On Mathematical Modeling of Pattern of Occurrence of Various Constitutional Components of Language

Hemlata Pande and H. S. Dhami*

***Abstract*—Despite the apparent freedom to express our views, substantial extent of written and spoken texts of natural languages generally have consistencies for the pattern of its various components i.e. they tend to obey some simple regulations which can be, by analyzing the texts, transmogrified in the form of mathematical models and equations to put in a theoretical framework. Present paper is an attempt to present an account for the applications of different mathematical models employed for modeling the pattern of occurrence of diverse constitutional components of natural languages.**

***Index Terms*—Model, frequency, distribution, language, corpus, text.**

***MSC 2010 Codes* – 97M80, 91F20**

I. Introduction

HUMAN language is a remarkable communication system which may be generally divided into two parts: the lexicon and grammar. Lexicon corresponds to the list of words, list of parts of words and of phrases with their associated meanings (which are used to form any sentence) and the grammar relates to the rules to unite different lexicon items into correct sentences and more complex phrases. If we look back into the history for path breaking researches in the field of linguistics then we come across the work of Harris [1], who presented the procedure for analyzing linguistic data depending on two steps: identification of phonological and morphological elements and description of the distribution of the elements relative to each other and of Noam Chomsky [2] , who in his book entitled "Syntactic Structures" has attempted for scientific theory-construction and presentation of comprehensive theory of language which according to the experts is considered identical to the theories of biological and chemical systems.

Similar to natural sciences, in language study also, different kinds of quantitative data can be collected and is required for example: in comparison of theoretical and observed data to present theory of language, in determination of the performance of different teaching methods by linguistic teachers, in study the pattern of texts in stylistics. Linguistics is the study of a language scientifically, and the discipline related to applications of mathematical tools and techniques for the scientific study of a language is known as Mathematical Linguistics. It includes work at every level of linguistic structure. The applications of mathematical techniques to language has been attempted by Lambek ([3]-[6]) ; Harris[7]; Jakobson (Ed.)[8]. As stated by Bolshakov and Gelbukh [9] "in the broader view, mathematical linguistics is the intersection between linguistics and mathematics, i.e., the part of mathematics that takes linguistic phenomena and the relationships between them as the objects of its possible applications and interpretations". "In Mathematical Linguistics, the objects of interest are the collection of words in a language, the collection of sentences, the collection of meanings etc.", as mentioned by Kornai ([10]: page 10). In statistical or quantitative linguistics, language is studied by the process of determining the frequencies of various structures in different texts. Altmann [11] has mentioned that in Quantitative linguistics, researcher tries to find a law in language as likely a physicist tries to do in the nature. The function of quantitative linguistics is the determination and formulation of the laws which explicate the observed and described facts of language(s). It supplies the techniques of making conclusion in text processing on the base of previously collected statistics. Meyer [12] has mentioned that "Quantitative Linguistics is concerned with accounting for quantifiable and measurable linguistic phenomena in terms of mathematical models such as curves, probability distributions, time series and the like".

According to Köhler and Altmann ([13], p.12) "Quantitative Linguistics studies the multitude of quantitative properties which are essential for the description and understanding of the development and the functioning of linguistic systems and their components". They have also mentioned in page 13 that the "properties of linguistic elements and their interrelations abide by universal laws, which can be formulated in strict mathematical way - in analogy to the laws of the well-known natural sciences"; according to them once a language or text has been viewed from quantitative point of view, the features and interrelations can be determined "which can be expressed only by numbers or rankings ....". Within the field of quantitative linguistics, languages can be viewed as structures which are subjected to the evolutionary methods in similar to biological organisms etc. Wilson [14] pointed out

Hemlata Pande is Post Doctoral Fellow in the Department of Mathematics, Kumaun University, S. S. J. Campus Almora, Almora-263601, Uttarakhand, India. (E-mail: hlpande@rediffmail.com)

H. S. Dhami is Professor in the Department of Mathematics and Campus Director, Kumaun University, S. S. J. Campus Almora,Uttarakhand, India. (E-mail: drhsdhami@gmail.com).

* The study reported in the paper has been supported by the University Grants Commission (UGC), New Delhi, INDIA under the 'UGC Dr. D. S. Kothari Post Doctoral fellowship scheme'to the first author. [Grant No. F.4-2/2006(BSR)/13-770/2012(BSR)].

that "the main task of quantitative linguistics is to attempt to explain, and express as general language laws, the underlying regularities of linguistic structure and usage". According to Manning and Schütze ([15], p. 5): "we can learn the complicated and extensive structure of language by specifying an appropriate general language model, and then inducing the values of parameters by applying statistical, pattern recognition, and machine learning methods to a large amount of language use".

For the purpose of formulation of pattern of occurrence of various linguistic components by different models and distributions, first there should be a source text and the data for the desired pattern can be accumulated from the text. A body of text is known as corpus and several such collections of texts form corpora. When we have a part of a text or a corpus, it can be analyzed quantitatively. For example, in the case of the text mentioned in the appendix, the text has lots of quantitative information in it as: in the text we have 106 word tokens, 78 word forms/types, 66 words occurring once, 07 words occurring twice; the text has 432 letters of English language alphabet with the letter 'a' occurring 33 times and so on. There is one 1-length word, 22 2-length words, 26 3-length words and so on if the length is measured in terms of the number of letters contained in the word etc. Different forms of quantitative linguistic data has been mentioned in the book by Butler[16].

A mathematical model is a depiction and explanation of any system with the help of mathematical concepts and language. A model can be useful in determining predictions about behaviour of a system, in determination of the effects of different components and explication of the system. Linguistic Models stand for patterns to illustrate a language and its various aspects (phonology, grammar, lexicon) in order to explain more accurately linguistic concepts and their relationships. Mathematical models in linguistics have been applied in different Natural language processing tasks for example the vector space model is used in information retrieval, Markov model used is parts-of-speech tagging (for a survey, we can cite the book of Manning & Schütze [15]) and for the formulation of pattern of different linguistic components in terms of different equations.

As the properties of languages are stochastic, quantification and probabilistic models play a vital role in their formulation. In the present paper, we have reviewed various mathematical models used for modeling the pattern of occurrence of different fundamental components of natural languages.

The ensuing section of the paper discusses the modeling of pattern of components by introducing a new factor 'rank' to the components. In this case the components are arranged in descending order of the values, to obtain a rank/frequency profile, and then the types of components in the frequency list are replaced with their frequency-based ranks, by assigning rank 1 to the most frequent type, rank 2 to the second most frequent etc. In the section III, various models for the formulation of occurrence of components directly from the components' statistics have been presented.

