Int. J. Adv. Res. Sci. Technol. Volume 4, Issue 2, 2015, pp.308-312.

IJARST

International Journal of Advanced Research in Science and Technology

journal homepage: www.ijarst.com

ISSN 2319 – 1783 (Print)

ISSN 2320 – 1126 (Online)

Mathematical formulation of the writing pattern of different texts based on text length and word length frequency in intra text perspective for stories of Hindi language and comparison for two authors

Hemlata Pande and H. S. Dhami*

Department of Mathematics, Kumaun University, S. S. J. Campus Almora, Uttarakhand, India

****Corresponding Author's Email:hlpande@rediffmail.com; Address for correspondence: Govt. P. G. College, Bageshwar.***

A R T I C L E I N F O

Article history:

Received 29 Jan. 2015

Accepted 05Feb. 2015

Available online 16 Feb. 2015

Keywords:

Text length, word length, model, authorship attribution

A B S T R A C T

Present paper is an innovative attempt in the direction to represent the relation among text length (determined in two different ways) and the total frequency of words of particular lengths in mathematical form. The texts written by two different authors have been compared on the basis of the determined formulation.

Introduction:

In Computational literary methods, linguistic variables are quantified, language is considered as the object for scientific study and extensive statistical and mathematical techniques are provided to examine literary data (Bagavandas and Manimannan, 2008). Dash (2008) has noted that application of the techniques, which "are the outcomes of an interface between computer and language corpora", generates "new evidence to describe language and its properties from different perspectives." Köhler and Altmann (2008) quoted that in Quantitative Linguistics discipline, quantitative interrelations are valuable in fundamental research and also have applications in different procedures, for example in natural language processing, in optimization of texts, language teaching and computational linguistics etc. Efforts have been made to represent different relations for linguistic elements in mathematical form by various researchers. In this regard following works can be cited to mention only a few: in the book by Popescu et al (2009, p.89) the proportion of auxiliaries within the given class (for example the class of words occurring once, twice..) have been formulated in terms of number of incidence of words; Pruscha (1998) has presented the relation between the text length and vocabulary of text, Zorbaz (2007) has examined the readability level of Turkish textbooks with the help of Ateşman's formula – depending on word and sentence length. Grzybek et al (2008) have tested the sentence length and word length relation with the help of Altmann-Menzerathian line. Buk and Rovenchak (2008) have examined the syntactic structure in Ukrainian and have determined the dependence of clause length on the sentence length.

In the current work we have attempted to determine the interrelation between the text lengths (determined in two different ways) and the number of words of particular lengths. The study is based on the data determined for 37 stories written by the two authors namely '*Rabindranath Tagore*', and '*Mohan Rakesh*'. The relation for the text length, measured in number of characters, has been expressed in terms of the total frequency of words of lengths two to four occurred in the text and the text length measured in terms of number of sentences.

Bagavandas and Manimannan (2008) have cited that: by counting and measuring stylistic attributes "it is possible to discover the characteristics of a particular author or a particular genre"; Popescu et al (2009, p. 250) have also referred to the fact that individual styles and authors "can be better distinguished by means of some frequency indicators than by means of qualitative indicators". Similarly Fengxiang (2007) has quoted that according to Butler (1985) word length and sentence length can be utilized as stylistic features. The works of Antic et al(2006), Kelih et al (2005, 2006) etc. can be cited regarding the applications of word or sentence length in the discrimination of genres and/or authors to mention only a few. We, in our earlier works (Pande and Dhami, 2013, 2016), have examined the role of word length features in classification of written texts of Hindi language and applied the sentence length frequency profile for authorship attribution for Hindi stories respectively. In the present study after

determination of the relation for text length, we have compared the writings of the two authors on the basis of the identified relation.

Methodology:

Data:

The work has been initiated with the selection of 37 stories, written by two different authors: '*Rabindranath Tagore*', and '*Mohan Rakesh*', for data analysis from the website of '*Hindi Samay*'. These stories have also been studied for the sentence length frequencies in our earlier work (Pande and Dhami, 2016). Data have been determined for the word length and text length as follows:

Sentence length frequency and corresponding text length: Similar to our earlier work (Pande and Dhami , 2016), the sentence length frequency data have been determined by the application of the 'Sentence Analyzer' (source available at the website of Central Institute of Indian Languages) after a few modifications of stories and with the help of the sentence length frequency data, the text length (LT_2) in terms of the number of sentences in the text has been determined.

