution of words.

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 root of the average frequency of the words' occurrence (or will decrease according to the square root of the rank).

* 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.

Although later we shall again return to the entire question of the "meanings of words" with the problem of defining the term *meaning*,[9] 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.*

Zipf (1935) Psycho-Biology 20-39

II

THE FORM AND BEHAVIOR OF WORDS

THERE is a decided advantage in beginning our investigation of language by restricting our attention to the form and behavior of words, since the essential phenomena in the dynamics of words are far more clearly apparent and readily apprehended than those of the smaller or larger speech-elements. If it cannot be truthfully averred, as is sometimes felt, that the stream of speech is primarily a stream of words rather than a stream of, say, phonemes or sentences, the word does seem, nevertheless, to occupy a middle terrain between the smaller elements which are its components and the larger phrasal, clausal, and sentence elements of which the word is in turn but a part. In studying the dynamics of words, then, we are studying what represents simultaneously either an aggregate of, or the component of, other speech-elements, and are hence incidentally approaching the dynamic problems of these other speech-elements at their most accessible side.

PART I

THE QUESTION OF FORM

1. THE LENGTH OF WORDS AND THEIR FREQUENCY OF USAGE

Probably the most striking feature of words is difference in length. A word may consist of a single phoneme (e.g. English *a* in *a book*), or it may represent a phoneme-sequence of considerable magnitude (e.g. English *transcendentalism*, *constitutionality*, *quintessentially*). The question naturally arises as to what, if any, is the significance of these observable differences in length.

If there is any connection between the length of a word and its meaning, the nature of the connection is certainly not at once apparent. The same idea is found adequately conveyed in different languages by words of decidedly different length (e.g. English *trade*, German *Handwerk*; English *work*, German *Arbeit*). Hence, at least for the present, we may disregard considerations of meaning in examining the significance of the factor of length in the form of words.

The question is natural as to the number of different long and short words in the vocabulary of a given person, or of a given dialect. As far as any *a priori* statement is permissible on this subject, it seems deducible only from the law of permutations which clearly gives the presumption in favor of a greater abundance of different longer words than of shorter. For, the possible permutations of a given number of phonemes taken, say, five at a time is far greater than those when only two are taken at a time. If some permutations of phonemes would be too difficult to pronounce with convenience (e.g. *tp* or *tpbgd* in English), unpronounceability is not restricted to long or short words. That is, short permutations may be unpronounceable as well as long permutations. If long permutations offer greater opportunity for unpronounceable arrangements of phonemes than do smaller configurations, they also offer, by the same token, greater opportunity for pronounceable arrangements. Hence, the palpable fact that some combinations of phonemes are impossible to articulate does not in itself invalidate the *a priori* statement which has just been made.

Nevertheless, empirical evidence does preclude the use of this *a priori* statement, for it would be accurately descriptive only if every language availed itself progressively of every possible legitimate short permutation before employing a larger permutation, a condition which is by no means the case. In English, for example, the long permutation *constitutionality* is a meaningful word while the shorter

and equally pronounceable permutations *puv*, *za*, *ut* have no meaning. Hence, the one *a priori* statement which it seems possible to make about the length of words may be discarded at the outset.

Interesting light might be shed empirically upon the significance in the length of words if it were possible to make a list of all words in the active-passive vocabulary of an individual or of a speech-community to ascertain the actual number of different words representing each of the different degrees of length. But a list of this type for any single language is impossible. Dictionary lists are generally inadequate because of their inclusion of obsolescent and obsolete words, and because of their exclusion of highly useful neologisms. And it seems practically impossible ever to make a completely adequate list, even of the vocabulary of an individual person, because so many words exist in the passive vocabulary which are used rarely, if at all, in the stream of speech which is alone perceptible. Then, too, the active-passive vocabulary, whether of individuals or of speech-groups, differs so in size and content and varies so from time to time, that an attempt even to estimate merely the limits of a vocabulary at any time is largely a matter of guesswork.[1] Hence we are obliged at the very beginning of our investigation to restrict our attention exclusively to the objective evidence of the stream of speech itself which, during the entire course of our investigation, will be the sole source of our data.

Now, in so far as data are already available from the stream of speech, it seems reasonably clear that shorter words are distinctly more favored in usage than longer words. That is, however large the stock of short and long words may be, the evidence of language seems to indicate unequivocally that the larger a word is in length, the less likely it is to be used. To illustrate this point, the data gathered by F. W. Kaeding[2] from samplings of connected written German, totalling 10,910,777 words (or 20 million

syllables) in length, are presented. Kaeding selected the syllable as the unit of length, and, in the following tabulation of his results, the left-hand column indicates the magnitude of each class of words as estimated by the number of syllables; the center column presents the number of occurrences (including repetitions) of words of each magnitude; and the column at the right notes the percentage of the occurrences of all the words (including repetitions) of each magnitude to the total number of words (10,910,777).