II. Rank Frequency Approach

Altman [17] has mentioned that: "Zipf's idea is the foundation stone of modern Quantitative linguistics", "His influence is not only restricted to linguistics but also incessantly penetrates other sciences". Zipf[18] has ranked words with respect to their frequency of occurrence in order to find a relation between the frequency of occurrence of a word and its rank. Under the title of Zipf's laws in the introductory chapter, Manning and Schütze [15] have mentioned that the rule applicable to counting of words of a language in a large corpus can be stated as: 'if the frequency of occurrence of word (type) is ranked in descending order of frequency, then

$$f \propto \frac{1}{r} \tag{1}$$

where f is the frequency of word of r^{th} rank[1] '. The generalization of this rule has been given again by him in the form:

$$f = \frac{a}{r^b} \tag{2}$$

where a and b are parameters of text and b is close to unity. Thus if the logarithm is taken of equation (2) both sides, Zipf's law predicts that rank/frequency profiles will appear as straight lines in double logarithmic space i.e., it describes log(frequency) as a function of log(rank). The authors ([15]) have also introduced the fact that Zipf's formula has been modified by Mandelbrot [19] to specify the rank frequency distribution for word's frequencies in the form:

$$f = \frac{a}{(r+c)^b} \tag{3}$$

where a, b and c are parameters. Later on the Zipf's law was tested for various languages by researchers, for example Hatzigeorgiu et al [20] in the case of Greek language, for common words and common lemmas, have utilized the Zipf law; Ha et al [21] applied it for 'English' and 'Mandarin' and Mikros et al [22] for Modern Greek. Köhler [23] determined the fact that for Hungarian text, the rank frequency distribution of words verifies the Zipf-Mandelbrot law. For the application of Zipf law and for the determination of Mandelbrot constants in the case of Turkish language, the work of Dalkiliç and Çebi [24] can be cited who have used large scale Turkish corpus for the purpose. Tuzzi et al [25] have tested the validity of Zipf's law in Italian texts by analyzing a corpus compiled by the end of year addresses delivered in the period 1949-2008 by ten presidents of Italian Republic. They have concluded that the Zipf's law is an adequate model and the corpus has a unique style and they have also found a position for each president on the synthetism/ analytism scale. Situngkir [26] reported the statistical observation of Zipf's law and Zipf Mandelbrot law to different human languages by using biblical texts, translated into many languages and observed the statistical properties of each language. The researcher pointed out that "statistical differences are discovered between English and widely used national and several ethnic languages

[1]In the entire article almost all symbols connected to f and F are corresponding to frequencies or normalized frequencies and symbols related to P and p correspond to probabilities except or otherwise defined in a separate form.

in Indonesia". By selecting four languages to represent Indo-Aryan and Dravidian families of languages: Hindi and Marathi representing Indo-Aryan and Kanada and Telugu representing Dravidian, Jayaram and Vidya[27] have applied the Zipfian approach and used the right-truncated zeta distribution of the form

$$P_x = \frac{1}{x^a T} \tag{4}$$

x =1, 2.....R, R is the truncation parameter at the right hand side of distribution and T is the normalizing constant and its value is $T = \sum_{j=1}^{R} j^{-a}$. For the purpose the researchers have collected the data under five main categories: 'aesthetics', 'commerce', 'natural physical and professional sciences', 'official and media languages' and 'social sciences'. They concluded that all texts from each language and each genre follow the right-truncated zeta distribution except for one text in the category of Telugu-Commerce.

The rank frequency approach was also applied to different components, other than words of language and the formulation of the frequency with the help of rank has been given by various equations. In the context we can cite the works of Sigurd [28] for suggesting a geometric series equation for the distribution of phoneme frequencies ('phoneme, in linguistics, smallest unit of speech distinguishing one word (or word element) from another'[2]) and of Good[29] for offering equation to describe the ranked distribution of phoneme and grapheme frequencies('grapheme is the smallest semantically distinguishing unit in a written language, analogous to the phonemes of spoken languages'[3]). As cited in the book by Popescu et al ([30], p.128),

$$P_x = pq^{x-1} \tag{5}$$

x=1,2....., represents the one displaced geometric distribution used by Sigurd [28]; and Good[29] has suggested equation to describe the ranked distribution of phoneme and grapheme frequency in the form

$$f = \frac{1}{n} \sum_{i=r}^{n} \frac{1}{i} \tag{6}$$

where n is the number of symbols.

Gusein-Zade [31] has given a model relating to the ranked frequency series of the letters in a language in the form:

$$F_r = \frac{1}{n}(\log(n+1) - \log r) \tag{7}$$

where n is total number of letters in the language alphabet. The researcher has shown that "the frequency distribution of the letters in the Russian language fits this model". Borodovsky and Gusein-Zade [32] have also suggested the equation (7) in the case of codon frequencies for total n symbols.

By taking two languages English and Mandarin, Ha et al [21] tested the Zipf law for single words and for n-gram word phrases and have investigated that in the case of single words, the law is applicable for high frequency words and if n-gram phrases and single words are merged together in one list and arranged according to the frequencies then the joint list follows Zipf's law for both languages. This application of Zipf's law for the combined data corresponding to single words and n-gram phrases has been mentioned by them as extended form of Zipf law. For the two languages: English and Chinese, Ha et al [33] concluded that in case the single word unigram distributions is taken, deviation occurs from Zipf law and when the frequency distribution of words is united with the distribution of 2-grams, 3-grams, 4-grams and 5-grams, the joint Zipf curve follows the Zipf law for all ranks and frequencies in cases of both the languages. They have also checked the nature of Zipf curve for Chinese syllables and English 2-byte and 3-byte substrings by taking n-grams and concluded that this 'extended form of Zipf's law also holds for the syllables of Chinese (as well as for 2-byte and 3-byte word fragments in English), even though the distribution of syllable unigrams is very different from the distribution for words'.Dalkiliç and Çebi [24] have tested the Zipf law for the monogram, bigram, trigram, tetragram and pentagram graphs for Turkish corpora and in the case of Turkish word frequencies they have concluded that Turkish satisfies Zipf Mandelbrot law.

