Physica A 392 (2013) 2775–2780

Contents lists available at SciVerse ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Menzerath–Altmann law for distinct word distribution analysis in a large text

Sertac Eroglu [*]

Department of Physics, Eskisehir Osmangazi University, Meselik, 26480 Eskisehir, Turkey

HIGHLIGHTS

- We showed the Menzerath–Altmann law describes the distinct word distribution in corpora.
- The observation/prediction comparison shows excellent accuracy.
- The distinct word distribution characteristics are language independent.
- We showed Menzerath–Altmann law is the special case of gamma distribution

ARTICLE INFO

Article history:
Received 16 July 2012
Received in revised form 13 February 2013
Available online 7 March 2013

Keywords:
Language
Distinct word distribution
Menzerath–Altmann law
Generalized gamma distribution

ABSTRACT

The empirical law uncovered by Menzerath and formulated by Altmann, known as the Menzerath–Altmann law (henceforth the MA law), reveals the statistical distribution behavior of human language in various organizational levels. Building on previous studies relating organizational regularities in a language, we propose that the distribution of distinct (or different) words in a large text can effectively be described by the MA law. The validity of the proposition is demonstrated by examining two text corpora written in different languages not belonging to the same language family (English and Turkish). The results show not only that distinct word distribution behavior can accurately be predicted by the MA law, but that this result appears to be language-independent. This result is important not only for quantitative linguistic studies, but also may have significance for other naturally occurring organizations that display analogous organizational behavior. We also deliberately demonstrate that the MA law is a special case of the probability function of the generalized gamma distribution.

1. Introduction

The MA distribution law has attracted considerable amount of scientific interest, particularly in the quantitative linguistic studies, for its existence in linguistic organizations at various levels, such as the length of words and the length of their morphemes observed in phonemes [1], the length of sentences and the length of their clauses observed in words [2], and the length of texts and the length of their constituting sentences observed in words [3].

Although the MA law has become one of the fundamental stochastic laws in quantitative linguistics, the use of the MA law reaches far beyond the detection of organizational regularity in linguistics. For example, it has been observed that the organizational regularity in musical texts [4] and in genomes [5–9] is measurable and the regularity obeys the MA law in various levels of structural organization.

In the quantitative linguistics literature, there are numerous studies discussing the word distributions in terms of words' occurrences (word frequencies) in a text by using statistical laws such as Zipf's law and Menzerath–Altmann law [10–13]. Nevertheless, Menzerath–Altmann regularity has not been studied comprehensively on the entire organizational levels of

[*] Tel.: +90 222 239 3750/2817.
E-mail address: seroglu@ogu.edu.tr.

0378-4371/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.physa.2013.02.012

natural languages. In this study, the objective is to apply the MA law in the framework of word distribution organization that has not yet been investigated, i.e., the distribution of distinct or different words (vocabulary stock) of a text rather than their occurrences.

The paper proceeds as follows: Section 2 introduces the MA law and discusses the forms and parameters of the law by means of other well known distribution functions. In Section 3 the sources of data and data acquisition procedure are described, and the data are presented. The agreement between the observed data; i.e., distinct word distributions of two corpora, and the prediction of the distributions described by the MA law are discussed in the final section. The final section also presents the distribution characteristics of the corpora, and some drawn conclusions on the interpretation of the MA law's parameters, which attempts to extend our understanding of the linguistic organization in the level of distinct words.

2. The MA law

In this section, we briefly introduce the MA law from historical and theoretical perspectives. Paul Menzerath, an experimental psychologist and phonetician, was one of the pioneering researchers who initiated quantitative linguistics research. In 1954, Menzerath concluded that there is a negative correlation between the length of a linguistic construct and the length of the construct's constituents; i.e., the longer a linguistic construct the shorter its constituents [14]. Altmann, in his seminal work "Prolegomena to Menzerath's Law" [15], converted Menzerath's significant observation into a mathematical form and this equation has since been called the MA law.

The MA law is a continuous probability distribution model that is used to describe a probabilistic relation between the discrete outcomes of quantities; i.e., the relation between the size of the construct, y, and the size of the construct's constituents, x, and its most general form is given by:

$$y\left(x\left|A,b,c\right.\right)=Ax^{b}e^{-cx}, \tag{1}$$

where A, b and c are free parameters to be determined empirically. There are two interesting cases:

(a) Where $b=0$ and $c\neq 0$ in Eq. (1) indicates that the construct size is an ordinary exponential function of the constituents' sizes given by two parameters:

$$y\left(x\left|A,c\right.\right)=Ae^{-cx}. \tag{2}$$

(b) Where $b\neq 0$ and $c=0$ in Eq. (1) yields the well-known equation for Zipf's law which is the particular class of the power law

$$y\left(x\left|A,b\right.\right)=Ax^{-b}. \tag{3}$$

In linguistics, Zipf's law establishes a very simple relationship between the frequency rank of a construct's constituents, x, and the corresponding occurrence frequencies $y(x)$ where the observed value of b is usually close to unity. The law principally states that the higher the rank, the fewer the occurrences [16]. Note that the sign of the parameter b in Eq. (1) is positive which represents increasing tendency of relevant property of the construct, while in Zipf's form (Eq. (3)) it is negative which corresponds to decreasing tendency of relevant property of the construct.

Although Eq. (3) has been applied on many linguistic levels [17,18] and other investigations [19,20] have discussed the linguistic elucidation of the parameters of the MA law given by Eq. (1), the explicit interpretation of the parameters in the linguistics framework remains somewhat controversial and there is a lack of consensus among researchers on accepted usage.

For this study, we used the most general form of the MA law Eq. (1), which in fact is a compact representation of the probability function $P(x)$ of the generalized gamma distribution; i.e., a general type statistical distribution function [21,22]. For $x\in[0,\infty)$, $P(x)$ is given by

$$P\left(x\left|\alpha,\beta,\theta\right.\right)=\frac{\beta}{\theta\Gamma\left(\alpha\right)}\left(\frac{x}{\theta}\right)^{\alpha\beta-1}e^{-\left(\frac{x}{\theta}\right)^{\beta}}, \tag{4}$$

where the parameters α, β and $\theta\geq 0$, and $\Gamma\left(\alpha\right)$ is the complete gamma function. Now consider the case $\beta=1$. In this case the generalized gamma distribution function reduces to the gamma distribution function which is from the family of two-parameter continuous distribution functions, and it is defined as

$$P\left(x\left|\alpha,\theta\right.\right)=\frac{x^{\alpha-1}}{\theta^{\alpha}\Gamma\left(\alpha\right)}e^{-\left(\frac{x}{\theta}\right)}. \tag{5}$$

The distribution parameters α and θ are often called the shape parameter and the scale parameter, respectively. The shape parameter determines the existence and height of the distribution peak, whereas the scale parameter is responsible for describing the spread of the distribution. In probability theory and statistics the gamma distribution (its mean and variance are equal to $\alpha\theta$ and $\alpha\theta^2$, respectively) is frequently implemented to model the waiting times between the occurrences, e.g., estimating the waiting times between earthquakes [23].

Note that, Eq. (1) can be recovered from Eq. (5) by redefining the parameters of the MA law in terms of the gamma distribution parameters as; $b\equiv(\alpha-1)$, $c\equiv(1/\theta)$ and $A\equiv\left[\theta^{\alpha}\Gamma\left(\alpha\right)\right]^{-1}$. Furthermore, the parameter A, in Eq. (1), serves as a normalizing constant for a given particular distribution.