Number of Syllables in Word	Number of Occurrences (Including Repetitions) of Words	Percentage of the Whole
1	5,426,326	49.76%
2	3,156,448	28.94
3	1,410,494	12.93
4	646,971	5.93
5	187,738	1.72
6	54,436	.50
7	16,993	.22
8	5,038	
9	1,225	
10	461	
11	59	
12	35	
13	8	
14	2	
15	1	
	10,906,235*	100.00%

These figures indicate that in German there is a decided preference in usage for short words, and that the magnitude of words stands in an inverse (not necessarily proportionate) relationship to the number of occurrences (including repetitions) of all words possessing that magnitude.

Though these statistics from Kaeding seem clear for all occurrences of all words when arranged in classes according to differences in syllabic magnitude it provides no information about the number of different words in each class. We do not know from the statistics for example whether the

* Subsequently corrected by Kaeding[1] to 10,910,777.

five million odd occurrences of words of one syllable represent the single occurrence of that many different words, or that many different occurrences of a single word. While Kaeding gives enough additional information in his entire treatise for us to determine with reasonable accuracy the general tendencies involved, we shall in lieu of this Kaeding material (which may be consulted in the notes [1]) present the results of three separate investigations, one of Plautine Latin,[2] one of modern colloquial Chinese [3] (Peiping dialect), and one of English,[4] which are far more precise than the Kaeding material on this subject, and which are corroborated in their main tendencies by the data of the Kaeding material.

The material for the investigation of colloquial Chinese of Peiping consists of twenty different samplings of connected speech, each a thousand syllables long. The number of different words * occurring in these 20,000 Chinese syllables are arranged in the ensuing table (p. 26) according to the relative frequency of their occurrence. The left-hand column of the Chinese statistics gives the times of occurrences; the center column presents the number of words assignable to a given frequency of occurrence; and the right-hand column, in parentheses, indicates the number of words in each frequency grouped according to the number of syllables (designated by a superior number). Thus, of the 2046 Chinese words of single occurrences, 315 contained only one syllable, 1571 two syllables, etc.

In the investigation of the Latin of Plautus a different procedure was adopted. With all the words of four Plautine plays (*Aulularia*, *Mostellaria*, *Pseudolus*, and *Trinummus*) selected for material, the average number of syllables in each frequency category was computed (p. 27). The

* It is commonly asserted that Chinese is a monosyllabic language in the sense that Chinese words on the whole represent compounds of monosyllabic roots (e.g. English *cupboard*) some of which do not occur except in compounds. That Chinese is not monosyllabic in the sense that every Chinese word is a monosyllable, is clear from the tabulation.

respective averages are in parentheses at the right. Thus, the average number of syllables of all words occurring once was 3.23, of those occurring twice, 2.92, etc.

The third investigation was made by R. C. Eldridge of four samples of American newspaper English totalling 43,989 words in length and representing the occurrences of 6002 different words.* The figures in parentheses at the right (p. 28) of the two columns represent the average number of phonemes in each frequency category. For example, the average number of phonemes in all words occurring once in this investigation is 6.656; this figure was derived by dividing the sum total (i.e. 19,809) of all phonemes (estimated according to the phonemic system in use in Cambridge, Massachusetts) of all words occurring once, by the number of words occurring once (i.e. 2976).

Thus in the investigations of the three different languages, three different yet apparently equally valid units of length were employed: the morpheme in Chinese, the average number of syllables in Plautine Latin, and the average number of phonemes in American newspaper English. The difference in the unit of magnitude does not disguise the presence of the prevailing tendency which, as we shall now clearly see, is equally manifest in each of the three languages. From the evidence of these tables it is clear: (1) that the magnitude of words tends, on the whole, to stand in an inverse (not necessarily proportionate) relationship to the number of occurrences; and (2) that the number of different words (i.e. variety) seems to be ever larger as the frequency of occurrence becomes ever smaller. That is, a statistical

* This material was in part re-examined to determine as far as possible the degree of its accuracy. Instead of the figures given above it was found that the investigation represented 5995 different words (instead of 6002) which aggregated in their occurrences 43,990 (instead of 43,989). If all the steps in Eldridge's entire investigation were conducted with the remarkably high degree of accuracy as evidenced in this one instance, which is presumable, the investigation is eminently trustworthy. Regrettably it is, however, not entirely adequate for our purposes because it disregards the occurrences of all numerals and proper nouns.

