Zipf's Law and Miller's Random-Monkey Model

Davis Howes


Stable URL:
http://links.jstor.org/sici?sici=0002-9556%28196806%2981%3A2%3C269%3AZLAMRM%3E2.0.CO%3B2-F

The American Journal of Psychology is currently published by University of Illinois Press.
NOTES AND DISCUSSIONS

ZIPF’S LAW AND MILLER’S RANDOM-MONKEY MODEL

Some years ago G. A. Miller published a note in this Journal^ purporting to prove that Zipf’s law for the distribution of word frequencies in natural languages^ can be explained as a consequence of random placement of the spaces or silences that demark successive words in text. His argument, along with the imputation that Zipf’s law is little more than a statistical artifact and therefore devoid of theoretical interest, has been widely accepted by psychologists interested in linguistic phenomena. In this note I wish to point out that the proof given by Miller actually rests on assumptions that are clearly untenable for natural languages.

Let us review Miller’s argument briefly. He considers a mechanism that generates random sequences of the letters of the alphabet and the space symbol. To give life to the argument Miller asks us to imagine the proverbial monkey striking the keys of a typewriter at random. A word is defined as any sequence of letters bounded by spaces. Miller then shows, by a straightforward argument based upon the calculus of probabilities, that the distribution of word frequencies produced by such a monkey will obey Mandelbrot’s equation^ for Zipf’s law,

\[ p(w) = b\{r(w) + c\}^{-d}, \]

where \( p(w) \) is the probability of the \( r \)-th most frequent word in the sample, and \( b, c, \) and \( d \) are constants. That people follow Zipf’s law so accurately, Miller concludes, “is not really very amazing, since monkeys typing at random manage to do it about as well as we do.”

If Zipf’s law indeed referred to the writings of ‘random monkeys,’ Miller’s argument would be unassailable, for the assumptions he bases it upon are appropriate to the behavior of those conjectural creatures. But to justify his conclusion that people also obey Zipf’s law for the same reason, Miller must perforce establish that the same assumptions are also appropriate to human language. In fact, as we shall see, they are directly contradicted by well-known and obvious properties of natural languages.

---

* Received for publication February 15, 1968.
1 G. A. Miller, Some effects of intermittent silence, this JOURNAL, 70, 1957, 311-314.
4 Miller, op. cit., 313.
The proof cited by Miller depends upon a strict ordering of word probabilities by length of word. All possible words of the same length are assumed to have the same probability, and every word of a given length is assumed to have a higher probability than any word of greater length. These assumptions appear explicitly in Miller's article. They are, moreover, logically necessary to his entire argument. The probability \( p(w,i) \) of any particular word composed of \( i \) letters, Miller says, is the probability of finding any string of \( i \) successive letters divided by the number of different words of that length that are possible with an alphabet of \( A \) different letters. This definition leads to the equation from which the rest of his proof follows,

\[
p(w,i) = \frac{p(*)p(L)^{i-1}}{A^i}
\]  

in which \( p(*) \) denotes the probability of the monkey's striking the space bar and \( p(L) \) the probability of his striking any of the \( A \) letter-keys. But Equation 2 obviously holds only if all \( A \) letters are equally probable and independent; that is, when all possible words of length \( i \) have the same probability, \( p(w,i) \).

Now, in natural languages it is well known that the frequencies of words of the same length vary over an enormous range. In English, for example, the three-letter word 'the' occurs more than 236,000 times as frequently as the three-letter word 'lop' according to the Lorge Magazine Count.\(^5\) One can scarcely imagine what would be the ratio for the three-letter word 'dzq,' which according to the random-monkey model also must have the same probability as 'the.' Longer words like 'understand' and 'government,' each of which occurs about 1000 times in the Magazine Count, are even more anomalous. By Miller's assumptions, each of these words (along with the other \( 10^{14} \) possible 10-letter words) should have a probability of about \( 2.74 \times 10^{-16} \). That any word with so small a probability could occur 1000 times in a sample the size of the Magazine Count is beyond belief. To be precise, the probability of that event, which can be calculated directly from Miller's assumptions, turns out to be less than \( 10^{-11468} \)—a figure whose only possible meaning is that the assumptions upon which it is based are beyond belief.

