ELSEVIER

Physica A 216 (1995) 148–150

PHYSICA A

# The distance between Zipf plots

Shlomo Havlin

*Department of Physics, Bar-Ilan University, Ramat-Gan 52900, Israel*

Received 15 December 1994

**Abstract**

We study the Zipf plots describing the occurrence of different elements in a given group as a function of their rank. We define a "distance" between two Zipf plots characterizing the differences between the two groups. We apply the distance concept on groups of words contained in books. Our results suggest that the distance between books written by the same author is smaller than the distance between books written by different authors.

Very recently, there has been considerable interest in the study of Zipf plots in various topics ranging from analyzing linguistic texts to studying of DNA base pair sequences [1–3]. The Zipf analysis considers the occurrence frequency, $f$, of each element in a given group [4]. The elements could be, for example, different words contained in a given book or different sets of consecutive symbols contained in a given sequence. All elements in the group are arranged in decreasing rank order, i.e. in decreasing order of occurrence. In the Zipf plot, the occurrence frequency, $f$, is plotted vs. the rank, $R$. Zipf has found that for languages the following approximate scaling behavior exists, $f \sim R^{-\zeta}$, with $\zeta \cong 1$.

Here, we define a "distance" between two Zipf plots characterizing two groups of elements. This distance is a quantitative measure of the rank differences of the same elements appearing in both groups. For simplicity let us define the distance between two books. First, one counts the occurrence of each word in each book (denoted as book 1 and book 2). In each book we assign a rank $R = 1$ to the most frequent word and a rank $R = 2$ to the second most frequent word and so on. If two different words have the same occurrence they obtain the same rank[1]. By this, we obtain a Zipf plot for each book, $f_1(R)$ for book 1, and $f_2(R)$ for book 2. The Zipf plots for both books look usually similar, however, the same words may have different rank in both books. Let

[1] An alternative definition is to give the two words two consecutive ranks, where the order is chosen randomly.

Elsevier Science B.V.
*SSDI* 0378-4371(95)00069-0

$R_1(\lambda)$ be the rank of word $\lambda$ in book 1 and $R_2(\lambda)$ be the rank of the same word in book 2. We define the distance $r_{12}(\lambda)$ between the ranks of $\lambda$ in the two books as

$$r_{12}(\lambda) \equiv \left[(R_1(\lambda) - R_2(\lambda))^2\right]^{1/2}. \tag{1}$$

The distance between the *two books* is defined as the mean square root distance between the ranks of all common words,

$$r_{12} \equiv \left[\frac{1}{N}\sum_{\lambda}(R_1(\lambda) - R_2(\lambda))^2\right]^{1/2} = \left[\frac{1}{N}\sum_{\lambda} r_{12}^2(\lambda)\right]^{1/2}. \tag{2}$$

Here, $N$ is the total number of common words, $\{\lambda\}$, that appear in both books, within a maximum rank, $R_{\max}$, which is a chosen parameter.

We calculate, using Eq. (2), the distance between each pair of books from a set of 9 books written by three different authors, 3 books by each author [5] [2]. We find that the distances between books written by the same author are usually smaller than the distances between books written by different authors. We calculate the mean distance between books written by different authors (27 cases) and between books written by the same author (9 cases). The mean distance is found to be $\langle r\rangle = 16.1 \pm 1.3$ between books written by the same author and $\langle r\rangle = 21.8 \pm 2.8$ between books of different author.

This result suggests that each author has his own hierarchy of words – i.e. typical rank for words, which he uses in his books. This measure might be useful to test how close (in the rank) are different books or even different languages. In particular, it might be tempting to use the distance concept to help to identify authors of books which authorship are questionable.

We wish to thank C.-K. Peng for bringing his attention, after this work has been completed, to Ref. [6] that provides a somewhat different definition of a distance between two Zipf plots. In Ref. [6] the total number of common words, $N$, in the definition of the distance between two texts, Eq. (2), does not appear. We wish to thank C.-K. Peng, S. Rabinowitz and H.E. Stanley for useful discussions and the Israel–USA Binational Science Foundation and the Israel Academy of Sciences for financial support.

## References

[1] R.N. Mantegna, S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng, M. Simons and H.E. Stanley, Phys. Rev. Lett. 73 (1994) 3169;
R.N. Mantegna, S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng, M. Simons and H.E. Stanley, Linguistic analysis of coding and noncoding DNA sequences, Phys. Rev. E (submitted);
A. Czirók, R.N. Mantegna, S. Havlin and H.E. Stanley, Correlations in binary sequences and generalized Zipf analysis, Phys. Rev. E (submitted).

[2] W. Ebeling and T. Pöschel, Europhys. Lett. 26 (1994) 241;
H. Herzel, E. Ebeling and A.O. Schmitt, Phys. Rev. E 50 (1994) 5061.

[2] In all cases we chose $R_{\max} = 100$, the number of common words in this range was $N = 100 \pm 20$ for the books of the same author and $N = 85 \pm 15$ for the books of different authors.

[3] M.H.R. Stanley, S.V. Buldyrev, S. Havlin, R. Mantegna, M.A. Salinger and H.E. Stanley, Zipf plots and the size distribution of firms, Economics Lett. (in press).

[4] G.K. Zipf, Human Behavior and the Principle of the Least Effort (Addison-Wesley, New York, 1949); C.E. Shannon, Bell Syst. Tech. J. 27 (1948) 379; 30 (1951) 50.

[5] The chosen books are by H.G. Wells: Dr. Moreau (43607 words total, 5299 different words), The Time Machine (32345, 4586), The War of the Worlds (60428, 6999); Jules Verne: 20,000 Leagues Under the Sea (101840, 8178), Around the World in 80 Days (64507, 6767), From the Earth to the Moon (93110, 7977); Mark Twain: The Adventures of Huckleberry Finn (114288, 6202), The Adventures of Tom Sawyer (73902, 7427), What is Man? and Other Essays (122263, 10543).

[6] W.H. Press, B.P. Flannery, S.A. Teukolsky and W.T. Vetterling, Numerical Recipes (Cambridge Univ. Press, Cambridge, 1988).