Grzybek and Kelih [34] have conferred theoretical model for grapheme frequencies in the case of Slavic alphabets and the negative hypergeometric distribution (NHGD) has been recommended as an adequate model. For the rank frequency distribution of grapheme frequencies, they have used the NHDG in 1-displaced form which can be expressed in the form:

$$P_x = \frac{\begin{pmatrix} M+x-2 \\ x-1 \end{pmatrix} \begin{pmatrix} K-M+n-x \\ n-x+1 \end{pmatrix}}{\begin{pmatrix} K+n-1 \\ n \end{pmatrix}} \tag{8}$$

K, M are parameters of a text , x is the rank, P_x is frequency of grapheme of rank x and n represents the total number of graphemes of Slavic alphabets. Zipf's approach has been applied for the letters instead of words by Eftekhari [35]. The researcher has defined two new terms 'Zipf's order' and 'Zipf's dimension' for letters by using English language texts. The order of letters formed by their arrangement in accordance with their frequencies of occurrence (ascending order) has been taken as the Zipf's order. For the text 'Hamlet' the Zipf's order mentioned by the researcher is 'j z x q v k b p g f c w y m u d l r n h s i a o t e' and the corresponding numerical value of Zipf's order (i) for these letters are taken numbers from 1 to 26 respectively. The researcher has applied the power law

$$P_i \sim \frac{1}{i^a} \tag{9}$$

with the value of the exponent a close to unity, for the relation between the Zipf's order 'i ' and corresponding frequency of occurrence 'P_i ' of a letter. The exponent a has been defined as the Zipf's dimension and the researcher proposed it for the relative study of texts besides the fractal dimension; as mentioned in the paper, the fractal dimension is obtained by sorting the letters in the alphabetical order and then application of power law relation between order and frequencies.

[2]From http://www.britannica.com/EBchecked/topic/457241/phoneme
[3]From Wikipedia http://en.wikipedia.org/wiki/Grapheme

Tambovtsev and Martindale [36] promulgated that in the case of phoneme frequencies, Yule equation

$$f = \frac{a}{r^b}c^r \tag{10}$$

where a , b, c are parameters, fits the distribution superiorly than any other distributions like Zipf, Good etc. Martindale et al [37] have compared rank frequency distributions of graphemes and phonemes by inspection of 32 text corpora from 18 languages. They have shown that letter and phoneme frequencies both are well explained by equation, first developed by Yule and the equation determined by Borodovsky and Gusein-Zade. These two equations have already been discussed above [equation (7) & equation (10)]. Macutek [38] has applied a discrete distribution, which is a generalization of the right truncated geometric distribution to model the frequencies of graphemes for the languages: Russian, Slovak, Slovene and Ukrainian in the form:

$$P_x = cp^{x-1}\left(1 + \frac{a}{n-x+1}\right) \tag{11}$$

x = 1, 2 ..... n, with parameters $p \geq 0$ & $a \geq -1$ and c is normalizing factor. The researcher has also applied the adapted probability distribution $\{Q_x\}$, for modeling frequencies with the help of corresponding ranks of graphemes in Tamil, of the form:

$$\begin{cases} Q_1 = 1 - \alpha(1 - P_1) \\ Q_x = \alpha P_x, \quad x = 2, 3, ...n \end{cases} \tag{12}$$

where $0 < \alpha < (1 - P_1)^{-1}$ and $\{P_x\}$ according to the researcher is the distribution defined by above equation (11). We [39] in our earlier studies have modified the Zipf Mandelbrot law and have determined the rank-frequency model in the form of following equation for the frequencies of occurrence of graphemes of English language alphabet in various texts

$$F_r = a(r^2 + k_r r + c)^{-b} \tag{13}$$

where F_r is the frequency corresponding to rank r; a, b ,c are parameters and the parameter k_r takes different values in different layers of rank r. In another work we ([40]) have applied the rank frequency approach to discuss the mode of occurrence of letters in texts and in the first position of the words of texts for Hindi texts, by using their Zipf's orders, rank frequency profile and model for their rank and frequency. For the frequencies of various letters in different texts and in initial positions of words of texts it has been concluded that frequencies of letters in a text as well as in combination of several texts follow Yule distribution given by equation (10) and in the case of their frequencies for word's initials, the frequencies can be more properly expressed by a distribution of kind:

$$F_r = \frac{a}{(r-\delta)^b}c^r \tag{14}$$

where δ is either zero or has a value between zero and one, and a , b and c are parameters. By using Zipf's order of letters and words' initials, we have concluded that 'क , त , न , म , र , स ,ह are seven most frequent letters of Hindi language' and 'क , प ,म , स , ह are five most frequent word's initials' for the words of different texts.

Li et al [41] tested the performance of several regression models on the log-frequency-log-rank scale for numerous ranked linguistic data, such as: ranked letter distribution, ranked inter-word spacing distribution and ranked word frequency distribution for the normalized frequencies from the 'Mody-Dick' English text of the author H. Melville's. They have applied Gusein-Zade equation, expressed as (7) in the present paper, Weibull equation:

$$p_{(r)} = C\left(\log\left(\frac{n+1}{r}\right)\right)^a \tag{15}$$

power-law equation, expressed previously as equation (2),

Exponential equation:

$$p_{(r)} = Ce^{-ar} \tag{16}$$

Beta equation:

$$p_{(r)} = C\frac{(n+1-r)^b}{r^a} \tag{17}$$

Yule equation, expressed previously as (10)

Altmann equation:

$$p_{(r)} = Cr^b e^{-\frac{a}{r}} \tag{18}$$

and Mandelbrot equation, expressed previously as (3).

They have concluded that Beta function is appropriate for the ranked frequency data for letters while Yule function fits the ranked distribution for word spacing and for the word ranked frequency distribution. Considering 'Mody-Dick' text for English and 'Don Quijo' text for Spanish, they have concluded that Altmann, Beta, Yule functions are better than the Zipf's power law, expressed by equation (2).

For ranked letter frequency distributions, Li and Miramontes [42] have examined the statistical model choice by utilizing ten functions for English and Spanish. In case of English they have used US Presidential Inaugural Speech texts for the 44 presidents (in the last 1-2 centuries) and for Spanish, 19 Mexican presidents' addresses to congress (Informes Presidenciales) from 1914 to 2006. The ten functions utilized by them were Gusein-Zade, power-law, exponential, logarithmic, Weibull, quadratic logarithmic, Yule, Menzerath-Altmann/Inverse-Gamma, Cocho/Beta and Frappat. Equations for functions, other than logarithmic, quadratic logarithmic, Inverse gamma and Frappat have already been discussed above in this section while the formulae for these four functions as used by the researchers (for the normalized frequency f)are:

Logarithmic:

$$f = C - a\log(r) \tag{19}$$

Quadratic logarithmic:

$$f = C - a\log(r) - b\left(\log(r)\right)^2 \tag{20}$$

Menzerath-Altmann/Inverse-Gamma :

$$f = C\frac{e^{-b/r}}{r^a} \tag{21}$$

Frappat :

$$f = C + br + ce^{-ar} \tag{22}$$

The Menzerath-Altmann/Inverse-Gamma equation (21) is same as of equation (18) for resetting the parameters.

We ([43]) have pointed out the fact that for Hindi language corpora the law corresponding to the rank frequency distribution of words is the Zipf-Mandelbrot-law, given by equation (3) in present paper. We have further investigated that the proper model for the determination of frequencies with the help of ranks, in case if only two parameters are taken into account, is:

$$F_r = \frac{1}{p + qr} \tag{23}$$

for the appropriate choice of the parameters p and q.

