First Edition 1935 by
Houghton Mifflin Company

First M.I.T. Press Paperback Edition, August, 1965
Second Paperback Printing, May, 1968

Printed in the United States of America

INTRODUCTION

The Psycho-Biology of Language is not calculated to please every taste. Zipf was the kind of man who would take roses apart to count their petals; if it violates your sense of values to tabulate the different words in a Shakespearean sonnet, this is not a book for you. Zipf took a scientist's view of language — and for him that meant the statistical analysis of language as a biological, psychological, social process. If such analysis repels you, then leave your language alone and avoid George Kingsley Zipf like the plague. You will be much happier reading Mark Twain: "There are liars, damned liars, and statisticians." Or W. H. Auden: "Thou shalt not sit with statisticians nor commit a social science."

However, for those who do not flinch to see beauty murdered in a good cause, Zipf's scientific exertions yielded some wonderfully unexpected results to boggle the mind and tease the imagination. Language *is* — among other things — a biological, psychological, social process; to apply statistics to it merely acknowledges its essential unpredictability, without which it would be useless. But who would have thought that in the very heart of all the freedom language allows us Zipf would find an invariant as solid and reliable as the law of gravitation?

Over the years Zipf's name has been linked to this particular statistical phenomenon until today one hears "Zipf curves" mentioned in the same offhand manner as "Bohr atoms" or "Skinner boxes" or "Bunsen burners." We pick up such tags and use them all too easily, often with little thought for the man or the work behind them. Yet each such term stands as an abbreviation for some important episode in the history of science. When a man's contribu-

tion is sufficient to earn him immortality — even this kind of anonymous terminological immortality — it is probably worth reprinting occasionally, if only to keep the record straight.

A "Zipf curve" expresses either (*a*) a relation between the frequency of occurrence of an event and the number of different events occurring with that frequency, or (*b*) a relation between the frequency of occurrence of an event and its rank when the events are ordered with respect to frequency of occurrence. All this is explained and illustrated in Chapter 2, and need not be repeated here. The point is that (for word events) Zipf found these curves to have a uniform shape under a remarkable variety of circumstances — for different topics, different authors, even for different languages — and he devoted most of his intellectual life to exploring and explaining what this regularity might signify. Although he was not the first to notice the fact, his many publications and ambitious hypotheses brought the matter to the attention of everyone with any scientific interest in language and eventually earned him the honor of having his name identified with the phenomenon he so diligently publicized.

Faced with this massive statistical regularity, you have two alternatives. Either you can assume that it reflects some universal property of the human mind, or you can assume that it represents some necessary consequence of the laws of probability. Zipf chose the synthetic hypothesis and searched for a principle of least effort that would explain the apparent equilibrium between uniformity and diversity in our use of words. Most others who were subsequently attracted to the problems chose the analytic hypothesis and searched for a probabilistic explanation. Now, thirty years later, it seems clear that the others were right. Zipf's curves are merely one way to express a necessary consequence of regarding a message source as a stochastic process.

Put it this way. Suppose that we acquired a dozen monkeys and chained them to typewriters until they had produced some very long and random sequence of characters. Suppose further that we defined a "word" in this monkey-text as any sequence of letters occurring between successive spaces. And suppose finally that we counted the occurrences of these "words" in just the way Zipf and others counted the occurrences of real words in meaningful texts. When we plot our results in the same manner, we will find exactly the same "Zipf curves" for the monkeys as for the human authors. Since we are not likely to argue that the poor monkeys were searching for some equilibrium between uniformity and diversity in expressing their ideas, such explanations seem equally inappropriate for human authors.

A mathematical rationalization for this result has been provided by Benoit Mandelbrot. The crux of it is that if we assume that word-boundary markers (spaces) are scattered randomly through a text, then there will necessarily be more occurrences of short than long words. Add to this fact the further observation that the variety of different words available increases exponentially with their length and the phenomenon Zipf reported becomes inescapable: a few short words will be used an enormous number of times while a vast number of longer words will occur infrequently or not at all.

So Zipf was wrong. His facts were right enough, but not his explanations. In a broader sense he was right, however, for he called attention to a stochastic process that is frequently seen in the social sciences, and by accumulating statistical data that cried out for some better explanation he challenged his colleagues and his successors to explore an important new type of probability distribution. Zipf belongs among those rare but stimulating men whose failures are more profitable than most men's successes.

The mathematical explanation for the form of the "Zipf

curve" leaves the whole question of word frequencies in an anomalous theoretical position. On the one hand we know that the form of his curves follows necessarily from the assumption that word-length is a random variable — or, in Zipf's words, that "a speaker selects his words not according to their lengths, but solely according to the meanings of the words and ideas he wishes to convey" — so that no further assumptions about an equilibrium between hypothetical forces working toward uniformity and toward variety of expression need be invoked. But, on the other hand, we recognize that any language in which short words were not the most frequent would be grossly inefficient. Moreover, psychologists have demonstrated that the frequency with which a word is used can powerfully influence the accuracy with which we hear it, read it, memorize it, associate to it, or use it appropriately in our own speech. It is impossible to believe that nothing more is at work to guide our choice of letter sequences than whatever random processes might control a monkey's choice, or that the highly plausible arguments Zipf puts forward have no relevance at all. In order to avoid overstating the case, therefore, we should put it negatively: whatever the social or psychological influences may be that move us toward efficiency in our use of words, the shape of the "Zipf curve" is irrelevant to them. If a statistical test cannot distinguish rational from random behavior, clearly it cannot be used to prove that the behavior is rational. But, conversely, neither can it be used to prove that the behavior is random. The argument marches neither forward nor backward.

