



A Novel Approach To Characterize Quantitatively Themes Of Literary Text- A Pilot Study With Shanon Entropy

¹Rini Bhattacharyya, ²Dipak Ghosh, ³Aritra Bhattacharyya

¹Assistant Professor, ²Emeritus Professor, ³Software Engineer

¹Department of Physics,

¹Techno Main Salt Lake, Kolkata, India

Abstract: In this paper we present a novel approach to characterize quantitatively the theme of different literary texts. For this we choose the parameter namely Shanon Entropy and apply it to two sets of literary texts. First one is a defined theme of suspense thriller authored by Agatha Christie and second set consists of texts of different themes of celebrated authors. The analysis clearly indicates the first set of the same theme possesses more or less the same Shanon entropy where the second set entropy value varies from book to book. This quantitative study reveals that 'Shanon Entropy' is a very useful and effective parameter for characterization and assessing the order of complexity in different literary texts.

Index Terms - Suspense story, Rhythm, Non Linear Analysis, Entropy

I. INTRODUCTION

Language is a structural system of words, sounds, gesture, patterns, symbols and others such as visual, aural, gestural and spatial by means of which human beings as member of a social group and participants in its culture, express themselves. The functions of language include communication, expression of identity, play imaginative expression and emotional release. Language grows with the biological and social evolution of human beings. Scientific studies of recent times focus that normal human brain can grasp language, a self-organized structured system easily and spontaneously[1] since childhood and language of a particular person reflects the inner orientation of his brain[2]. A language reflects how organized a society or a nation is[3-6]. Analysis of structural patterns of language of different literary works of celebrated authors of a particular language by using scientific technique gives us quantitative measure of their writing styles as well as idea about their development of linguistic abilities. And this type of research will provide us quantitative information required for understanding as regards the complexity of the functioning of the human brain.

A sophisticated dynamics is associated with each language[7]. It would be a fundamental mistake to consider a language as an amorphous mixture of symbols viz, (letters, phonemes, morphemes, words etc.). In fact, a language is a highly organised sequence in which particular symbols are ordained according to specific rules most of which are guided by the grammar of that particular language[8,9]. Therefore, language can be considered as complex system[10,11] with reference to the language of grammar. Some researchers have expressed words by nodes and their relations by edges for English, Portuguese and Chinese[12-16].

2.BACKGROUND

Attempts are made to discover structural patterns of an author embedded in his texts series produced during a considerable length of period [17-23] and the way in which a typical linguistic network is generated[24-26]. Study of Yang et al.[17] reveals the evolution of scaling behavior embedded in sentence applying Zif's law[27-28], according to which the occurrence frequency of the most frequent word in a text obeys a power law according to its rank, , if the words are ranked in a descending order of their occurring frequencies. In 1966 Mandelbrot modified Zipf's law [29] in order to fit more adequately the region corresponding to the lowest ranks by introducing fractal [30] structure of the lexical trees. In 2001, Marcelo pointed out certain limitations as regards Zipf-Mandelbrot law relating to prediction of anomalies associated with high-rank words[31]. In another study Kulig et al. [32] showed that in linguistic networks the punctuation marks obey the same laws as words.

Words divulge a man's thought, knowledge[33], emotions, feelings, desire, expectation etc. despite restricted capacity of brains to store lexicons. The concepts of successful communications inspires a man to establish complex, syntactic and semantic relations between limited set of stored words and symbols. Since human language is a carrier of highly complex information about thoughts, feelings, knowledge etc., it must come under competing requirement of allowing high information rate and remarkably appreciable communication error [34]. Statistical analysis of human language is required to characterize universality, specificity and evolution of cultures[35]. Recently it has been advocated to explore the analysis of literary text with the help of entropy.

3.OBJECTIVE:

In this paper we aim to characterize and compare literary text of distinctively different themes using Shannon Entropy. For that we take two sets of text, one set consisting of several books of Agatha Christie on her well known detective theme other some books of famous literary persons on different themes. We want to explore whether it is possible to characterize the linguistic pattern with the help of a parameter of information science namely Shannon Entropy.

Generally, entropy is defined as a measure of randomness or disorder of a system. The concept of entropy was introduced by Rudolf Clausius, a German Physicist in 1850. From a thermodynamic view point of entropy, microscopic details of a system are not considered, instead entropy is used to describe the behaviour of a system in terms of thermodynamic properties, such as temperature, pressure, and heat capacity. The thermodynamic description took into consideration, a state of equilibrium of systems. Meanwhile, a statistical definition which was developed at a later stage focussed on the thermodynamic properties which was defined in terms of the statistics of molecular motions of a systems.

At present there is a growing tendency in using this term of thermodynamics namely entropy in other scientific fields. The concepts of physical and informational entropy are being applied to linguistics, logic, financial, numeric, social and military systems[36,37,38].

The concept of information entropy was introduced by Claude Shannon in his work[39] in 1948. Such concept of information entropy is also referred to as Shannon entropy. He actually considered problems such as order and disorder, errors and errors' control, possibilities and actualization of possibilities, ambiguity and ambiguity limits. The Shannon's idea of entropy ultimately allowed the formulation of information entropy. In this context the Shannon entropy has been considered as an effective tool in different disciplines. Only few, attempts have been made to apply the concept of entropy in linguistics

Given a discrete random variable X , with possible outcomes $x_1, x_2, \dots, x_n$ which occurs with probability $P(x_1), P(x_2), \dots, P(x_n)$, the Shannon entropy $H(X)$ is defined as follows:

$$H(X) = - \sum_{i=1}^n P(x_i) \log P(x_i) \quad \dots \dots \dots (1)$$

$\sum$ denotes the sum over the variable's possible values. The choice of base for log varies for different applications.