Besides various rank frequency models and distributions diverse mathematical relations have also been applied and defined by researchers for governing frequencies and probabilities of different components of language(s). The next section 3 discusses several models and equations used for the formulation of the probabilities and frequencies of various components in different texts directly with the help of components' statistics without considering their ranks.

III. Mathematical Models with the Help of Components' Statistics

Pruscha [44] has given the relation between the length of a text and its vocabulary with the help of parametric non-linear regression method. In order to determine the relation between the length x and the vocabulary y of texts, the researcher has applied the residual sum of squares to compute the goodness of fit and has tested the fitting results of the models obtained as the solutions of the following differential equations:

$$y' = \frac{\alpha}{y^{\gamma}} \tag{24}$$

$$y' = \alpha e^{-\frac{y}{\gamma}} \tag{25}$$

$$y' = \alpha - \gamma y \tag{26}$$

$$y' = \alpha y - \gamma y^2 \tag{27}$$

The solutions of these differential equations shall have 'three-parameter-models' with a parameter c of integration and 'two-parameter-models' without integration parameter.

It has been shown by the researcher that the relationship $y = Ax^b$ with $b \sim 0.5$ executes well.

Manning and Schütze [15] in the chapter 'Topics in information retrieval' have emphasized that by characterization of a word for retrieval one must understand that a model for the distribution of a word be developed and then it should be used. They have discussed the applications of the Poisson distribution, two-poisson model and Katz's K mixture in the form of following equations:

Poisson distribution:

$$p(k; \lambda_i) = e^{-\lambda_i} \frac{\lambda_i^k}{k!} \tag{28}$$

parameter λ_i has been taken as the average number of occurrence of word w_i per document.

The two-Poisson model (equation (29)): According to the authors, the assumption of the model is that there are two classes of documents related with a term where one class is corresponding to the documents in which the term has high average number of occurrences (mentioned as the privileged class) and the another one is for the low average number of occurrences (non-privileged class).

$$p(k; \pi, \lambda_1, \lambda_2) = \pi e^{-\lambda_1} \frac{\lambda_1^k}{k!} + (1 - \pi) e^{-\lambda_2} \frac{\lambda_2^k}{k!} \tag{29}$$

Where λ_1 and λ_2 have been taken as the average number of occurrence of word in the privileged and non-privileged classes, respectively. π was defined by them as the probability of a document being in the privileged class and similarly $(1 - \pi)$, the probability of a document being in the non-privileged class.

The K mixture:

$$p_i(k) = (1 - \alpha)\delta_{k,0} + \frac{\alpha}{\beta + 1} \left(\frac{\beta}{\beta + 1} \right)^k \tag{30}$$

where $\begin{cases} \delta_{\mathrm{k},0} = 1 & \text{iff} \quad \mathrm{k} = 0 \\ \delta_{\mathrm{k},0} = 0, & \text{otherwise} \end{cases}$ and the parameters α and β can be determined by observed mean and with the help of document frequency (number of documents in the collection of documents in which the word w_i occurs).

In above three expressions, $p(k)$ is the proportion of times that a particular word w_i appears k times in a document. By using the Kolmogorov criterion, Guilpin and Guilpin [45], have verified the hypothesis of a uniform law for the occurrence of a particular word at different positions in a text for the Greek text. Aoyama and Constable [46] have showed that the relation between the mean number of syllables per word and the number of sequences of words totaling a given number of syllable $(n, n = 1, 2, 3 \ldots)$is dependent on the geometric frequency of word length totals. Sigurd et al [47] have studied data from English, Swedish and German to discover a theoretical distribution corresponding to word length and frequency. They have expressed the word length frequency distribution in the form

$$f_{\exp} = aL^b c^L \tag{31}$$

where a, b and c are parameters and f_{exp} , according to them, is the predicted word frequency on the basis of word length in letters (L). Lupsa and Lupsa[48] have generated the parameterized function illustrating the relation between the length of words and the absolute frequency of the words (i.e. the number of different words for each probable word length) in the form of following equation (32) by taking diverse basic word forms that are found as entries in the dictionary (or language vocabulary):

$$LV(x; c, k, \theta) = cx^k e^{-\frac{x}{\theta}} \tag{32}$$

where k, θ and c are parameters determined experimentally and in a language vocabulary $LV(x; c, k, \theta)$ is the law that depicts the absolute frequency of dictionary form of words with x length by selecting the experimental data for the Romanian and English languages. They have also concluded that the relation (mentioned in the form of equation (32)) also approximates the frequency of different word lengths in a corpus. Pande and Dhami [49], by using Zipf's order approach, have determined

that the parametric relation for the relative frequencies of words of different lengths in texts is of the form:

$$\begin{cases} f = A(1 + B\sqrt{l} + Ce^{-l})^b \\ f = 0,\ if\ 1 + B\sqrt{l} + Ce^{-l} < 0 \end{cases} \tag{33}$$

where f is frequency of words of length l. It has also been concluded that words of length 2 to length 6 are words of highest Zipf's orders in different texts and the frequencies of words of these lengths and their total frequency in different texts vary with the text length N according to the power law relation:

$$frequency = pN^q \tag{34}$$

where p and q are parameters and the parameter q takes value nearly equal to 1.

Another approach to study the length and frequency is depends on the consideration of the proportionality between two consecutive classes in the form $P_x = g(x)P_{x-1}$, where P_x and P_{x-1} are the probabilities of occurrence of elements with the values of element x and $x-1$ respectively. For example: Best [50] has determined the frequencies for the different lengths of rhythmic units and has investigated the 1-displaced Hyperpoisson distribution of the form:

$$P_x = \frac{a^{\mathrm{x}-1}}{b^{(x-1)}{}_1F_1(1;a;b)} \tag{35}$$

x =1, 2, 3. . . ...., for the relation between length of rhythmic units and corresponding frequencies in short prose of German language, where P_x is the theoretical probability of occurrence corresponding to the length of rhythmic unit x. The value of x has been selected by the researcher 1 if a stressed syllable follows another stresses syllable; $x = 2$ when one stressed syllable between two stresses and so on. The relevance of the hyper Poisson distribution for word length frequencies can be seen in the works of Best [51] for Old Icelandic songs and prose texts, Röttger [52] for Ciceronian letters, Dittrich [53] for German letters; Rottmann [54] for word length data from Old Church Slavonic and of Wilson [14] for present-day lower Sorbian newspaper texts.