Because his fame has been so closely tied to the puzzle of the curves, the fact that these curves were but one facet of Zipf's work is often forgotten. He was above all else a man with a vision — a naturalistic vision — of human language, a vision he attempted to actualize in statistical form. Without the support his word counts gave it, his vision

must now seem less compelling to us than it did to him, but it is still worthy of study and discussion.

Begin with the assumption that a concept is a bundle of semantic features, or "genes of meaning," as Zipf called them. If a particular bundle occurs frequently in the cognitive life of a community, they will assign to it some phonological representation, or "word." If it occurs infrequently, no word will be available, so the bundle will have to be made up as needed from strings of words arranged in phrases. As the language evolves, a relation will develop between the lengths of the phonological representations and the frequencies of occurrence of the bundles they express. These ideas still have considerable currency among students of psycholinguistics, on grounds having little or nothing to do with "Zipf curves" or probability theory. It is difficult to see how to develop or test them without a much better theory of syntax than was available in 1935, but nevertheless they are ideas that merit consideration in their own right. It is even possible that statistical analysis may yet prove relevant if we can learn how to look behind those surface symptoms called "words" to the deeper cognitive processes that underlie them.

So it is good to have this book in print again. Perhaps now we are ready to look beyond its statistical puzzles to some of the underlying issues it raises concerning the cognitive aspects of linguistic behavior. That may not be a prospect to inspire a poet, but it is difficult to imagine any way a scientist could probe closer to what is uniquely human about human beings.

George Kingsley Zipf was born in Freeport, Illinois, on January 7, 1902. He graduated *summa cum laude* from Harvard College in 1924 and spent the following year in Germany, studying at Bonn and Berlin. He returned to Harvard and received his Ph.D. in Comparative Philology

in 1930; then became Instructor in German until 1936, Assistant Professor of German until 1939, and University Lecturer until 1950.

His Ph.D. dissertation was concerned with relative frequency of use as a determinant of phonetic change in the evolution of language, a topic revisited in Chapter 3 of the present book. *The Psycho-Biology of Language* was his first attempt to relate his linguistic ideas to man's experience as a whole. In 1941 he published *National Unity and Disunity,* which applied his statistical methods to the study of the sizes of cities and movements of population. His most ambitious work, *Human Behavior and the Principle of Least Effort* (1949), was a further study of semantics, psychology, sociology, geography; it abounds with illustrations of the probability distributions he first noticed in his statistical studies of vocabulary.

Zipf died on September 25, 1950.

GEORGE A. MILLER

Harvard University
April, 1965

PREFACE

NEARLY ten years ago, while studying linguistics at the University of Berlin, it occurred to me that it might be fruitful to investigate speech as a natural phenomenon, much as a physiologist may study the beating of the heart, or an entomologist the tropisms of an insect, or an ornithologist the nesting-habits of a bird. That is, speech was to be regarded as a peculiar form of behavior of a very unusual extant species; it was to be investigated, in the manner of the exact sciences, by the direct application of statistical principles to the objective speech-phenomena. The stream of speech, whatever it might represent to the historical grammarian, the comparative philologist, or the descriptive phoneticist, was to be viewed as but a series of communicative gestures. The findings of the extensive investigation that resulted are now presented in full. They are presented, moreover, intentionally in such a manner that they will, I think, be readily available, not only to the professional linguist, but to any serious reader interested in linguistic phenomena, whether his interest be from the angle of the biological, sociological, or psychological sciences, or from the angle of aesthetics and *belles lettres*.

Perhaps nothing will more conveniently illustrate the nature, scope, and appeal of the material about to be discussed than the brief presentation of a few typical examples from our findings. For example, it can be shown that the length of a word, far from being a random matter, is closely related to the frequency of its usage — the greater the frequency, the shorter the word. It can furthermore be shown either from speech-sounds, or from roots and affixes, or from words or phrases, that the more complex any speech-element is phonetically, the less frequently it occurs. As an illustration of the high degree of orderliness with which linguistic forces operate, the frequency distribution of words in English may

the psycho-biology of language: an introduction to dynamic philology

george k. zipf

THE PSYCHO-BIOLOGY OF LANGUAGE:
An Introduction to Dynamic Philology

by George K. Zipf

This introduction to the field of dynamic philology presents the findings of an unusual and extensive study of speech as a natural phenomenon. The author has approached his subject in the manner of the exact sciences – by treating speech as a specific form of behavior in a unique species and by directly applying statistical principles to the objective speech-phenomena. Regarding the stream of speech as simply a series of communicative gestures has resulted in a number of suggestive findings, for example: the author discovered an interesting correlation between the length of a word and the frequency of its usage – "the greater the frequency the shorter the word." He also found that in any given language a fundamental condition of equilibrium exists between the form and function of speech habits or patterns. In addition, evidence points to the fact that the desire to preserve or restore this equilibrium is a significant contributory cause of the continual change in language. "By change is meant not only changes in phonetic form and accent, but changes in meaning, in emotional intensity, in syntactical arrangement." Since the need to express meanings and emotions dictates the content of our speech, this study attempts to examine "the problems of meaning, emotion, and of mental behavior in general . . . to investigate the forces of the mind by viewing linguistic phenomena in the stream of speech as manifestations of the forces of the mind in the process of functioning."

This unique approach to the field of linguistics will interest any serious reader concerned with the subject, whether he interprets the findings from the biological, sociological, psychological, or aesthetic point of view.

George K. Zipf was Lecturer at Harvard University for many years and the author of *Selected Studies of the Principle of Relative Frequency* and *National Unity and Disunity.*

THE M.I.T. PRESS
Cambridge, Massachusetts 02142