CHINESE OF PEIPING

Number of Occurrences	Number of Words	Number of Words with their Syllables
1	2046	(315^1 1571^2 1443^3 144^4 1^5 1^6)
2	494	(110^1 358^2 23^3 3^4)
3	216	(59^1 147^2 9^3 1^4)
4	100	(24^1 73^2 3^3)
5	99	(39^1 58^2 2^3)
6	66	(24^1 41^2 1^3)
7	41	(16^1 25^2)
8	25	(10^1 14^2 1^3)
9	30	(13^1 15^2 1^3 1^4)
10	20	(13^1 7^2)
11	25	(14^1 11^2)
12	22	(15^1 7^2)
13	10	(6^1 4^2)
14	14	(7^1 7^2)
15	13	(5^1 8^2)
16	10	(4^1 5^2 1^4)
17	10	(6^1 4^2)
18	6	(2^1 4^2)
19	5	(4^1 1^2)
20	5	(5^1)
21	4	(3^1 1^2)
22	2	(2^1)
23	5	(4^1 1^3)
26	3	(2^1 1^2)
28	4	(3^1 1^2)
29	4	(1^1 3^2)
30	6	(4^1 2^2)
32	6	(4^1 2^2)
33	2	(1^1 1^2)
34	1	(1^1)
35	1	(1^1)
36	1	(1^1)
37	1	(1^1)
38	1	(1^1)
41	4	(4^1)
43	2	(2^1)
44	2	(1^1 1^2)
45	3	(1^1 2^2)
46	1	(1^1)
47	2	(2^1)
50	1	(1^1)
52	1	(1^1)
55	2	(2^1)
57	1	(1^1)
58	1	(1^1)
60	1	(1^1)
66	2	(1^1 1^2)
68	1	(1^1)
72	1	(1^1)
73	1	(1^1)
75	1	(1^1)
78	1	(1^1)
81	1	(1^1)
83	1	(1^1)
101	2	(2^1)
102–905	12	(12^1)
13,248	3,332	

LATIN OF PLAUTUS

Number of Occurrences	Number of Words	Average Number of Syllables
1	5429	(3.23)
2	1198	(2.92)
3	492	(2.77)
4	299	(2.05)
5	161	(2.60)
6	126	(2.53)
7	87	(2.39)
8	69	(2.44)
9	54	(2.35)
10	43	(2.32)
11	44	(2.29)
12	36	(2.30)
13	33	(2.30)
14	31	(2.09)
15	13	(2.07)
16	25	(2.40)
17	21	(2.09)
18	21	(2.04)
19	11	(2.18)
20	15	
21	10	
22	8	(2.08)
23	8	
24	9	
25	11	
26	7	
27	9	(2.00)
28	12	
29	4	
30	4	
31	8	
32	3	
33	4	
34	6	(2.05)
35	3	
36	5	
37	7	
38	2	
39	4	
40	3	
41	3	
43	4	
44	1	
45	1	
46	1	
47	3	(1.70)
48	1	
49	1	
50	2	
51	2	
53	4	
54	1	
55	1	
56	2	
58	1	
61	3	
62–514	71	(1.40)
33,094	8,437	

relationship has been established between high frequency, small variety, and shortness in length, a relationship which is presumably valid for language in general.

AMERICAN NEWSPAPER ENGLISH
(According to R. C. Eldridge)

Number of Occurrences	Number of Words	Average Number of Phonemes	Number of Occurrences	Number of Words	Average Number of Phonemes
1	2976	(6.656)	31	6	
2	1079	(6.151)	32	4	
3	516	(6.015)	33	6	
4	294	(6.081)	34	2	
5	212	(5.589)	35	5	
6	151	(5.768)	36	3	
7	105	(5.333)	37	2	
8	84	(5.654)	39	2	
9	86	(5.174)	40	4	
10	45	(5.377)	41	1	(3.903)
11	40	(4.825)	42	7	
12	37	(5.459)	43	1	
13	25	(5.560)	44	4	
14	28	(5.00)	45	1	
15	26	(4.807)	46	2	
16	17	(5.058)	47	5	
17	18	(4.166)	48	1	
18	10	(6.100)	49	3	
19	15	(4.733)	50	3	
20	16	(4.687)	51	1	
21	13		52	3	
22	11		54	1	
23	6		55	1	(3.333)
24	8		56	1	
25	6		58	2	
26	10	(3.455)	60	1	
27	9		61–4290	71	(2.666)
28	6				
29	5				
30	4				

2. THE QUESTION OF A CAUSAL RELATIONSHIP BETWEEN THE LENGTH AND FREQUENCY OF WORDS

The question now arises as to the nature of a possible causal relationship between the length of a word on the one hand, and the relative frequency of its occurrence on the other. If there is a causal relationship between relative

frequency and length which accounts for the statistical relationship just discussed, there are only two possible explanations: (1) the length is a cause of the frequency of usage, or (2) the frequency of usage is a cause of the length. That is, for example, the shortness of, say, the most frequent English word, *the*, is either (1) a cause of its high frequency of occurrence, or (2) a result of its high frequency of occurrence.

It seems that on the whole the comparative length or shortness of a word cannot be the cause of its relative frequency of occurrence because a speaker selects his words not according to their lengths, but solely according to the meanings of the words and the ideas he wishes to convey. Occasionally, of course, out of respect for the youth, inexperience, or low mentality of a particular auditor, a given speaker may seek to avoid long or unusual words. On the other hand, speakers are sometimes found who seem to prefer the longer and more unusual words, even when shorter more usual words are available. Yet in neither case are the preferences for brevity or length followed without respect for the meanings of the words which are selected. Hence there seems no cogent reason for believing that the small magnitude of a word is the cause of its high frequency of usage.