In the case of modern Welsh prose texts; Wilson [55] has expressed the relation for the frequency corresponding to a particular word length as:

1-displaced Singh-Poisson distribution in the form

$$P_x = \begin{cases} 1 - \alpha + \alpha e^{-a}, & x = 1 \\ \frac{\alpha a^{x-1}e^{-a}}{(x-1)!}, & x = 2, 3, 4..... \end{cases} \tag{36}$$

where x is the number of syllables in the word and P_x the expected probability of words with x syllables in the text. Antić et al [56], for the word length frequency distribution in several Slovenian texts have also used the Singh-Poisson distribution. The Singh-Poisson distribution has previously also been utilized by Frischen [57] and Barbaro [58] for Jane Austen's letters and Italian Letters respectively. Different models applied for the word length can be found in Grzybek [59]. The author has presented the results of various distributions for the word length data of different languages, used by various researchers. Important distributions discussed by the author are given in the form of following equations (equation (37) to equation (44)):

1-displaced Conway-Maxwell-Poisson distribution:

$$P_x = \frac{a^{x-1}}{(x-1)!^b T_1} \tag{37}$$

$x = 1, 2, 3.....$ and $T_1 = \sum_{j=1}^{\infty} \frac{a^j}{(j!)^b}$

1-displaced Poisson distribution:

$$p_i = e^{-\lambda}\frac{\lambda^{i-1}}{(i-1)!} \tag{38}$$

i=1,2,3.....

(1-Displaced) Dacey-Poisson Distribution:

$$p_i = (1-\alpha)\frac{e^{-\lambda}\lambda^{i-1}}{(i-1)!} + \alpha\frac{e^{-\lambda}\lambda^{i-2}}{(i-2)!} \tag{39}$$

, i=1,2,3....

3-Parameter Fucks-Distribution:

$$\begin{cases} p_1 = e^{-\lambda}(1-\alpha) \\ p_2 = e^{-\lambda}\left[(1-\alpha)\lambda + (\alpha-\beta)\right] \\ p_i = e^{-\lambda}\left[(1-\alpha)\frac{\lambda^{i-1}}{(i-1)!} + (\alpha-\beta)\frac{\lambda^{i-2}}{(i-2)!} + \beta\frac{\lambda^{i-3}}{(i-3)!}\right] & i \geq 3 \end{cases} \tag{40}$$

$a-$ displaced form of negative binomial distribution:

$$f(x;k;p) = \begin{pmatrix} \mathrm{k+x}-a-1 \\ x-a \end{pmatrix} p^k q^{x-a} \tag{41}$$

where $q = 1-p$.

Buk and Rovenchak[60] have also used the one displace form of negative binomial distribution for the dependence of the number of sentences versus the number of constituting clauses.

1-displaced form of Poisson-uniform distribution:

$$\begin{aligned} P_x \quad = \quad & \frac{1}{\lambda_2-\lambda_1}\left(e^{-(\lambda_1-1)}\sum_{j=1}^{x}\frac{(\lambda_1-1)^{j-1}}{(j-1)!}\right. \\ & \left. -e^{-(\lambda_2-1)}\sum_{j=1}^{x}\frac{(\lambda_2-1)^{j-1}}{(j-1)!}\right) \end{aligned} \tag{42}$$

Kromer [61] has also obtained word length distribution based on Poisson-uniform distribution for German text by testing it in 10 newspaper texts.

Generalised Poisson distribution:

$$\begin{cases} P_0 = e^{-a} \\ P_x = \frac{a(a+bx)^{x-1}e^{-(a+bx)}}{x!}, & x = 1, 2, 3..... \end{cases} \tag{43}$$

and Hyper-Poisson distribution in one displaced form:

$$P_x = \frac{a^{x-1}}{{}_1F_1(1;b;a)b^{(x-1)}} \tag{44}$$

x=1,2..... where ${}_1F_1$ is the confluent hypergeometric function ${}_1F_1(1;b;a) = \sum_{j=0}^{\infty}\frac{a^j}{b^{(j)}}$ and

$b^{(0)} = 1$

$b^{(j)} = b(b+1)(b+2).....(b+j-1)$

Jayaram and Vidya [62] have investigated patterns of five Indian languages namely Assamese, Bengali, Hindi, Marathi Kannada and Tamil from the categories: 'Aesthetics', 'Commerce', 'Natural, Physical & Professional Sciences', 'Official

and Media Languages', 'Social Sciences'and 'Translated Material' with respect to word length distribution. They have determined that across categories there is a single distribution-which fits in majority of samples: in the case of Hindi 1 displaced extended positive binomial, in Kannada positive Cohen Poisson, and in Marathi and Tamil Dacey-negative binomial and for Assamese language, according to them Dacey-Poisson fits samples from two categories, Decay-negative binomial fits for two other categories and Extended positive binomial fits samples from the categories of official and media languages. The formulae used by them for the extended positive binomial, Cohen-Poisson, and Dacey negative binomial distribution are respectively as mentioned in following equations:

$$P_x = \begin{cases} 1-\alpha & x=1 \\ \alpha \begin{pmatrix} n \\ x-1 \end{pmatrix} \frac{p^{x-1}q^{n-x+1}}{1-q^n}, & x=2,3.....,n+1 \end{cases} \tag{45}$$

$$P_x = \begin{cases} \frac{(1-\alpha)a}{e^a-1-\alpha a} & x=1 \\ \frac{a^x}{x!(e^a-1-\alpha a)}, & x=2,3,4..... \end{cases} \tag{46}$$

$$P_x = (1-\alpha)\begin{pmatrix} k+x-2 \\ x-1 \end{pmatrix} p^k q^{x-1} + \alpha \begin{pmatrix} k+x-3 \\ x-2 \end{pmatrix} p^k q^{x-2} \tag{47}$$

, x=1,2,.. . .

The applications of Cohen-Poisson distribution for word length have also been pointed out in the works of Pawl-Rottmann [63] for East Slavonic and Abbe [64] for Arabic Letters and the relevance of extended positive binomial distribution for Czech letters has been established in Uhlírová [65].

Ishida and Ishida [66] have applied the Hyperpascal distribution of the form:

$$P_x = \frac{\begin{pmatrix} k+x-1 \\ x \end{pmatrix}}{\begin{pmatrix} m+x-1 \\ x \end{pmatrix}} q^x P_0 \tag{48}$$

x = 0, 1, 2..... where $P_0^{-1} = {}_2F_1(k,1;m,q), \quad F(.)$ is hypergeometric function and $P_x = \frac{k+x-1}{m+x-1} q P_{x-1}$, to the data of Japanese sentence length measured in terms of morphemes for the whole texts, and for the dialogue and narrative parts of the texts. The formula for Dacey-Poisson distribution is of the type of equation (39), discussed earlier in this paper.

