Zipf's Law and the Structure and Evolution of Languages

Least effort entities and the most common words

BY A.A. TSONIS, C. SCHULTZ, AND P.A. TSONIS

A.A. Tsonis and C. Shultz are at the Department of Geosciences, University of Wisconsin-Milwaukee, Milwaukee, WI 53201. P.A. Tsonis is at the Department of Biology, University of Dayton, Dayton, OH 45469.

The power law $y=cx^a$ can be viewed as the solution of the differential equation $dy/y=a dx/x$, which says that if x changes from x_1 to x_2 ($x_2>x_1$), then y is magnified by a factor $(x_2/x_1)^a$. Such properties are appropriate to scale invariant processes [3] arising in complex systems and are often thought of as least effort entities in natural phenomena [6,7]. In 1949, George K. Zipf argued that human behavior is often dictated by the principle of least effort [8]. He reasoned that since languages are means of transmitting information, their structure should be optimal, thus allowing transmission with the least effort. He demonstrated this by showing that languages obey the power law $f \alpha r^a$ where $a \approx -1$, f is the frequency of each word, and r is the word rank (the most frequently used word having rank 1, the second most frequently used word having rank 2, and so on).

This view has since been challenged by studies involving random text properties. It was shown that random texts (i.e., texts produced by randomly selecting letters from an alphabet of M nonspace and one-space characters) should obey similar power laws [2,4]. Recently it was shown that the true distribution of words in random texts is not a power distribution but a lognormal distribution with an inverse power law tail [5]. In such cases, the true lognormality is obscured and an apparent power law emerges unless the sample size is very large. In either case, one may argue that Zipf's law can be derived without appeal to least effort.

However, even though the random text studies provide a general framework for the study of artificial words, they hardly relate to actual languages. First, in random texts, all combinations of N letters constitute a word of length N. This is not true with actual languages. As such, the theoretically derived word distributions do not approximate the actual distributions.

Second, in random texts, the probability of a given word is a function of its length. This is definitely not true in real languages. For example, if one considers M=26, then it follows (since the space character has a probability equal to 1/27) that, in random texts, very short and very long words will be the least probable. In the English language, the most common words consist of one to three letters (e.g., *the, I, to, of*).

Finally, random texts cannot account for the structure and evolution of modern languages. In fact, if one thinks about it, letters have nothing to do with languages. Letters have been devised to delineate the sounds and phonetics of words. Zipf's law can, in fact, be derived without letters by simply using a tape recorder. As such, we believe that the least effort angle cannot be dismissed on the basis of results from random text studies, and we proceed by providing a slight revision to the least effort principle suggested by Zipf.

Since Zipf's publication, many other studies have reported power laws in modern languages (see [1] for a commentary with examples). In all studies, the linear regression of lnf on lnr appears to be very good over the whole range of ranks. Because of this overall good

fit, an important deviation at small scales (leading ranks) has gone uninvestigated. A careful look at the results of all studies (including Zipf's) reveals that the power law $f \alpha r^{-1}$ overestimates the frequency of the most common words. If one accepts the statistical characteristics of random texts, one may argue that this is the result of trying to fit a power law to a lognormal distribution. Given, however, our previous arguments concerning random text word distributions, we propose here another explanation. According to our explanation, this feature is not part of the scaling in modern languages, but rather an important aspect of language evolution and structure.