There are, however, copious examples of a decrease in the magnitude of a word which results, as far as one can judge, solely from an increase in the relative frequency of its occurrence, as estimated either from the speech of an individual in which the shortening may occur, or in the language of a minor group, or of the major speech-group. Shortenings of this sort may be termed abbreviations; these are of two types: (1) truncations, (2) substitutions, whether permanent or temporary. A consideration of these two types of abbreviation reveals that they account for practically the entire statistical relationship between magnitude and frequency, and suggest unmistakably that high frequency is the cause of small magnitude.

a. Abbreviatory Acts of Truncation

That truncation occurs primarily with frequent long words, presumably for the purpose of saving time and effort, is a proposition which is too self-evident to require demonstration. When any object, act, relationship, or quality becomes so frequent in the experience of a speech-community that the word that names it develops a high frequency of occurrence in the stream of speech, the word will probably become truncated. A development of this sort is reflected in the histories of the words *movies*, *talkies*, *gas*, which are shortenings of *moving pictures*, *talking pictures*, *gasoline*. The shortenings result from frequent usage, a frequency due to the rapid increase of frequency of movies, talkies, and gas in our daily experience.* Longer words than these, such as *constitutionality*, *quintessentially*, *idiosyncrasy* are not truncated because they are not frequently used.

There are, however, two aspects of truncation which deserve mention at this time: first, the risk of a possible homonymy arising from truncation, and second, the influence of small speech-groups upon truncations within the larger speech-community. Though the two are essentially unrelated phenomena, yet the influence of the small speech-group in minimizing the risk of homonymy, which may in turn conceivably restrain truncation, justifies their being treated together.

That the truncation of a longer word may result in an abbreviation which is homonymous with another word already in the language is not inconceivable nor will it of necessity lead to a confusion of meaning. One may safely assume that all languages have homonyms, such as English

* Frequently used proper names are very susceptible to truncation (e.g. *Whitsuntide* from *Whitsunday-tide* or *Dorchester* from *Dornwara-ceaster*). The truncation of place-names is especially interesting because often the middle member of the composition is truncated (see Otto Ritter, *Vermischte Beiträge zur englischen Sprachgeschichte*, Halle: Niemeyer, 1922, pp. 88–90).

hole and *whole*, *hear* and *here*, which are of identical phonetic form but of different usage. The differences in usages seem in most instances sufficient to obviate any serious confusion in meaning.

But though differences in syntactical usage are frequently sufficient to keep separate and distinct two words of like form and different meaning, this is not always the case. The simple statement 'I want some gas,' could in itself signify a desire for illuminating gas, for gasoline, or for 'laughing gas' (nitrous oxide). Another instance is the two homonyms *hypo*, both of which are the results of truncations: *hypo* may be a truncation of a *hypodermic injection* (*A* below), or it may be an abbreviation of *'hyposulphite of soda'* (*B* below), a name erroneously applied by early photographers to a well-known fixative (hypothiosulphate of soda). Both *hypos* (*A* and *B*) are of the same part of speech; in real life a confusion of the two might easily be disastrous. To photographers *hypo* means one thing, to physicians and trained nurses another, to perhaps the majority of English speakers it is without any meaning at all. If we were examining the vocabularies of photographers, physicians, and the general public in respect to *hypo*, we should find at least one salient difference in usage. Photographers use *hypo* (*B*) in their speech presumably much more frequently than *hypo* (*A*); physicians use *hypo* (*A*) more frequently than *hypo* (*B*); the public uses either rarely if ever. A patient suffering from heavy metal-poisoning would be more likely to receive a 'hypo of hypothiosulphate of soda' than a 'hypodermic injection of hypo' although both amount to the same thing; he certainly would not receive a 'hypo of hypo.'

From the above we may perhaps conclude that a longer word may be truncated if it enjoys a high relative frequency, not only if this high relative frequency obtains throughout the entire speech-community (*movies* for *moving pictures*) but if its use (as *hypo A* and *B*) is frequent within any special

group inside this large and inclusive speech-community. To the photographer, *hypo* means 'fixative' and he very likely calls a *hypodermic injection* by its full name. At a filling station, *gas* means 'gasoline'; at a plant producing illuminating gas, *gas* means 'illuminating gas'; in a dental clinic, *gas* means 'nitrous oxide.' And the mutually exclusive nature of many groups tend to minimize the danger of confusion which might otherwise arise from homonymy resulting from truncations.

The influence of the special group also doubtless explains the short form of many words of comparatively rare occurrence in the general stream of speech, such as English *volt*, a unit of measure of electricity. Though rare in the general speech it is doubtless of high frequency in the group of electricians and physicists by whom the word *volt* was introduced into the general stream of usage.[1] To understand its short form we must remember its high frequency in this special group in which the impetus toward brevity took place. Though a speech-community is a unit group in itself, it is also a complex of many different minor social, political, professional, economic, and even geographical groups, in each one of which there are deviations in relative frequency of usage of words from that found in the general vocabulary of the total speech community. Not only do truncations of words occur and persist in these minor groups, as we have seen in *hypo A* and *B*, but a truncated form originating in a minor group may become adopted into the language of the large community. Viewed from the average language of the larger group, the rare word appears to be unjustifiably short; yet viewed from the special group where the word is used, the word has a frequency sufficiently high to justify its shortness. The influence of minor groups, then, as we shall repeatedly observe, must be borne in mind as a possible modifying factor in the behavior of the stream of speech of the general speech-community.