For Czech language, Uhlirová [67] has tested the extended positive binomial distribution for the distribution of the $(n+1)^{th}$ word, given the length of n^{th} word form for the one-syllable (syllable is generally considered as a unit of pronunciation voiced without interruption), 2-syllable, 3-syllable and 4-syllable word forms which according to the researcher are the four most frequent word classes. Rottmann [68] has investigated syllable lengths in Russian, Bulgarian, Old Church Slavonic and Slovene and have founded that the adequate models are: Hyper-Poisson in Old Church Slavonic, modern Bulgarian and Slovene; and Conway-Maxwell-Poisson or Morse in modern Russian. Tamaoka and Altmann [69] have worked for the Kanji Strokes in Japanese. They have considered the Kanji as the smallest unit of meaning, the 'morpheme' and have used the 1-displaced negative hypergeometric distribution for the mathematical modeling for kanji strokes. They have also modeled the relationship between kanji strokes and kanji frequency in the form of the equation $y = ax^b e^{-cx}$ where a, b and c are paramerters , y is the frequency corresponding to the stroke number x. Pande and Dhami [70] have determined the distributions of various parts-of-speech for their occurrence in different parts of texts and occurrence in various texts. For the distributions, it has been have determined that: in case of the parts of speech- 'noun', 'pronoun', 'adjective', 'adverb', 'verb', 'to', 'determiner', 'preposition/conjunction', 'model', Binomial or Poisson distributions are more suitable and if parts of speech 'cardinal numbers', 'particle', 'interjection', 'pre determiners' and 'existential there' are considered, distributions shall be two-poisson or K mixture.

Köhler [23] has cited that "another kind of power law model frequently investigated in linguistics is the Menzerath-Altmann law". Menzerath-Altmann law (MAL) is a law that states the length of a linguistic component as a function of the length of the construct which it constitutes. Menzerath's law on the word-morpheme level has been specified in the work of Krott [71]. As mentioned in Grzybek et al [72], the most general form of MAL (cf. Altmann [73]:1) is as follows:

"The longer a language construct the shorter its components (constituents)".

They have also mentioned that "In quantitative and synergetic linguistics, the relations between linguistic units of different levels usually are treated in the framework of the Menzerath-Altmann law". They further cited that Altmann [73] proposed the formula to be the most general form:

$$y = Ax^b e^{-cx} \tag{49}$$

with two special cases: for $c = 0$ equation

$$y = Ax^b \tag{50}$$

and for b=0,

$$y = Ae^{-cx} \tag{51}$$

Here, in above three equations, y corresponds to the constituent's size (for example syllable length) and 'x ' the sizes of the linguistic construct that is examined (for example number of syllable per word or word length in terms of number of syllables). By applying the relation of the form of equation (50) and the linear relation, for texts of different genres, for sentence length (SL) and word length (WL) of Russian texts they (Grzybek et al , [72]) have concluded that "there is only a weak relation between the means of WL and SL". Köhler [23] has also discussed that the Menzerath-Altmann law (MAL) is characterized by the formula $y = Ax^B e^{Cx}$ for the length y of the constituent and the length x of the corresponding unit calculated in number of their constituents. In case of Hungarian, the researcher has established the MAL for the syllable length as a function of word length measured in terms of syllables, by using dictionary study.

Grzybek et al [74] cited that Wimmer and Altmann ([75], [76]) have extended the approach of Menzerath-Altmann law by using the differential equation of the form $\frac{dy}{y} =$

$\left(a_0 + \frac{a_1}{x} + \frac{a_2}{x^2} + \frac{a_3}{x^3} \ldots\ldots\right) dx$ and have given three new equations:

$$\begin{cases} y = ae^{d/x} \\ y = ax^{-b}e^{d/x} \\ y = ax^{-b}e^{cx}e^{d/x} \end{cases} \quad (52)$$

They (Grzybek et al [74]) have tested the sentence length word length relation among Altmann-Menzerathian line by applying the relation of the type of equation (50) for testing separately the Menzerahian tendency for short sentences, for long sentences, for minimal frequency and textual heterogeneity. Buk and Rovenchak [59] have tested the syntactic structure in Ukrainian. For Ukrainian the researchers have determined the number of clauses for each sentence by the formula: $number\ of\ clauses = \max(N_1 + N_2 + N_3 + N_4; N_C + 1)$ where N_1 , according to them represents the number of 'all verbal forms except an infinitive'; similarly N_2 is the number of 'participles preceded by a comma'; N_3 number of 'predicative words'; N_4 the number of dashes 'standing for the missing verb in compound predicates', and N_C represents the number of 'conjunctions preceded by a comma'. They have analyzed the novel by 'Ivan Franko' and have determined the dependence of clause length on the sentence length (measured in terms of number of clauses) in the form of an equation of the kind of second equation in (52) above by setting $y = f(x) = clause\ length$ in this case.

IV. Conclusion

The function of statistical means to natural language processing has been exceptionally successful. Statistical tactics employ various mathematical techniques and are frequently used in large text corpora to make approximate generalized models of linguistic facts based on concrete examples of these phenomena provided by the text corpora without adding major linguistic or world knowledge. The broad availability of linguistic resources as well as of speech and text corpora has assisted a vital role in their success. For all learning techniques, these techniques and methods generally rely on data. In the present paper a study of various models and distributions, which are applied for the formulation of linguistic components of different languages, has been presented. These have been used by researchers at different levels of linguistic elements for many languages. This presentation shall be helpful in understanding the treatment of the language from quantitative point of view and paves the way of further investigations to determine the pattern of occurrence of various unanswered components for different languages.

Appendix
Example text

The first paragraph of the ebook 'THE CRUISE OF THE DOLPHIN' of author 'Thomas Bailey Aldrich'

"Every Rivermouth .....of waters."

