Probability distribution in a quantitative linguistic problem

F. Calderón, S. Curilef* and M. L. Ladrón de Guevara†
Departamento de Física, Universidad Católica del Norte,
Av. Angamos 0610, Antofagasta, Chile

(Received on 23 March, 2009)

In the present contribution, we propose a possible way to discuss the distributions of words in a given text. We have devoted our study to discuss some relevant properties observed in Spanish texts of Latin-American writers. We start analyzing the appearance of distributions of the frequency of occurrence in the Zipf perspective. We identify two regions of behavior separated by a special point. In order to correctly define such a point, we work beyond the Zipf law, defining other probability distribution that takes the frequency of repetition of a particular word among other different words into account. At this point, we take the linguistic problem to a statistical level. We make an effort to characterize the point of separation between two regions, via the Binder cumulant of fourth order, as it is made in the characterization of critical points in phase transitions of physical systems.

Keywords: Probability distribution, Zipf law, Binder cumulants

1. INTRODUCTION

A large number of mutually interacting parts characterizes complex systems, which are often open to their environment and they self-organize their internal structure and their dynamics with sometimes surprising macroscopic properties. The probability profile is often the first quantitative characteristics of complex systems. A first approximation for this profile covers the Gaussian law, the power law and the stretched exponential distributions. Nowadays, much attention has been paid to power law and its relevance is explained because it approaches to critical phenomena from the statistical perspective[1–5].

One example of power law distributions [6] constitutes the Zipf law, and its several deformations[2]. This law describes a variety of distributions in nature; for instance, diversity of biology newsgroups, time intervals of earthquakes, processes in collections of journal papers, size of cities, distribution of words in texts, and others. Applications to quantitative linguistics is understood owing to an empirical observation on certain statistical regularities of human writings that has become the most well-known statement of this topic. There are several approaches devoted to the understanding of the quantitative linguistics[7, 8], however a consistent approach to explain it remains open.

In this contribution, we briefly review the Zipf law in the context of the linguistic problem. We analize the validity of this law for the distributions of frequencies of words in books of Latin-American writers. In all the studied cases, we observe two regions of behavior in the frequency distributions around a special point. To understand this feature we go beyond the Zipf law, defining a new probability distribution, which is the probability of finding a word in a given text. Thus, we lead the linguistic problem to the statistical mechanical description. In addition, we characterize the point of change of behavior via the Binder cumulant of fourth order, as it is made in the characterization of critical point in phase transitions of physical systems. In this way, we expect that any possible irregularity in the appearance of the representation of typical parameters of the present system might be correctly characterized by these cumulants.

2. FREQUENCY OF OCCURRENCE: ZIPF LAW

We start introducing some historical definitions in quantitative linguistics; in particular, these are the frequency of occurrence and the rank of words. The frequency of occurrence is the number of times that a particular word appears in a given text corpus. The rank is the place that a particular word takes in a histogram ordered by decreasing frequency. Thus, Zipf proposed an approximate mathematical relation between the frequency of occurrence of a word and its rank by the following power law

$$f(s) \propto \frac{1}{s^{\alpha}}, \qquad (1)$$

where $f(s)$ is the frequency, the label s is the rank of words, and α a fitting parameter that characterizes the text in the Zipf law.

Previous efforts have been made in order to verify the validity of this power law in texts that are written in languages such as English (see for instance Ref.[8]), Indian[9], etc., and some deviations from the Zipf law have been already appointed[2, 3, 10].

We have analyzed several books by different authors, but we present results of four of them. In FIG. 1, the frequency of occurrence $f(s)$ is depicted as a function of the rank s for the distribution of words in given texts of writers of our interest. We observe that the frequency of occurrence exhibits a departure from the standard trend of the Zipf law. The most frequent words seem to be separated from the less frequent words by a particular point around $s = 30$, which we are interested in defining and characterizing correctly. Therefore, we can suggest that the Zipf law is approximately verified whenever the full trend is divided in two parts of behaviour. Each part is characterized by a different fitting parameter α. In the inset of FIG. 1, we depict the same feature for normalized distributions. This unique profile suggests a particular use of the Spanish language by Latin-American writers. The

*scurilef@ucn.cl
†mlladron@ucn.cl

set of the most frequent words is distributed in a different way in comparison with the less frequent words.