Occasionally homonyms of identical usage but of decidedly

different meaning may arise through some process of linguistic change, either in a special group or in the whole speech community. What then? An example of this is the Shakespearean *let* 'to hinder' and *let* 'to allow' — words of the same part of speech and of almost opposite meaning. Today the verb *let* 'to hinder' is obsolete. On the ground of being less frequent and on occasion susceptible to confusion with *let* 'to allow,' *let* 'to hinder' was presumably dropped in favor of the synonymous, unambiguous, equally familiar, and incidentally longer *hinder*.

In concluding our brief discussion of the phenomenon of truncation, it may be said that abbreviatory acts of truncation seem to arise on the whole as a consequence of the increased frequency in usage of a word, whether within the entire speech-community or within certain minor groups thereof. The accumulated effects of abbreviatory acts of truncation during the long periods of years in which language has slowly evolved are probably responsible for the shortness of many of the frequently occurring words in speech today, and responsible, as we shall presently see, in many other ways than that which we have just observed.

b. Abbreviatory Acts of Substitution

The substitution of shorter words for longer words, such as *car* for *automobile*, or *it* for *Christmas*, has much the same net effect as truncation on the magnitude of words, and doubtless contributes extensively to the preponderance in usage of short frequent words in the stream of speech. The abbreviatory acts of substitution fall into two types: (1) the more durable substitutions which often involve a change in meaning, and (2) the temporary substitutions which we shall see are largely contextual in nature.

i. *Durable Abbreviatory Substitutions*

Durable abbreviatory substitutions may occur throughout the entire speech-community (e.g. *car* for *automobile*) or within minor groups within the entire speech-community (e.g. *juice* for *electricity*, *soup* for *nitroglycerine*, *spuds* for *potatoes*). Though one effect of substitutions of this sort may be a more or less permanent renaming (see page 274), we are now chiefly interested in the effect of such substitutions on the frequency-magnitude relationship of words in the stream of speech. If it cannot be directly proved by means of statistics that abbreviatory acts of substitution are the direct result of high relative frequency of occurrence, we can nevertheless apprehend the existence and nature of this causal relationship between high frequency and abbreviatory substitution by viewing typical examples of abbreviatory substitution against the general background of the statistics already presented.

The influence of high frequency upon the more durable substitutions is most readily observable in the substitution of a single word for a complex of words. For example, let us take the two complexes, *sweet potatoes* and *Irish potatoes*, which designate two distinctly different vegetables familiar both in the northern and southern states of the United States. In the northern states, sweet potatoes are called *sweet potatoes*, but Irish potatoes simply *potatoes*; in the southern states the reverse is true. In the South *potatoes*, 'sweet potatoes,' has been dialectally and colloquially abbreviated to *taters*; in the North *potatoes*, 'Irish potatoes,' has been similarly abbreviated to *spuds*. In the South, the sweet potato is a far more familiar article of diet than the Irish potato, and being more familiar in experience is undoubtedly more frequent in the stream of speech; in the North, the reverse is true. Surely in these two instances where all significant factors are constant except differences in frequency, one cannot but believe that the preponderant frequency in each case has led to the shortening.

In the two transitive phrases, *strike with the chin* and *strike with the foot*, there is no difference in the degree of complexity of verbal arrangement or of clarity of meaning of the concept. Yet is not the greater frequency of the second (*to strike with the foot*) indicated by the very existence of a convenient abbreviatory substitute, *kick*, which is lacking to the first? In the two concepts, *brother* and *uncle's second wife's tenth child by her first marriage*, we find the first described by one word, the second by nine. The difference in frequency of these concepts in the normal stream of speech seems alone to account for the differences in length; were the second as frequent in occurrence as the first, we should doubtless possess a single word for it — an abbreviatory substitution caused by high frequency. Though the longer example seems to be an "inherently more complex" concept than that of the single word, yet what we may term "inherent complexity of the concept" does not seem alone capable of preserving the speech-element from abbreviation; few things are more complex in nature than electricity, yet we not only have a single word for the total phenomenon but even a colloquial substitute, *juice*, a substitution presumably made because of the high frequency of usage of the concept without any respect for its complexity whatsoever.

In the minor speech-groups within the large speech-community, abbreviatory substitutions occur more frequently; witness the technical jargon and slang of the various professional, social, political, and commercial groups.[1] To the outsider the most striking aspects of the jargon, aside from its picturesqueness, are the shortness of the clique-terms, the frequency of their usage, and the unusualness of the meanings which they convey. But to the insider, the meanings are familiar and the high frequency and short length unnoticed.