Equation 2 shows further that Miller's model requires every word of length \( i \) to be of higher probability than every word of greater length,

since the denominator increases and the numerator decreases with $i$. But in natural languages many relatively long words, such as 'government' and 'understand,' have much higher frequencies than most shorter words, which, like 'lop' or 'dzq,' have very small frequencies. Zipf demonstrated that in natural languages the average frequency of words of length $i$ is greater than the average frequency of words of length $(i + 1)$.

Miller has converted this empirical finding into the assumption that all words of length $i$ have probabilities greater than all words of length $(i + 1)$, which is obviously untenable.

The full significance of the assumption that all sequences of letters are equiprobable appears to have escaped Miller. In his concluding paragraph (p. 314), he twice states that his derivation follows from the "simple" assumption of "random placement of spaces" without mentioning the more critical—and vulnerable—assumption of random sequences of letters. The same error is apparent when he speaks of the amount of redundancy as affecting the parameters of Equation 1. Since that expression is derived from Equation 2, it assumes equal and independent letter-probabilities. Redundancy is then zero by definition; its introduction in any amount therefore would contradict the postulates from which the equation was derived.

The proof Miller gives is actually a special case of the more general derivation of Zipf's law previously published by Mandelbrot. Indeed, Mandelbrot had already indicated the line of argument taken by Miller as well as the reasons for discarding it. He then went on to develop a proof that does not require the assumptions that are contrary to fact. To do so he assumed that the language-mechanism maximizes information, in the Shannon sense, for a given cost of coding. (He also proved the theorem from other similar assumptions.) Miller feels that these assumptions "strain one's credulity"; Mandelbrot's use of a maximizing assumption particularly bothers him. But it is not quite accurate to claim, as Miller does on p. 313, that his own assumption of random spacing "replaces" Mandelbrot's maximizing assumption. Information is by definition maximized when alternatives are assumed to be equiprobable. As we have noted, the crucial assumption of Miller's argument is not that spaces occur at random, but that all sequences of letters are equiprobable. In other words, Miller has not really replaced the assumption of maximization, he only introduces it in disguise. It is no coincidence that Miller's

---

*Mandelbrot, loc. cit.
'random monkey' follows Mandelbrot's equation: the creature has been defined as an information-maximizer.

To account for Equation 1 as an expression for Zipf's law, then, we must return to Mandelbrot's original theory. The use of a maximizing criterion in his proof can hardly be accepted as a serious objection. The theory is broadly conceived and deserves more attention than it has received. Whether Mandelbrot's equation is the most satisfactory description of Zipf's law, however, is open to question. Simon's birth-process model does not fit the data as well as Mandelbrot's, as Miller says,$^9$ but the log-normal distribution does.$^{10}$ This is a matter that can only be resolved by empirical investigation. In any case, it is clear that Zipf's law cannot be dismissed as merely a statistical artifact, the senseless product of a monkey's random typing.

Aphasia Research Center
Boston University School of Medicine

FRAGMENTATION OF A PROLONGED, STRUCTURED AFTERIMAGE VIEWED BY NAIVE OBSERVERS

When a geometrical figure such as a square is seen in steady fixation, partially stabilized with a contact lens, or as a prolonged afterimage, O commonly reports the intermittent disappearance and reappearance of the figure as a whole or in part.$^1$ The intermittent disappearance and reappearance of parts of the figure has been termed 'fragmentation.' Most studies of this phenomenon typically have involved a small number of sophisti-

---


* Received for publication September 25, 1967. The author expresses his indebtedness to the Ontario Department of Education and to the University of Waterloo for providing the facilities for this work, and to M. P. Bryden, G. E. MacKinnon, E. Zurif and J. Forde, all of the Department of Psychology, University of Waterloo, for their help in collecting the data.