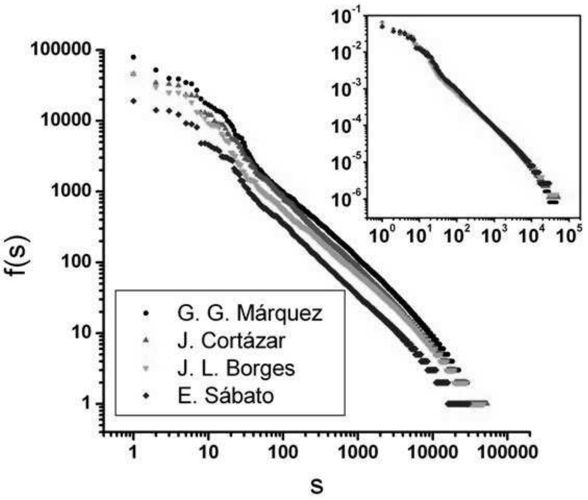

FIG. 1: The frequency of occurrence $f(s)$ as a function of the rank s is depicted for complete plays of four Latin-American writers. In the inset, the normalized frequency of occurrence $f(s)/N$ as a function of s is depicted to emphasize the similarity of the profile of the particular using of the Spanish by different writers.

3. FREQUENCY OF REPETITION: BEYOND THE ZIPF LAW

Without forgetting the historical perspective on linguistics, but going beyond the Zipf law, we can define a new histogram. Extracting from the text the quantity of words between two equal words, we define a frequency of repetition of a word separated by a determined number of words. In the present model, we propose a possible expansion for each frequency of occurrence in terms of the frequency of repetition $C_s(r)$, considering r as label of the number of words between two equal words. As before s represents the rank of the word. This allows to rewrite the Eq.(1) as follows:

$$f(s) = \sum_{r=1}^{N_s} C_s(r), \tag{2}$$

where N_s is the number of couple of successive words labeled by s.

We have numerically obtained the profile of the distributions $C_s(r)$. Certainly, the corresponding normalization requirement is given by

$$1 = \frac{1}{f(s)} \sum_{s=1}^{N_\alpha} \sum_{r=1}^{N_s} C_s(r), \tag{3}$$

where N_α is the number of different words of the text. Thus, we can define a new probability distribution in the following manner:

$$p_s(r) = \frac{C_s(r)}{f(s)}, \tag{4}$$

which represents the probability of finding two equal s words separated by r other different words. Obviously, at this point, we lead the linguistic problem to a statistical level. Several properties of statistical interest can be defined and evaluated in a similar manner to the standard statistical systems. The word “de” is the most frequent word in almost all the considered books and we choose it to represent the typical profile of the distribution. In FIG. 2 the probability distributions of finding the word “de” and “él” (in English “of” and “he”, respectively) are depicted as a function of r, the number of words between two consecutive words “de” or “él”. From the plot shown in FIG. 2, we may remark the following:

- The shape of the probability is a distribution with a peak. In each distribution, the value of r_{max} seems to increase as s increases. On the contrary, the height of the peak decreases as s increases.
- The tail of the function is a straight line, which suggests an exponential tail. Its slope decreases as s increases.
- The tendency of the curve for frequent words as “de” is clearer than for less frequent words as “él”.

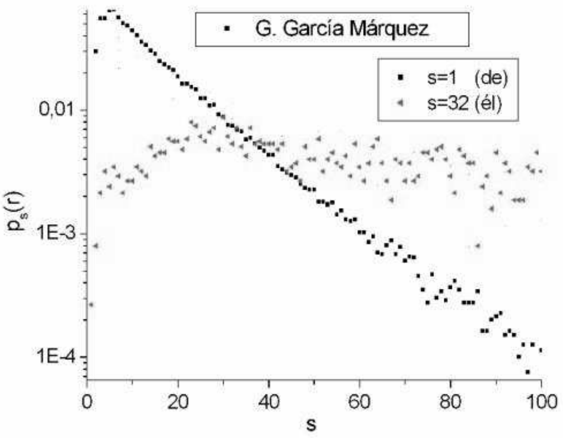

FIG. 2: A typical distribution of the probability of repetition $p_s(r)$, as a function of the number of words r between two equal words labeled by the rank s, is depicted for complete plays of one author. We show the probability of repetition $p_s(r)$ for two different words ($s = 1$, and 32), where the peak in r_{max} apparently increases as s increases

In order to define some properties that we have observed in the behavior of this kind of systems, we evaluate some statistical quantities that we discuss in the following Section.