It does not, of course, follow that every substitution constitutes a shortening, or that the primary conscious impulse thereto is always one of time-saving. Substitutions may be made for the sake of increased vividness of expression,[1] or of increased articulatedness of meaning. Nevertheless the frequent use of slang and technical terms, words on the whole apparently more convenient than standard language, takes place because it saves time in expression; slang and technical terms save time because these terms represent, by and large, abbreviatory substitutions for frequent concepts which, if fully articulated in standard language, would be excessively long.

The sole point of present concern, however, is not a consideration of the question of change in meaning, which will be treated later (see page 274), nor of the influence of the group upon the speech of the whole community, but rather the fact that many substitutions are shortenings resulting from high frequency. Until it can be shown that lengthenings occur from frequency or shortenings from rarity, we may reasonably presume: (1) that, where frequency and abbreviatory substitution are connected, the frequency is the cause of the abbreviatory substitution; and (2) that the accumulated effect of acts of durable abbreviatory substitution during the evolution of a language is in part reflected by the frequency-magnitude relationship of words today.

ii. *Temporary Abbreviatory Substitutions*

Substitutions of the second type, such as a pronoun for a noun (e.g. *it* for *Christmas*) or a simple adverb for an adverbial phrase, are likewise the result of high frequency. But they differ from the first (more durable) type of substitutions in one salient respect: the more durable substitutions of the first type reflect a general increase in the average relative frequency of a concept within the entire speech-community or within a minor group, while the temporary or transitory substitutions of the second type reflect merely a temporary increase in relative frequency resulting from the topic of conversation. Thus in the substitution of *car* for *automobile*, there is a high average frequency of occurrence of the concept; but in the substitution of *it* for *Christmas* in the sentence 'Christmas is a great day, *it* comes but once a year,' the substitution is the result of only a high temporary frequency which is occasioned by the nature of the context. Similarly with the substitution *there* for *down in Florida* in the sentence 'They are *down in Florida* because it is so warm *there*.' If substitutions of the first type are intimately connected with the phenomena of naming, substitutions of the second type will be found closely bound up with questions of syntax and style (see Chapter V).

Likewise with temporary abbreviatory substitutions one cannot prove statistically that frequency is the inevitable cause of all substitutions of shorter forms. Nevertheless, our feelings assure us that the substitutions of *it* for *Christmas* and *there* for *down in Florida* in the above typical sentences were made to avoid a too great repetitiveness or frequency of *Christmas* and *down in Florida* within a short period of time. The unusual frequency of occurrence of the concepts precipitated the substitutions which were in fact shorter words. It is unquestionably true that from the point of view of grammar, either the chief or a major function of many of these shorter words, such as pronouns, adverbs, and auxiliaries, is to act as substitutes. But for our present purposes it is sufficient to observe that their use may generally be viewed as abbreviatory substitutions which result from a high though transitory frequency of occurrence of the concepts for which they stand.

3. CONCLUSION: THE LAW OF ABBREVIATION

In view of the evidence of the stream of speech we may say that the length of a word tends to bear an inverse relationship * to its relative frequency; and in view of the influence of high frequency on the shortenings from truncation and from durable and temporary abbreviatory substitution, it seems a plausible deduction that, as the relative frequency of a word increases, it tends to diminish in magnitude. This tendency of a decreasing magnitude to result from an increase in relative frequency, may be tentatively named the Law of Abbreviation.

The law of abbreviation seems to reflect on the one hand an impulse in language toward the maintenance of an equilibrium between length and frequency, and on the other hand an underlying law of economy as the *causa causans* of this impulse toward equilibrium. That the maintenance of equilibrium is involved is clear from the very nature of the statistics. That economy, or the saving of time and effort, is probably the underlying cause of the maintenance of equilibrium is apparent from the fact that the purpose of all truncations and transitory contextual substitutions is almost admittedly the saving of time and effort. If one cannot argue with complete certainty in favor of economy as the sole cause of the more durable abbreviatory substitutions, one cannot readily advance any other factor as a general precipitating cause, nor escape the inference that the result of durable abbreviatory substitutions is frequently an economy of time and effort, even though this may conceivably not be the purpose. Unquestionably other factors are involved in the general phenomena of abbreviation, some of which will be subsequently discussed in considerable detail as they manifest themselves in the typical

* Not necessarily *proportionate*; possibly some non-linear mathematical function.

behavior of the phoneme, morpheme, and sentence. And from these several angles we shall also observe that the law of abbreviation is by no means restricted in its scope to the length of words.

PART II

THE BEHAVIOR OF WORDS

1. THE FREQUENCY DISTRIBUTION OF WORDS IN THE STREAM OF SPEECH

Manifestations of a tendency toward the maintenance of equilibrium in the behavior of words is not restricted to the relationship between their length and the frequency of their usage; the orderliness of the frequency distribution of the words themselves in the stream of speech suggests an analogous tendency toward the maintenance of equilibrium. But before turning to the evidence in support of this statement, let us momentarily digress in order to define a certain aspect of the term *word*.

a. The Word and the Lexical Unit

In the statistical analyses of Chinese, English, and Plautine Latin (pages 26 f.) and in the Kaeding analysis of German (pages 22 f.) the unit into which the streams of speech were anatomized was the word. For example, in English the word *child* may be considered as one word, *children* as another, *give* a third, *gives* a fourth, *given* a fifth, — five different words for each of which the respective frequencies in a given sample may be established. On the other hand, a dictionary compiled on the basis of this evidence would view *child* and *children* as but two forms of one word, and *give*, *gives*, *given* as but three forms of one word;

Zipf (1935) 172-176

tive frequency, though this important aspect of accent has never been statistically investigated.

