Comments to "Bell Curves and Monkey Languages", J. Casti, Complexity,1,12-15(1995). by Wentian Li

Whether there are universal laws or principles in complex systems is a fascinating and important question. Prof. John Casti uses the case of Normal Distribution ("bell curves") to illustrate that such universal principle is perhaps out there waiting to be discovered [1]. He suggests Zipf's law as a candidate for such universal principle. But as the author of one of the three publications to prove that "monkey languages" exhibit Zipf's law [2] (B. Mandelbrot and J. Nicolis wrote the other two [3,4]), I tend to believe that Zipf's law is a "law of transformation" (if it can be called a "law") rather than a "law of complex systems".

I restrict the name "Zipf's law" to rank-frequency distributions that are power-law function $P(r) \sim 1/r^{\alpha}$, $\alpha \approx 1$ (r is the rank, and $P(r)$ is the frequency of events corresponding to that rank). Besides the rank-frequency distribution of English words and "monkey typed words", other rank-frequency distributions that obey power-law functions with exponent close to 1 were also known, including the population of cities [5].

The non-linguistic examples given in Ref.[5] may not have an explanation yet, but the Zipf's law in "monkey languages" has already been explained and understood. This explanation also shed light on the Zipf's law in English language. To summarize in one sentence, the Zipf's law in "monkey languages" is caused by an exponential transformation of variables, and if the first variable follows an exponential distribution, the second would follow a power-law distribution. To be more specific, the first variable is the word's length, and the second variable is the word's rank.

The details of this proof can be found in Ref.[2], but I can sketch the proof here easily: suppose there is an exponential function of variable x': $f(x') = Aexp[-Bx']$, and a transformation from x' to x: $x = Cexp[Dx']$, the new function in variable x will be:

$$f(x) = A\left(\frac{C}{x}\right)^{B/D}$$

which is a power-law function! The exponent of the power-law function is close to 1 when $B \approx D$

We can also consider a probability density function: $p(x')dx' = Aexp[-Ax']dx'$ (B is replaced by A so that the probability density function is normalized to 1) with x' ranging from 0 to ∞. When the same transformation $x = Cexp[Dx']$ is applied, making $p(x')dx' = p(x)dx$, the probability density function for variable x becomes:

$$p(x) = \frac{AC^{A/D}}{D}\left(\frac{1}{x}\right)^{A/D+1},$$

again, a power-law function.

Although I do not have a proof, but it is conceivable that other Zipf's laws can be derived by similar variable transformations.

To end this letter, I would like to comment on the claimed "Zipf's law in non-coding DNA sequences" [6] as mentioned in Ref.[1]. The rank-frequency distribution of short segments of DNA sequences with a fixed length as determined in [6] can be hardly called a Zipf's law: not only the exponent is not close to 1, but also the power-law function is not well obeyed. The distribution is actually better modeled as a "Yule distribution" [7].

References

1. J. Casti, "Bell curves and monkey languages," Complexity, 1(1), 12-15 (1995).
2. W. Li, "Random texts exhibit Zipf's-law-like word frequency distribution," IEEE Transactions on Information Theory, 38(6), 1842-1845 (1992).
3. B. Mandelbrot, "An informational theory of the statistical structure of language", in *Communication Theory*, ed. W. Jackson (Academic Press, 1953).
4. J. Nicolis, *Chaos and Information Processing* (World Scientific, 1991).
5. G.K. Zipf, *Human Behavior and the Principle of Least Effort* (Addison-Wesley, 1949).
6. C. Mantegna, et. al. "Linguistic features of noncoding DNA sequences", Physical Review Letters, 73, 3169-3172 (1994).
7. C. Martindale, A.K. Konopka, "Oligonucleotide frequencies in DNA follow a Yule distribution", to be published in Computers and Chemistry (1996).