Available at: http://www.gutenberg.org/files/1757/1757-h/1757-h.htm

References

[1] Z. S. Harris, *Methods in Structural Linguistics.*, Chicago: University of Chicago Press, 1951.
[2] N. Chomsky, *Syntactic Structures.*, The Hague: Mouton, 1957.
[3] J. Lambek, "The mathematics of sentence structure", *Americal Mathematical Monthly*, vol. 65, pp. 154-170, 1958.
[4] J. Lambek, "Contribution to a mathematical analysis of the English verb-phrase", *Journal Canadian Linguistic Association* , vol. 5, pp. 83-89, 1959.
[5] J. Lambek, "On the calculus of syntactic types", *In R. Jakobson (Ed.). Structure of Language and Its Mathematical Aspects*, pp. 166-178,1961.
[6] J. Lambek, "On a connection between algebra, logic and linguistics", *Diagrammes*, vol. 22, pp. 59-75, 1989.
[7] Z. S. Harris, *Mathematical Structure of Language. New York. John Wiley and Sons*, 1968.
[8] R. Jakobson, (Ed.), *Structure of Language and Its Mathematical Aspects. Providence: AMS*, 1961.
[9] I. A. Bolshakov and A. Gelbukh, *Computational Linguistics Models, Resources, Applications. Ciencia De La Computacin, Mexico.*, 2004.
[10] A. Kornai, *Mathematical Linguistics. Springer*, 2008.
[11] G. Altmann, "On the symbiosis of physicists and linguists", *Romanian Report in Physics*, vol. 60(3), pp. 417-422, 2008.
[12] P. Meyer, "Laws and Theories in Quantitative Linguistics", *Glottometrics*, vol. 5, pp. 62-80, 2002.
[13] R. Köhler and G. Altmann, "Aims and Methods of Quantitative Linguistics", *In G. Altmann, V. Levickij and V. Perebyinis (Eds.) Problems of Quantitative Linguistics*, pp. 12-41, 2005.Available at http://www.ram-verlag.de/lexico5.pdf
[14] A. Wilson, "Word-Length Distribution in Present-Day Lower Sorbian Newspaper Texts", *In P. Grzybek(Ed.) Contributions To The Science Of Text And Language Word Length Studies and Related Issues, Springer, Netherlands*, pp. 319-327, 2006.
[15] C. D. Manning, and H. Schütze, *Foundations of Statistical Natural Language Processing, MIT Press, Cambridge*, 1999.
[16] C. Butler, *Statistics in Linguistics. Oxford, Basil Blackwell*, 1985.
[17] G. Altmann, "Zipfian linguistics", *Glottometrics* , vol. 3, pp. 19-26,2002.
[18] G. K. Zipf, *Human Behaviour and the Principle of Least Effort , Reading, MA: Addison-Wesley Press*, 1949.
[19] 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*, pp. 151-168, 1966.
[20] N. Hatzigeorgiu , George Mikros and George Carayannis, "Word Length, Word Frequencies and Zipf's Law in the Greek Language", *Journal of Quantitative Linguistics* , vol. 8:3, pp. 175-185, 2001.
[21] L. Q. Ha, E. I. Sicilia-Garcia , J. Ming and F. J. Smith, "Extension of Zipf's Law to Words and Phrases", *In Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, Available at: http://aclweb.org/anthology/C/C02/C02-1117.pdf, 2002.
[22] G. Mikros, N. Hatzigeorgiu and G. Carayannis, "Basic Quantitative Characteristics of the Modern Greek Language Using the Hellenic National Corpus", *Journal of Quantitative Linguistics*, vol. 12:2-3, pp. 167-184, 2005.
[23] R. Köhler, "Power law models in linguistics: Hungarian", *Glottometrics*, vol. 5, pp. 51-61, 2002.
[24] G. Dalkiliç and Y. Çebi, "Zipf's Law and Mandelbrot's Constants for Turkish Language Using Turkish Corpus (TurCo)", *In T. Yakhno(Ed.) Advances In Information Systems. Lecture Notes in Computer Science*, vol. 3261, pp. 273-282,Springer, 2004.
[25] A. Tuzzi, I. I. Popescu and G. Altman, "Zipf's laws in Italian texts", *Journal of Quantitative Linguistics*, vol. 16(4), pp. 354-367, 2009.
[26] H. Situngkir, "An Observational Framework to the Zipfian Analysis among Different Languages: Studies to Indonesian Ethnic Biblical Texts", *Available at: http://cogprints.org/5481/1/2007a.pdf*, 2007.
[27] B. D. Jayaram and M. N. Vidya, "Zipf's law for Indian languages", *Journal of quantitative linguistics, 15(4), 293-317*, vol. 15(4), pp. 293-317, 2008.
[28] B. Sigurd, "Rank frequency distribution of phonemes", *Phonetica*, vol. 18, pp. 1-15, 1968.