It is perhaps worth indicating that in our investigation of the development of Latin accent from Indo-European to Classical times, the degree of accent has seemed to be a function of relative frequency and not *vice versa*. That is, it seems that a speech-element became less accented or more accented because it became relatively more frequent or less frequent, and not the reverse. So, too, it seems that the higher degree of crystallization and lower degree of distinctness of meaning have attended upon loss of accent and hence are dependent upon relative frequency. Though the author feels that this is a true statement of the causal relationship, yet in view of the difficult nature of the evidence (coming from remote times) he offers it as but his surmise.

6. RELATIVE FREQUENCY AND THE LENGTH OF MORPHEMES

Having seen that the accentual intensity of morphemes depends to a considerable extent upon their relative frequency of occurrence, let us now turn to a brief investigation of the comparative length of morphemes. In light of Chapters II and III it is not presumptuous to expect that the length or total magnitude of complexity of a morpheme will depend to a considerable extent upon its relative frequency, and for the same reasons which were valid for the magnitude of words and phonemes.

The ideal data for our present purposes would be a statistical investigation of the relative frequencies of morphemes as they occur in the stream of speech of some language over a considerable period of time; that is, an investigation of morphemes similar to, say, Eldridge's investigation of words in English. Yet an investigation of this sort has, to my knowledge, never been made. Hence, until such valuable data are

available we are forced to use the material provided for modern German by Kaeding. Though the Kaeding material has, as we shall see, decided shortcomings, it possesses, nevertheless, sufficient merit to justify presentation, and unmistakably indicates that the length of a morpheme tends to bear an inverse ratio to its relative frequency of occurrence.

In the 20 million syllables of connected German, Kaeding investigated the frequencies of (1) prefixes (*Vorsilben*), (2) stems (*Haupt-* or *Stammsilben*), and (3) suffixes (*Nachsilben*), for which he gives figures on pages 464, 424, and 468 respectively. Let us briefly consider Kaeding's treatment of each category.

In his investigation of German prefixes,[1] Kaeding included not only the autochthonous German prefixes such as *ge*, *be*, *ver*, *zu*, etc. but also prefixes borrowed into German from Latin (e.g. *extra*, *super*, *praeter*) and from the Greek (e.g. *kata*, *ortho*, *mono*, *eu*, etc.). These borrowed prefixes, though in use in German today, can scarcely be considered living morphemes, as they were in Latin and Greek; in German they are generally an integral part of the words in which they occur, without any individuality of their own. Thus, in the words *Paragraph*, *Katarrh*, *Monopol*, and the like, the average German speaker very probably is unaware of even the existence of an underlying morphological structure, and certainly of its nature. Hence we may exclude these foreign prefixes from consideration. Since they are of comparatively rare occurrence, their exclusion, moreover, matters but little.

The second disadvantage of Kaeding's investigation is for our purposes more serious though again not irreparable. Taking the morpheme *vor* for illustration, we find an entry for *vor* under the list of stems (page 423) giving 25,953 as the number of occurrences of *vor* as an independent word; on page 464 in the list of prefixes we find separate entries for the morpheme *vor*, i.e. *vor* (59,132), *voraus* (1831), *vorüber* (693), *bevor* (657), *vorher* (330), etc. An investigation undertaken for the specific purposes of establishing the relative frequencies of mor-

phemes in German would have one entry for the morpheme *vor* which would include its occurrences as an independent word and as a prefix and as a part of a prefix. Kaeding's list of suffixes and endings (page 468 ff.) meets with a similar objection: for example, Kaeding gives the frequency of the suffix-ending *-e* (1,287,899) and *-ig* (197,679) and also for the combination of the two, *-ige* (32,222). Though the additional information about the frequencies of occurrence of certain morpheme combinations is extremely valuable, information about the frequencies of occurrence of the uncombined morphemes is essential.

Nevertheless, the data furnished by Kaeding are of considerable value if used with discretion. If we limit our investigation, for example, to the thirty most frequent autochthonous German prefixes we have a list which includes 25 per cent of the different available autochthonous prefixes, and over 95 per cent of the occurrences of all prefixes. In these 30 prefixes, only *zurück* (7560) and *zusammen* (6621) consist of more than one morpheme.* If there is, then, any correlation between the lengths and frequencies of morphemes we should expect to find in this arbitrarily selected category of prefixes that the average length of the thirty most frequent prefixes would increase as we proceed down the list which in Kaeding is arranged in the order of diminishing frequency. And this correlation is indeed substantiated by Kaeding's figures. Arranging the thirty prefixes into six groups of five each, and counting the total number of phonemes (standard pronunciation) in each group, we find

Group	*Total Number of Phonemes*	*Average Number of Phonemes*
I (1 through 5)	11	2.2
II (6 " 10)	12	2.4
III (11 " 15)	12	2.4
IV (16 " 20)	16	3.2
V (21 " 25)	16	3.2
VI (26 " 30)	20	4.0

* *Zurück* is frequently condensed into what seems to be one morpheme in rapid colloquial discourse, e.g. *zrück*.