4.DATA ANALYSIS:

Data were obtained from the original texts of Agatha Christie and other eminent authors. First, the texts are arranged in a chronological order. In the present work, we have analyzed the following books which were written from 1920 to 1966. We also include different other famous books written over a wide span of time (1829-2005):

The list of book is given below:

- | | |
|------------------------------------|---|
| 1. The Mysterious Affair at Styles | 1. David Copperfield |
| 2. The Mysterious Mr Quin | 2. War and Peace |
| 3. Parker Pyne Investigates | 3. Anna Karenina |
| 4. And Then There Were None | 4. Resurrection |
| 5. At Bertram's Hotel | 5. The Copernican Revolution |
| 6. Third Girl | 6. Game of Thrones |
| | 7. Ideas A History of Thought and Invention, from Fire to Freud |

In each book number of words in a sentence were counted and a frequency distribution of the number of words were obtained. Thus we get a probability distribution in respect to the number of words in a sentence which were used as a input for calculation of Shannon's entropy as given by the formula (1). The figures 1 to 6 show the probability distribution for six books of Agatha Christie and figure 7-11 show similar distribution for other books.

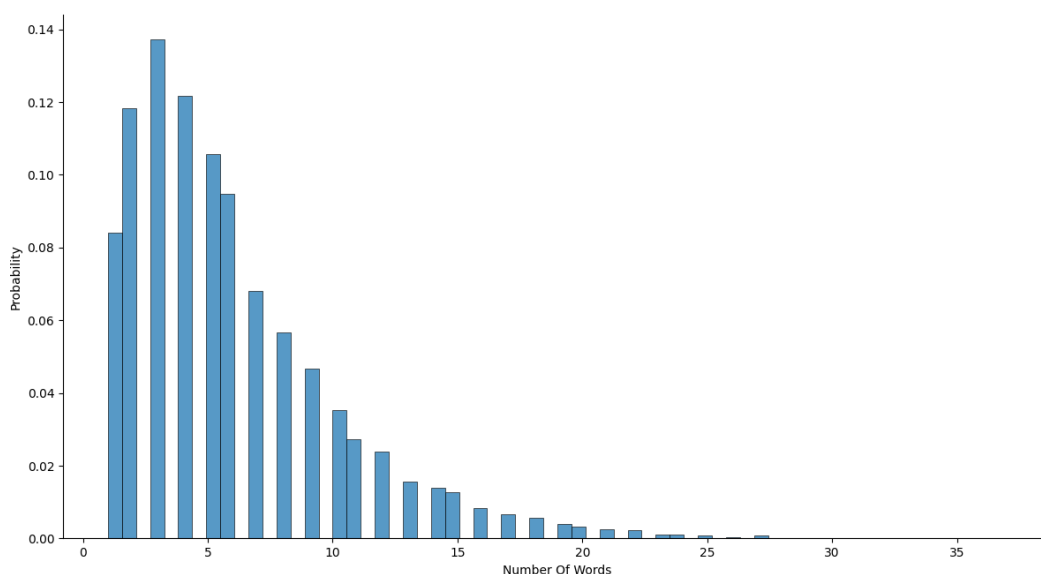


Fig1:Probability distribution for the book The Mysterious Affairs at Styles

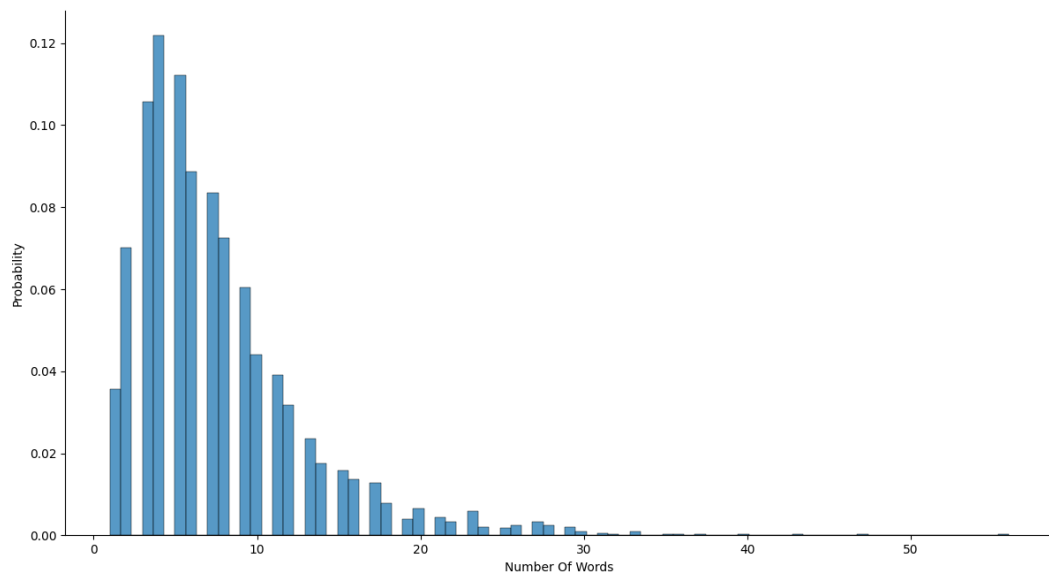


Fig 2: Probability distribution of the book The Mysterious Mr Quin

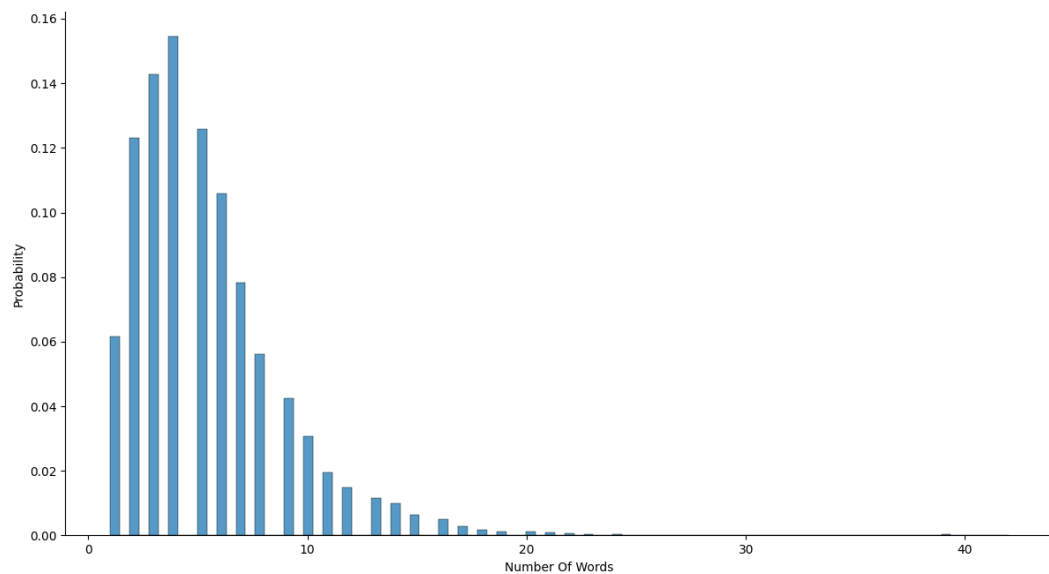