4. MOMENTA AND BINDER CUMULANTS

We can assess momenta by using the distribution of the Eq.(4) obtained nuumerically. They are given by

$$\left\langle r^l \right\rangle = \sum_r r^l p_s(r), \tag{5}$$

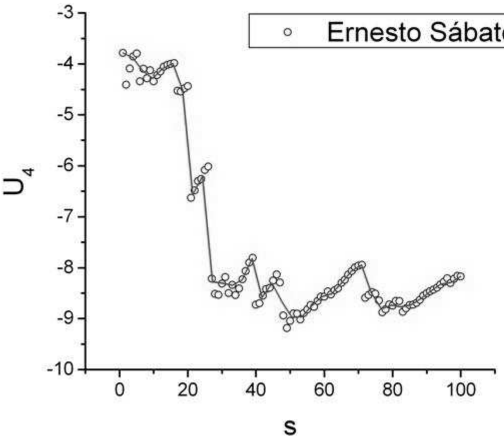

FIG. 3: A typical Binder cumulant of fourth order is depicted as a function of s. A change of behaviour is observed close to $s = 20$.

where l is the order of the momenta. Now, with the aim of applying the description that leads to a treatment similar to the critical point in thermodynamics for physics description of systems we evaluate the Binder cumulant of fourth order. This is

$$U_4 = 1 - \frac{\left\langle (r - \langle r \rangle)^4 \right\rangle}{3 \left\langle (r - \langle r \rangle)^2 \right\rangle^2}. \quad (6)$$

In FIG. 3, U_4 is depicted for a particular author in order to illustrate the behaviour of this cumulant. In magnetic system, the rank of U_4 is $[0,1]$. This is not true in the present case because the behaviour of the distribution of our order parameter is different to the distribution of the order parameter for magnetic systems. Our evidences declare that Latin-American authors, who we are here analyzing, use the language in a particular way. They use a few words giving a breakdown of the tendency of the distribution of the frequency of words in the corpus of the text. We identify this change close to $s = 20$. This behavior is similar to all authors who we are analyzing.

From a grammatical point of view, in the first region we mainly find Spanish prepositions, as “de”, and other monosyllables as conjuctions, articles or pronouns.

5. SUMMARY

We have made an effort to develop a possible way to describe this kind of systems defining a new probability distribution, which corresponds to the probability of repetition. As a consequence of this fact, we can bring our discussion to a statistical level.

In this picture we may identify two regions, where the curve of the frequency of occurrence presents mainly two different slopes. In the region of the most frequent words we find some of them that are characteristic of the language, which are common for all authors that we have studied. However, in the second region, we expect to find words of particular use of the author.

The trend of the standard deviation is not clear. This feature suggests an additional calculation, which is the Binder cumulant of fourth- order. Due to the trends of this cumulant we can ensure the existence of the two different regions in the curve.

Acknowledgment: Authors would like to thank partial financial support by FONDECYT, grants 1051075 and 1080660.

[1] S. Abe and N. Suzuki, Physica **D 193**, 310 (2004)

[2] M. Montemurro, Physica **A 300** 567 (2001)

[3] R. F. I. Cancho, Eur. Phys. J. **B 44**, 249 (2005)

[4] R. F. I. Cancho, Eur. Phys. J. **B 47**, 449 (2005)

[5] T. M. Cover, AND R. C. King, IEEE Trans. on Inf. Theory **24** (1978)

[6] M. E. J. Newman, Contemporary Physics 46, 323(2005)

[7] T. M. Cover and R. C. King, IEEE Transactions on Information Theory 24, 413 (1978)

[8] A. P. Masucci and G. J. Rodger, Phys. Rev. E 74, 026102 (2006)

[9] B. D. Jayaram and M. N. Vidya, Journal of Quantitative Linguistics 15, 293 (2008)

[10] B. Mandelbrot, Information theory and psycholinguistics: a theory of words frequencies, in: P. Lazafeld, N. Henry (Eds.), Readings in Mathematical Social Science, MIT Press, Cambridge, MA, 1966.