There seems here a clear tendency for average length to increase as average relative frequency decreases.

Similarly with the stems. If we select the 400 most frequent stems, which include all stems occurring more often than 5010 times which are arranged (pages 424–425) according to diminishing frequency, we have eight groups of stems, each group containing 50 stems. Again the average number of phonemes per stem in each group seems to increase as the frequency decreases:

Group	*Total Number of Phonemes*	*Average Number of Phonemes*
I (1 through 50)	136	2.72
II (51 " 100)	156	3.12
III (101 " 150)	154	3.08
IV (151 " 200)	158	3.16
V (201 " 250)	167	3.34
VI (251 " 300)	171	3.42
VII (301 " 350)	171	3.42
VIII (351 " 400)	176	3.52

Similarly with the suffixes. Disregarding Kaeding's additional notations of combinations of morphemes in the list (pages 468–469) which is arranged according to decreasing frequencies, and dividing the thirty most frequent suffixes into six groups of five each, we find:

Group	*Total Number of Phonemes*	*Average Number of Phonemes*
I (1 through 5)	9	1.8
II (6 " 10)	12	2.4
III (11 " 15)	10	2.0
IV (16 " 20)	12	2.4
V (21 " 25)	12	2.4
VI (26 " 30)	14	2.8

Here again, though not so marked, there seems a tendency for an increase in magnitude to be coupled with a decrease in relative frequency.

In spite of the obvious limitations of these Kaeding figures, they reveal a reasonably clear correlation between

the relative frequency of morphemes and their magnitudes of complexity, in this sense: the magnitude of complexity of a morpheme bears an inverse * relationship to its relative frequency. Naturally the evidence from which this conclusion is drawn is not as perfect from the point of view of our science as we might wish, but it is adequate, and the method of procedure to be adopted in making an ideal frequency count of morphemes is, I think, sufficiently clear so that in the future Dynamic Philology may check the accuracy of the above conclusion by independent investigations.

The next obvious step would be to establish a causal relationship between the magnitude of complexity of morphemes and their relative frequency in the manner adopted for words (page 29) and phonemes (page 81). But limited space forbids tarrying on this point; let us therefore assume that the magnitude of complexity depends upon relative frequency of occurrence, an assumption which in all probability is correct, and one which could be tested in the same general manner adopted in the analogous cases for words and phonemes.

The value of the assumption of the above causal relationship is twofold: first, it sheds light on the phenomenon of accentual phonetic changes (page 90), and second, it offers a possibility of determining the comparative magnitudes of many vowel phonemes. Since both of these problems may be not without some interest, let us turn to a brief discussion of them.

a. Accentual Phonetic Changes

On page 90 it was remarked that vowels sometimes undergo change in an unaccented syllable. An example of this is the familiar changes in prehistoric Latin in which an unaccented short *a* shifted in an open syllable to short *i*, while

* Not necessarily proportionate; possibly some non-linear mathematical function.

an unaccented short *a* shifted in a closed syllable to short *e*. Thus, in prehistoric times

cón-fă-cere became *cón-fĭ-cere*

cón-făc-tus " *cón-fĕc-tus*

This phonetic change in Latin was phonemically significant, since both before and after the change the short vowels *a*, *e*, and *i* were three different phonemes.

The reason for these phonetic changes in Latin was probably the general shift of accent in Latin, of which the causes were discussed on pages 163 ff. Now, if we remember (1) that the degree of accent of a morpheme tends to vary in an inverse relationship with its relative frequency, and (2) that the magnitude of complexity of a morpheme tends to stand in an inverse relationship to its relative frequency, it does not seem an overbold step to correlate these two factors in these vowel changes of prehistoric Latin. That is, the total magnitude of complexity of the root-morpheme *fac*, a typical example, was diminished, not by truncation of an entire phoneme, but by a decrease in the total magnitude of complexity of the root-vowel. In a roundabout way, then, this change in Latin can be correlated with relative frequency.

But we must be cautious in universalizing this probable correlation in Latin; for, first of all, not all vowels were changed in unstressed syllables in Latin, for example, the long vowels *ā* and *ē* (e.g. **dēlātus* > *delātus*, and **cónfēci* > *confēci*). Furthermore, modern languages that use stress as a distinctive feature are not consistent in their treatment of unstressed vowels. English, for example, generally modifies unstressed vowels; compare the pronunciation of the vowels of the similar syllables of the pairs: *cóncert, concérted; áddress* (noun), *addréss* (verb); *átom, atómic; rélay, retúrn*. On the other hand, languages like Italian, Spanish, Czechish, Polish, do not use special variants for any of the unstressed vowels: the vowel is the same, with or without stress. In the Ger-