Fig3: Probability distribution of the book Parker Pyne Investigates

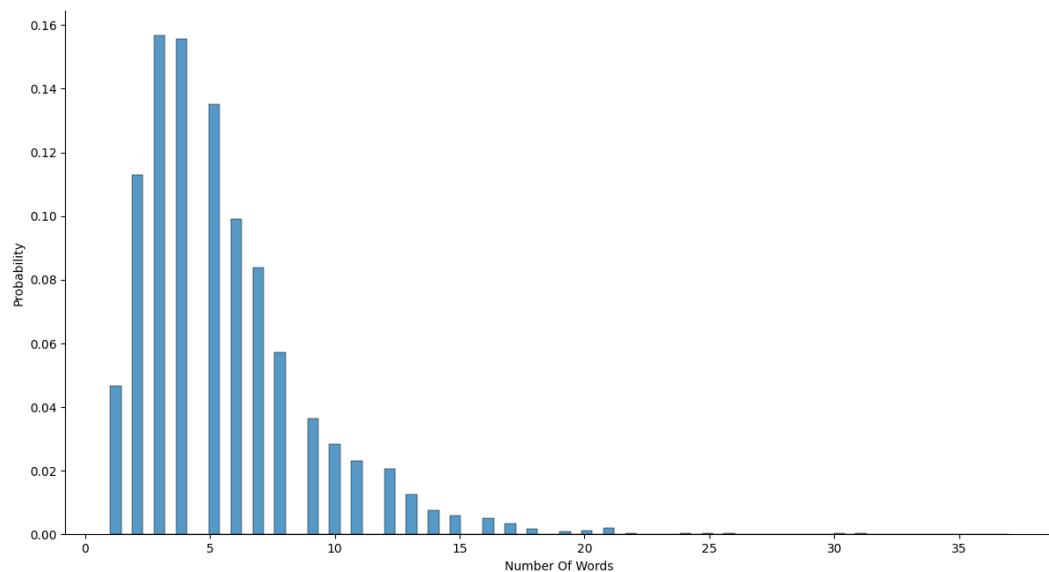


Fig 4: Probability distribution of the book And Then There Were None

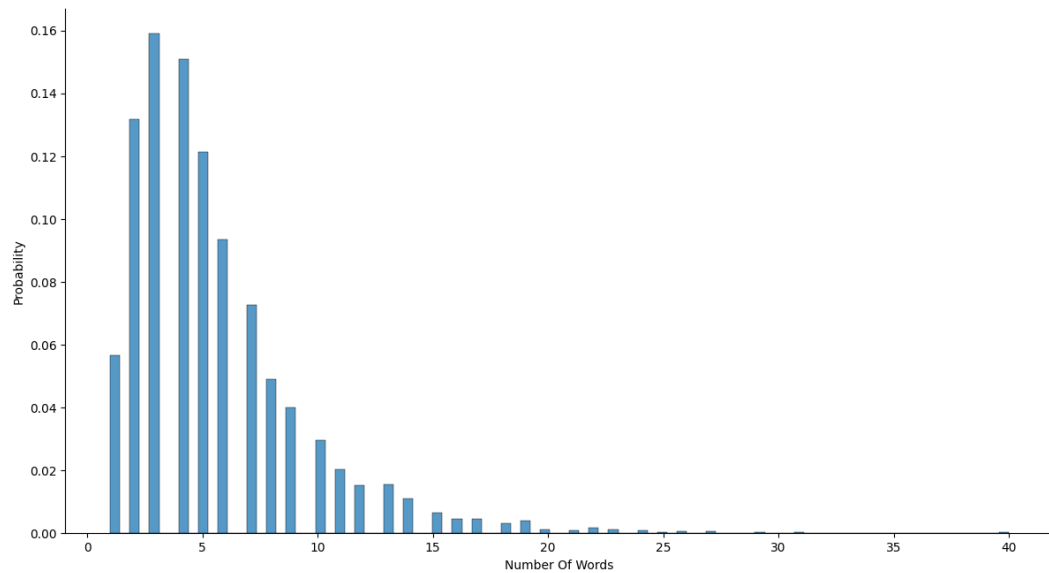


Fig 5: Probability distribution of the book At Bertram's Hotel

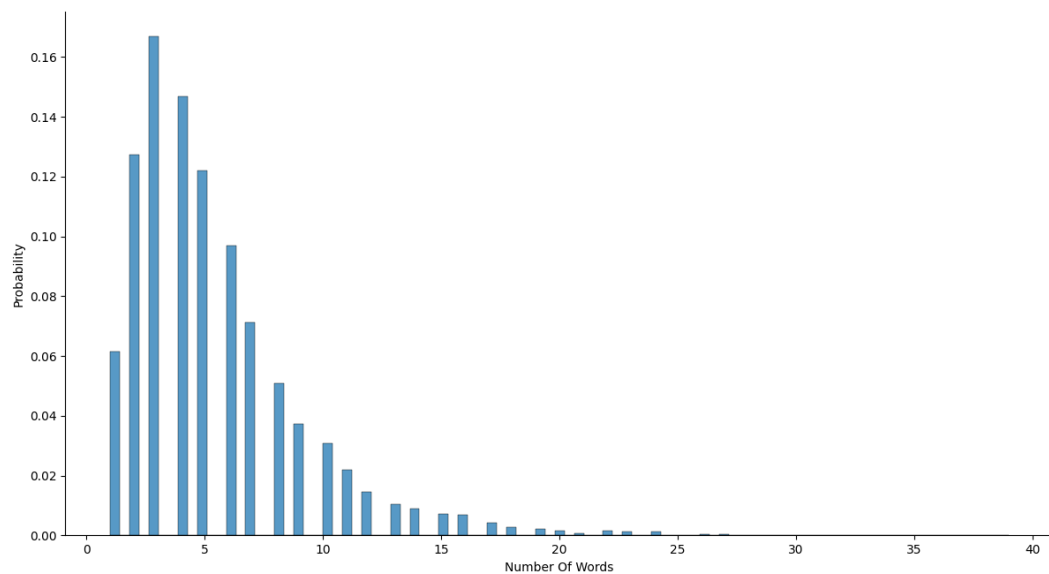


Fig 6: Probability distribution of the book Third Girl

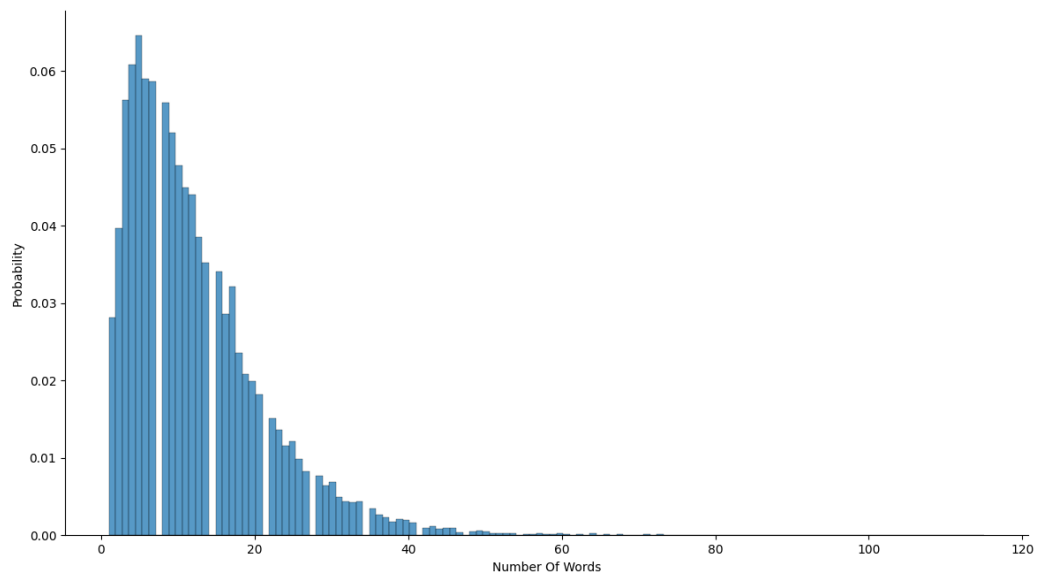


Fig7: Probability distribution of the book David Copperfield

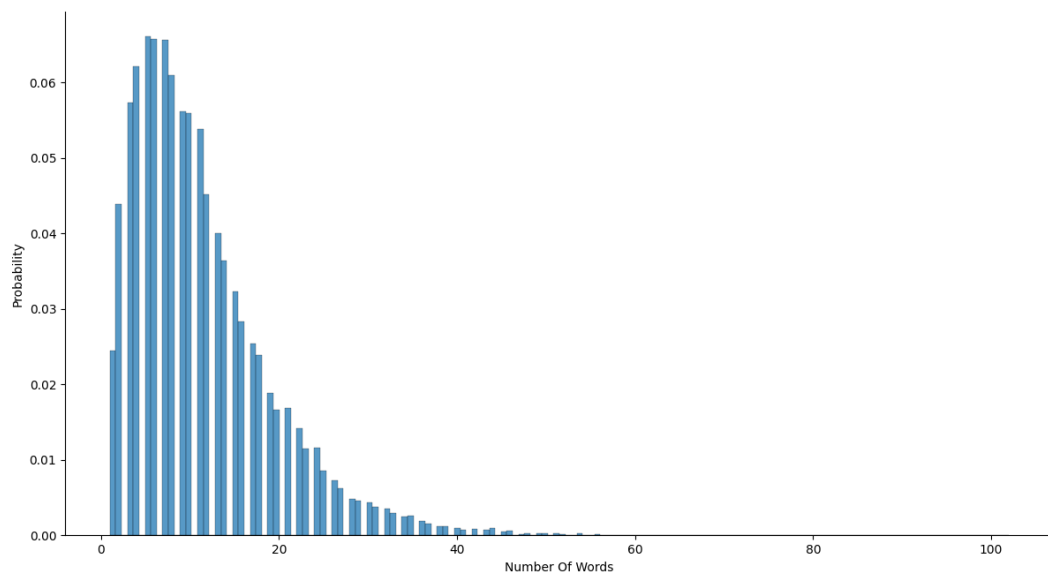


Fig 8: Probability distribution of the book War and Peace

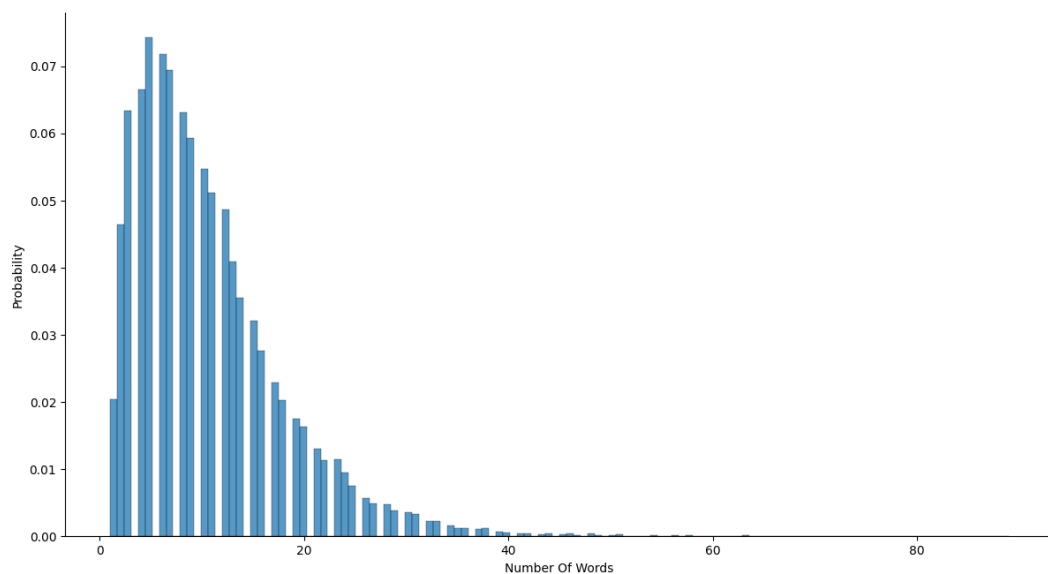


Fig 9: Probability distribution of the book Anna Karenina

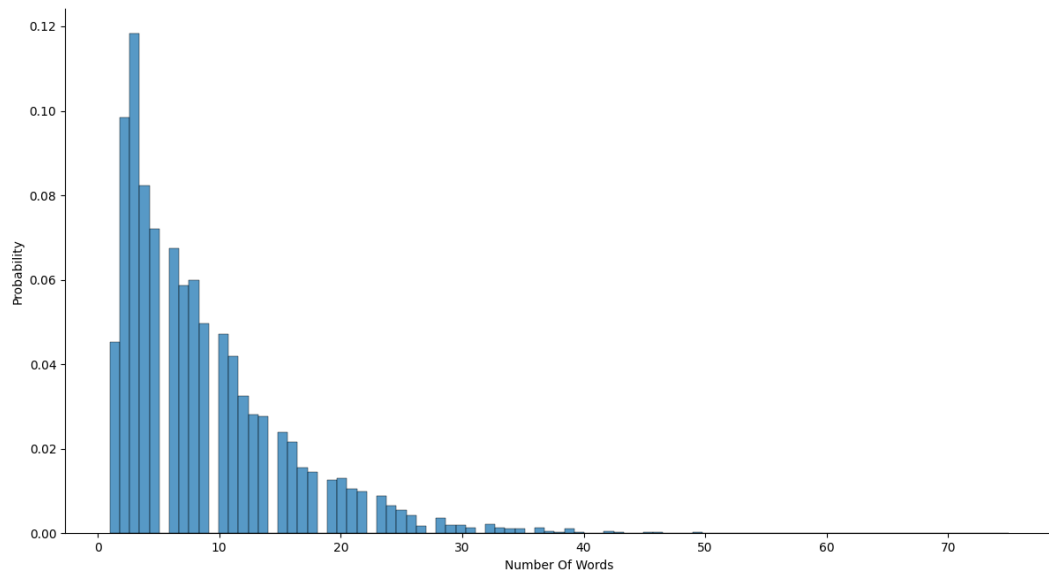