[29] I. J. Good, "Statistics of language", *In A. R. Meetham and R. A. Hudson, (Eds.) Encyclopaedia of linguistics, information and control . Oxford: Pergamon*, pp. 567-581, 1969.
[30] I. I. Popescu, G. Altmann, P. Grzybek, B. D. Jayaram, R. Köhler, V. Krupa, J. Macutek, R. Pustet, L. Uhlirová, and M. N. Vidya, *Word Frequency Studies. Mouton De Gruyter.*, 2009.
[31] S. M. Gusein-Zade, "Frequency Distribution of Letters in the Russian Language", *Problems of Information Transmission* , vol. 24(4), pp. 338-342, 1988.
[32] M., Y. Borodovsky, and S. M. Gusein-Zade, " A general rule for ranged series of codon frequencies in different genomes", *Journal of Biomolecular Structure and Dynamics*, vol. 6, pp. 1001-1013, 1989.
[33] L.Q. Ha , E. I. Sicilia-Garcia , J. Ming and F. J. Smith, "Extension of Zipf's Law to Word and Character N-grams for English and Chinese", *International Journal of Computational Linguistics & Chinese Language Processing* , vol. 8(1), February 2003, pp. 77-102, 2003 .
[34] P. Grzybek and E. Kelih, "Towards a general model of grapheme frequencies in Slavic languages", *In: Garabík, R. (Ed.), Computer Treatment of Slavic and East European Languages . Bratislava: Veda*, pp. 73-87, 2005.
[35] A. Eftekhari, "Fractal geometry of texts: An initial application to the works of Shakespeare", *Journal of Quantitative Linguistics*, vol. 13(2-3), pp. 177-193, 2006.
[36] Y. Tambovtsev and C. Martindale, "Phoneme frequencies follow a Yule Distribution", *SKASE Journal of Theoretical Linguistics*, vol. 4(2), pp. 1-11, 2007.
[37] C. Martindale, S. M. Gusein-Zade, D. Mekenzie and M. Y. Borodovsky, "Comparison of equations describing the ranked frequency distributions of graphemes and phonemes", *Journal of Quantitative Linguistics*, vol. 3(2), pp. 106-112, 1996.
[38] J. Macutek, "A generalization of the geometric distribution and its application in quantitative linguistics", *Romanian Reports in Physics*, vol. 60(3), pp. 501-509, 2008.
[39] H. Pande and H. S. Dhami, "Generation of a model for grapheme frequencies and its refinement and validation by group theoretic aspects", *Journal of Quantitative Linguistics* , vol. 16(4), pp. 307-326, 2009.
[40] H. Pande and H. S. Dhami, "Mathematical Modelling of Occurrence of Letters and Word's Initials in Texts of Hindi Language", *SKASE Journal of Theoretical Linguistics*, vol. 7(2), pp. 19-38, 2010.
[41] W. Li, P. Miramontes and G. Cocho, "Fitting Ranked Linguistic Data with Two-Parameter Functions", *Entropy* , vol. 12(7), pp. 1743-1764, 2010.
[42] W. Li, and P. Miramontes, "Fitting Ranked English and Spanish Letter Frequency Distribution in U.S. and Mexican Presidential Speeches", *Journal of Quantitative Linguistics* , vol. 18(4), pp. 359-380, 2011.
[43] H. Pande and H. S. Dhami, "Mathematical modeling of the pattern of occurrence of words in different corpora of Hindi language", *Journal of Quantitative Linguistics (to appear)*, vol. 20(1), 2013.
[44] H. Pruscha, "Statistical models for vocabulary and text length with an application to the NT corpus.", *Literary and Linguistic Computing*, vol. 13(4), pp. 195-198, 1998.
[45] P. Guilpin and C. Guilpin, "Linguistic and statistical analysis of the frequency of a particular word at different times (diachrony) or in different styles (synchrony).", *Journal of Quantitative Linguistics*, vol. 12(2-3), pp. 138-150, 2005.
[46] H. Aoyama and J. Constable, "Word length frequency and distribution in English: Part I. Prose", *Literary and Linguistic Computing* , vol. 14(3), pp. 339-358, 1999.
[47] B. Sigurd, M. Eeg-Olofsson, and J. van Weijer, "Word length, sentence length and frequency- Zipf revisited.", *Studia Linguistica* , vol. 58(1), pp. 37-52, 2004.
[48] D. A. Lupsa and R. Lupsa, "The Law of Word Length in a Vocabulary", *Studia Univ. Babes-Bolyai, Informatica* , vol. L, Number. 2, pp. 69-80, 2005.
[49] H. Pande and H. S. Dhami, "Model generation for word length frequencies in texts with the application of Zipf's order approach", *Journal of Quantitative Linguistics*, vol. 19(4), pp. 249-261, 2012.
[50] K. H. Best, "The distribution of rhythmic units in German short prose", *Glottometrics*, vol. 3, pp. 136-142, 2002.
[51] K. H. Best, "Word length in Old Icelandic songs and prose texts ", *Journal of Quantitative Linguistics*, vol. 3:2, pp. 97-105, 1996.
[52] W. Röttger, "Distribution of word length in Ciceronian letters ", *Journal of Quantitative Linguistics*, vol. 3:1, pp.68-72, 1996.
[53] H. Dittrich, "Word length frequency in the letters of G.E. Lessing", *Journal of Quantitative Linguistics*, vol. 3:3, pp. 260-264, 1996.
[54] O. A. Rottmann, "WordLength counting in Old Church Slavonic", *Journal of Quantitative Linguistics* , vol. 4:1-3, pp. 252-256, 1997.
[55] A. Wilson, "Word length distributions in modern Welsh prose texts", *Glottometrics* , vol. 6, pp. 35-39, 2003.
[56] G. Antić, E. Stadlober, P. Grzybek and E. Kelih, "Word Length and Frequency Distributions in Different Text Genres", *From Data and Information Analysis to Knowledge Engineering, Springer, Berlin Heidelberg*, pp.310-317, 2006.
[57] J. Frischen, "Word length analysis of Jane Austen's letters", *Journal of Quantitative Linguistics*, vol. 3:1, pp. 80-84, 1996.
[58] S. Barbaro, "Word Length Distribution in Italian Letters by Pier Paolo Pasolini", *Journal of Quantitative Linguistics*, vol. 7:2, pp. 115-120, 2000.
[59] P. Grzybek, "History and methodology of word length studies", *P. Grzybek (Ed.): Contributions To The Science Of Text And Language. Dordrecht: Springer*, pp.15-90, 2006.
[60] S. Buk and A. A. Rovenchak, "Menzerath-Altmann law for syntactic structures in Ukrainian", *Glottotheory*, vol. 1, No. 1, pp. 10-17, 2008.
[61] V. Kromer, "Word length model based on one-displaced Poisson-uniform distribution", *Glottometrics*, vol. 1, pp. 87-96, 2001.
[62] B. D. Jayaram and M. N. Vidya, "Word length distribution in Indian languages", *Glottometrics*, vol. 12, pp. 16-38, 2006.
[63] O. A. PawlRottmann, "Word and Syllable Lengths in East Slavonic", *Journal of Quantitative Linguistics* , vol. 6:3, pp. 235-238, 1999.
[64] S. Abbe, "Word Length Distribution in Arabic Letters", *Journal of Quantitative Linguistics*, vol. 7:2, pp. 121-127, 2000.
[65] L. Uhlírová, "Word Length Modelling: Intertextuality as a Relevant Factor?", *Journal of Quantitative Linguistics*, vol. 6:3, pp. 252-256, 1999.
[66] M. Ishida and K. Ishida, "On distribution of sentence lengths in Japanese writings", *Glottometrics*, vol. 15, pp. 28-44, 2007.
[67] L. Uhlirová, "On language modelling in automatic speech recognition", *Journal of Quantitative Linguistics*, vol. 7(3), pp. 209-216, 2000.
[68] O. Rottmann, "Syllable lengths in Russian, Bulgarian, Old Church Slavonic and Slovene", *Glottometrics*, vol. 2, pp. 87-94, 2002.
[69] K. Tamaoka and G. Altmann, "Mathematical Modelling for Japanese Kanji Strokes in Relation to Frequency, Asymmetry and Readings", *Glottometrics*, vol. 10, pp. 16-29, 2005.
[70] H. Pande and H. S. Dhami, "Distributions of different parts of speech in different parts of a text and in different texts", *The Modern Journal of Applied Linguistics*, vol. 2:1, pp. 152-170, 2010.
[71] A. Krott, "Some remarks on the relation between word length and morpheme length", *Journal of Quantitative Linguistics*, vol. 3:1, pp.29-37, 1996.
[72] P. Grzybek, E. Stadlober, E. Kelih, "The relationship of word length and sentence length: the inter-textual perspective", *In: R. Decker and H.-J. Lenz (Eds.): Advances in Data Analysis, Berlin: Springer*, pp. 611-618, 2007.
[73] G. Altmann, "Prolegomena to Menzerath's Law", *Glottometrika*, vol. 2, pp. 1-10, 1980.
[74] P. Grzybek, E. Kelih and E. Stadlober, "The relation between word length and sentence length: an intra - systemic perspective in the core data structure", *Glottometrics*, vol. 16, pp. 111-121, 2008.
[75] G. Wimmer and G. Altmann, "Unified derivation of some linguistic laws", *In Köhler, R., Altmann, G., Piotrowski, R. (Eds.), Quantitative Linguistik-Quantitative Linguistics. in Internationales Handbuch-An International Handbook, Berlin/New York:de Gruyter*, pp. 791-807, 2005.
[76] G. Wimmer and G. Altmann, "Towards unified derivation of some linguistic laws", *In: Grzybek, P. (Ed.). Contributions to the Science of Text and Language: Word Length Studies and Related Issues, Dordrecht: Springer, Netherlands*, pp. 329-337, 2006.