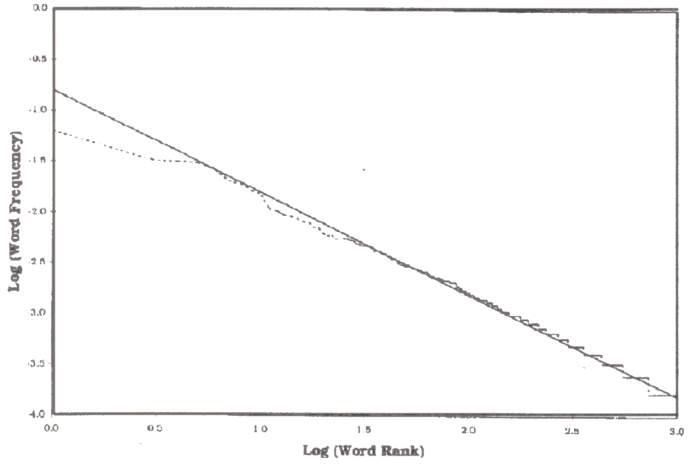
Figure 1 shows word frequency versus word rank (dashed line) for $1 \le k \le 1000$. The solid line is the power law fit in that range of ranks. Including ranks greater than 1000 does not influence the results at all. The data in this figure are based on a chapter of the book *Complexity* by M. Mitchell Waldrop, which is about 20,000 words long. The slope of the regression is -1.001, as expected. In agreement with previous studies, we find *the, of, a, to, and* to be the five most common words. Also in agreement with the other studies, we observe the deviation of the power law close to the origin. The sum of squares $\varepsilon^2 = \Sigma(f_r - \hat{f}_r)^2$ (where f_r is the observed frequency at rank r, and $\hat{f}_r$ is the predicted frequency by the power law fit) is equal to 0.0110921. This is a very small error considering that n=1000 and indicates that the power law is a good fit to the data. However, it turns out that an exponential law ($f=ce^{ar}$) or a linear model ($f=c+ar$) produces comparable ε^2 values (0.0101859 and 0.0100336, respectively). This result indicates that even though the power law may be an adequate fit, its ability to predict the actual values is not, on the average, better than other laws. The single reason for this result is that, compared to the other models, the power law does a very poor job in representing the values at the origin.

When the same analysis is repeated for $6 \le r \le 1000$, we find that the exponent *a* changes to -1.06 and we estimate that $\varepsilon^2 = 0.0002764$. A dramatic improvement in ε^2 is observed as it decreases by a factor of about 40. Excluding words with ranks greater than five results in a further, but not as dramatic, improvement. This kind of behavior is not observed with the other two models (in the range $6 \le r \le 1000$, ε^2 for the linear model is equal to 0.0024278 and for the exponential model is equal to 0.0022594). Thus, the power law becomes a superior fit only when the most common words are excluded .

Accordingly, we conclude that Zipf's law is valid but not for the most common words. If it were, then the English language would have three times more *the*'s, two times more *of*'s, etc. This would have made the language rather awkward and possibly not an efficient means of communication. Since the majority of the most common words are, as called by linguists, "function" words, we hypothesize that languages evolved in a manner consistent with the principle of least effort; however, in that process, a few "coupling" words outside the power law had to be invoked. Those words depended on the structure of the language and, as a result, they may not be the same in every language. This will explain why, for example, the article *the* is the most common word in English, appears in more than ten different forms in Greek, but does not even exist as a word in Japanese.

An interesting question that arises from this commentary is whether such findings apply to other complex systems that exhibit least effort properties (such as ecosystems) and Zipf-like power laws. Similar breaks at small scales may provide insights into the dynamics and evolution of these systems.

FIGURE 1



Zipf's Law for the English language for word ranks from 1 to 1000.

REFERENCES

1. J. L. Casti: Bell curves and monkey languages: when do empirical relations become a law of nature? Complexity 1(1): pp. 12-15, 1995.
2. B. B. Mandelbrot: On the theory of word frequency and on related Markovian models of discourse. In: Structures of language and its mathematical aspects. R. Jacobson (Ed.) American Mathematical Society, New York, 1991.
3. B. B. Mandelbrot: The fractal geometry of nature. W.H. Freeman and Co., New York, 1983.
4. W. Li: Random texts exhibit Zipf's-law-like word frequency distribution. IEEE transactions on information theory. 38: pp. 1842-1845, 1992.
5. R. Perline: Zipf's law, the central limit theorem, and the random division of the unit interval. Phys. Rev. E. 54: pp. 220-223, 1996.
6. A. A. Tsonis: Some probabilistic aspects of fractal growth. J. of Phys. A. 20: pp. 5025-5028, 1987.
7. A. A. Tsonis: Dynamical systems as models for physical processes. Complexity. 1(5): pp. 23-33, 1996.
8. G. Zipf: Human behavior and the principle of least effort. Addison-Wesley, Cambridge, MA, 1949.