Fig 10: Probability distribution of the book Resurrection

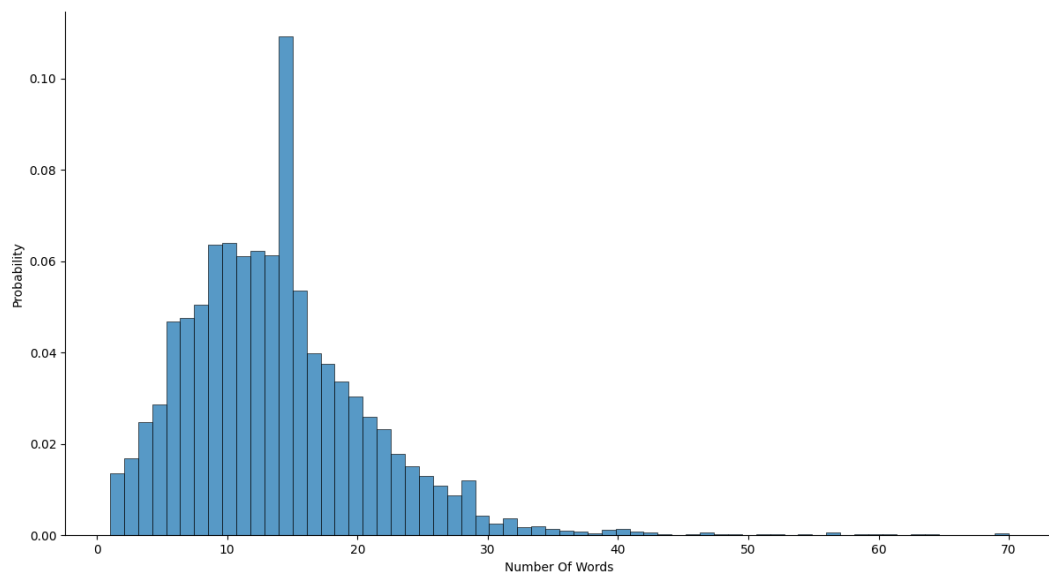


Fig 11: Probability distribution of the book Copernicus Revolution

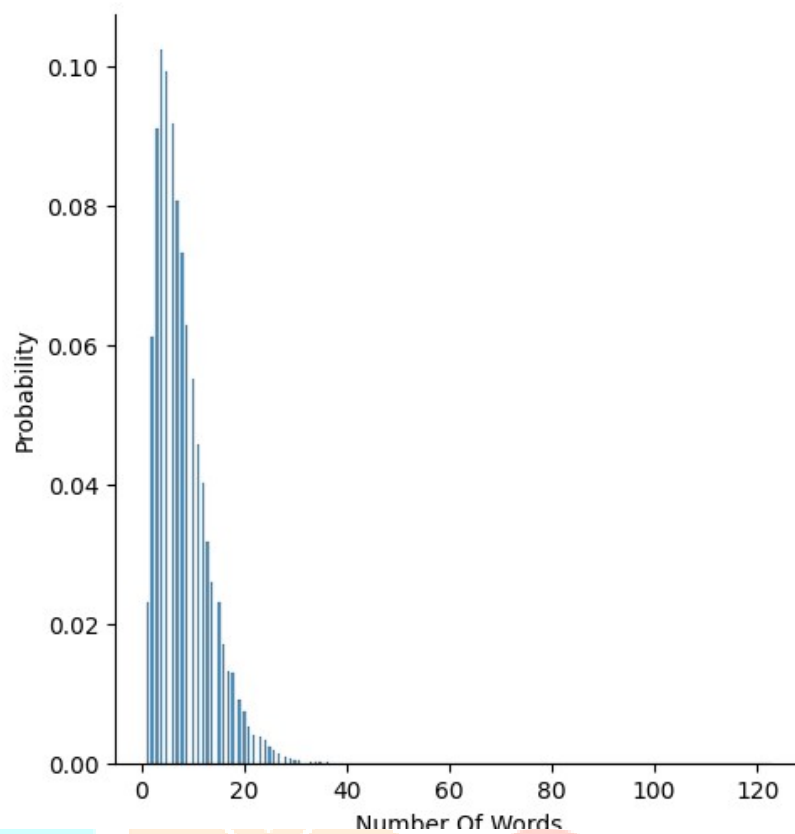


Fig 12: Probability distribution of the book Game of Thrones

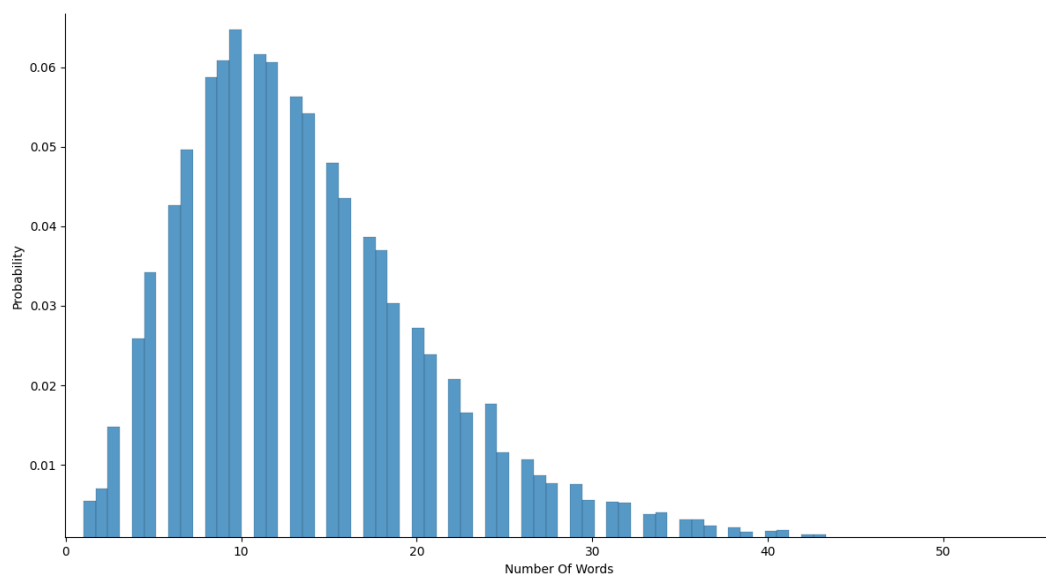


Fig 13: Probability distribution of the book Ideas A History of Thought and Invention, from Fire to Freud

Next we calculated Shanon Entropy using formula 1 for two sets of books. Results are shown in following tables- the table 1 shows the results of Augatha Christi, table 2 for other books.

TABLE 1:

Name of Book	Year of Publish	Entropy
1. The Mysterious Affair at Styles by Agatha Christie	1920	1.08
2.The Mysterious Mr Quin	1930	1.21
3.Parker Pyne Investigates	1934	1.15
4. And Then There Were None	1939	1.06
5.At Bertram's Hotel	1965	1.08
6. Third Girl	1966	1.02

TABLE2:

Name of Book	Year of Publish	Entropy value
David Copperfield by Charles Dickens	1849	1.61
War and Peace By Leo Tolstoy	1869	1.28
Anna Karenina By Leo Tolstoy	1878	1.41
Resurrection By Leo Tolstoy	1899	1.27
The Copernican Revolution by Thomas S Khun	1992	1.76
Game of Thrones by <u>David Benioff, George R. R. Martin</u>	1996	1.19
Ideas A History of Thought and Invention, from Fire to Freud by Peter Watson	2005	1.29

5.CONCLUSION:

It is highly interesting to note that entropy values of the books of Agatha Christie written over a period of 40 years are very close to each other(1.02 to 1.21) [table 1].On the other hand, entropy values of other books written on different theme varies from 1.19 to 1.76 (shown in table2).

Needless to say all books of Agatha Christie deal with detective thrillers and this is reflected in entropy values revealing more or less same entropy. Thus , the paper provides new data on the use of information science more precisely, Shanon entropy, gives a good quantitative indication to identify different genres of writing in a fruitful way. In other words, Shanon entropy is indicative of different intellectual skills of the same author. The higher entropy values also indicate a higher degree of unpredictability and complexity of events. This work enlarges the idea of further exploration of literary texts from different perspectives by using Shanon entropy as an effective parameter.

6. REFERENCES:

- [1] M.H. Christiansen, N. Chater, Language as shaped by the brain, *Behav Brain Sci.* 31 (2008) 489
- [2] I. Grabska-Gradzińska, A. Kulig, J. Kwapień, P. Oświęcimka, S. Drozd, Multifractal analysis of sentence lengths in english literary texts, *AWERProcedia Inform. Technol. Comput. Sci.* 03 (2013) 1700.
- [3] R. Jackendoff, Possible stages in the evolution of the language capacity, *Trends Cogn. Sci.* 3 (1999) 272.
- [4] M.D. Hauser, N. Chomsky, W. Tecumseh Fitch, The faculty of language: What is it, who has it, and how did it evolve?, *Science* 298 (2002) 1569.
- [5] P. Foster, A. Toth, Toward a phylogenetic chronology of ancient gaulish, celtic, and indo-european, *Proc. Natl. Acad. Sci. USA* 100 (2003) 9079.
- [6] D.G. Russell, D.A. Quentin, Language-tree divergence times support the Anatolian theory of Indo-European origin, *Nature* 426 (2003) 435.
- [7] I. Grabska-Gradzińska, A. Kulig, J. Kwapień, S. Drozd, Complex network analysis of literary and scientific texts, *Internat. J. Modern Phys. C* 23 (2012) 1250051.
- [8] M.A. Nowak, D.C. Krakauer, The evolution of language, *Proc. Natl. Acad. Sci. USA* 96 (1999) 8028.
- [9] M.A. Nowak, J.B. Plotkin, V.A.A. Jansen, The evolution of syntactic communication, *Nature* 404 (2000) 495.
- [10] P.W. Anderson, More is different, *Science* 177 (1972) 393.
- [11] J. Kwapień, S. Drozd, Physical approach to complex systems, *Phys. Rep.* 515 (2012) 115.
- [12] A.P. Masucci, G.J. Rodgers, Network properties of written human language, *Phys. Rev. E* 74 (2006) 026102.
- [13] S.M.G. Caldeira, T.C. Petit Lobão, R.F.S. Andrade, A. Neme, J.G.V. Miranda, The network of concepts in written text, *Eur. Phys. J.* 49 (2006) 523.
- [14] L. da Fontoura Costa, O.N. Oliveira Jr, G. Travieso, Analyzing and modeling real-world phenomena with complex networks: a survey of applications, *Adv. Phys.* 60 (2011) 329.
- [15] S. Zhou, G. Hu, Z. Zhang, J. Guan, An empirical study of Chinese language networks, *Physica A* 387 (2008) 3039.
- [16] R.M. Roxas-Villanueva, M.K. Nambatak, G. Tapang, Characterizing english poetic style using complex networks, *Internat. J. Modern Phys. C* 23 (2012) 1250009.
- [17] Y. Yang, C. Gu, Q. Xiao, H. Yang, Evolution of scaling behaviors embedded in sentence series from A story of the stone, *PLOS ONE* 12 (2017) e0171776.
- [18] J. Laherrere, D. Sornette, Stretched exponential distributions in nature and economy: fat tails with characteristic scales, *Eur. Phys. J. B* 2 (1998) 525.
- [19] E.G. Altmann, J.B. Pierrehumbert, A.E. Motter, Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words, *PLOS ONE* 4 (2009) e7678.
- [20] M.A. Montemurro, P.A. Pury, Long-range fractal correlations in literary corpora, *Fractals* 10 (2002) 451.
- [21] E.G. Altmann, G. Cristadoro, M.D. Esposti, On the origin of long-range correlations in texts, *Proc. Natl. Acad. Sci.* 109 (2012) 11582.
- [22] M. Ausloos, Generalized hurst exponent and multifractal function of original and translated texts mapped into frequency and length time series, *Phys. Rev. E* 86 (2012) 031108.

- [23] M. Ausloos, Measuring complexity with multifractals in texts: Translation effects, *Chaos Solit. Fract.* 45 (2012) 1349.
- [24] R.F. Cancho, R.V. Sole, The small world of human language, *Proc. Roy. Soc. Lond. B* 268 (2001) 2261.
- [25] J. Cong, H. Liu, Approaching human language with complex networks, *Phys. Life Rev.* 11 (2014) 598.
- [26] A. Kulig, S. Drozd, J. Kwapien, P. Oświecimka, Modeling the average shortest-path length in growth of word-adjacency networks, *Phys. Rev. E* 91 (2015) 032810.
- [27] G.K. Zipf, *Human Behavior and the Principle of Least Effort*, Addison-Wesley, Cambridge, 1949.
- [28] G.K. Zipf, *The Psycho-Biology of Language, an Introduction to Dynamic Philology*, MIT Press, Cambridge, MA, 1965.
- [29] B. Mandelbrot, Information theory and psycholinguistics: a theory of words frequencies, in: P. Lazafeld, N. Henry (Eds.), *Readings in Mathematical Social Science*, Cambridge, MA, 1966.
- [30] B. Mandelbrot, *The Fractal Structure of Nature*, Freeman, New York, 1983.
- [31] M.A. Montemurro, Beyond the zipf–mandelbrot law in quantitative linguistics, *PhysicaA* 300 (2001) 567.
- [32] A. Kulig, J. Kwapien, T. Stanis, S. Drozd, In narrative texts punctuation marks obey the same statistics as words, *Inform. Sci.* 375 (2017) 98.
- [33] A. Mehri, S.M. Lashkari, Power-law regularities in human language, *Eur. Phys. J. B* 89 (2016) 241.
- [34] M.A. Montemurro, D.H. Zanette, Complexity and universality in the long-range order of words, in: M. DegliEsposti, E. Altmann, F. Pachet (Eds.), *Creativity and Universality in Language*, in: *Lecture Notes in Morphogenesis*, Springer, Cham, 2016.
- [35] L. Lü, Z.K. Zhang, T. Zhou, Deviation of Zipf's and Heaps' Laws in human languages with limited dictionary sizes, *Sci. Rep.* 3 (2013) 1082.
- [36] R. Arnheim, *Entropy and Art. An Essay on Disorder and Order* (Berkley and Los Angeles: University of California Press, 1971).
- [37] K.G. Denbigh, J.S. Denbigh, *Entropy in Relation to Incomplete Knowledge* (Cambridge: Cambridge University Press, 1985).
- [38] R. M. Gray, *Entropy and Information Theory* (New York: Springer-Verlag, 1990).
- [39] C.E. Shannon, "A Mathematical Theory of Communication", *Bell System Technical Journal*, № 27 (1948), p. 379-423; p. 623 – 656