Word length frequency: The data for the word length frequency distribution has been obtained with the help of MS Word tools by considering the length of a word as the total number of characters contained in the word. The number of words of particular length (2-4) occurred in the text has been calculated with the help of the determined data. Before the determination of the data for the word lengths, all the consonants of the *Devanāgari* alphabet which occurred in the modified form (mentioned in Pande and Dhami, 2015), with a dot below them, have been replaced by corresponding consonants.

Text Length in number of characters: The text length is measured in terms of total number of characters occurred in the text in forming the alphabetic words of the text. The space character, numerals and punctuation marks have not been considered in counting the text length. The length of text is determined with the help of word length frequency data as $TL_1 = \sum_{l=1}^{p} l f_l(w)$, where $f_l(w)$ is the frequency of words of length *l* occurred in the text and *p* is the highest length corresponding to which the frequency of occurrence of words in text is non zero.

Mathematical Formulation:

On the basis of the obtained data it has been noted that the total proportion of the words of lengths 2, 3, 4 in different texts is more than 65% (of total words) in all the considered texts and the average value of their total proportion for the considered texts is 76.67%. Thus the words of the mentioned lengths form a considerable part of stories of Hindi language. We have tried to relate their proportion in text with the text lengths. The aim of the present paper is to identify the pattern of variation of the text length (in characters) with the frequencies of the words above discussed lengths and number of sentences in text. It is not claimed that the text length (in number of characters) depends only in these variables but we want to demonstrate that if these variables are selected then what is the formulation and how it can be useful in determination of authorship?

We have attempted to formulate the relation of text length LT_1 in terms of text length LT_2 and the total frequency of the words (say, f_w) having lengths two, three or four ($f_w = f_2(w) + f_3(w) + f_4(w)$) in the intra text perspective. The data for LT_2, f_w and LT_1 have been determined for each of the considered texts with the help of sentence length and word length frequency tables. To obtain the required relation, we have first calculated the value of coefficient of determination (R^2) for the linear relation between LT_1 and f_w as well as between LT_1 and LT_2 with the help of the obtained empirical values of the variables for all considered texts. The values of coefficient of determination in semi logarithmic and log-log scale have also been determined. Calculated values of R^2 have been listed in the Table: 1.

Table: 1. Values of the R^2 for various linear relations

Relation	Value of R^2	Relation	Value of R^2
$LT_1 = 4.242 f_w + 953.3$	0.986	$LT_1 = 45.4 LT_2 + 1474$	0.916
$\log(LT_1) = 0.00013 f_w + 3.689$	0.865	$\log(LT_1) = 0.001 LT_2 + 3.711$	0.778
$LT_1 = 28171 \log(f_w) - 82025$	0.883	$LT_1 = 29045 \log(LT_2) - 54699$	0.838
$\log(LT_1) = 0.937 \log(f_w) + 0.878$	0.984	$\log(LT_1) = 0.946 \log(LT_2) + 1.835$	0.911

Thus the Table (Table: 1) depicts that the values of R^2, obtained corresponding to the double logarithmic scale, in the case of both word length and text length (in number of sentences), are nearly equal to the corresponding values obtained in simple scale (linear relation between variables). But as we want to determine a relation which will be applied to compare the stories of the two authors, therefore we have preferred the relation corresponding to the double logarithmic scale as it provides better estimation also in cases when the values of f_w and LT_2 are small; while the obtained linear relation provides the text length $LT_1 > 950\ and\ 1450$ even in the cases when $f_w = 1\ and\ LT_2 = 1$ respectively. Thus we have taken up the power law relation between LT_1 and f_w and between LT_1 and LT_2.

By combining the two power laws, we have assumed the relation for the above three variables in the form:

$$LT_1 = a(f_w)^b (LT_2)^c \qquad (1)$$

where *a*, *b* and *c* are parameters. The model in the form of equation (1) has been fitted to the data of 37 stories for the two defined text lengths, frequencies of words having length two, three or four with the help of Datafit 9.1.32. Determined model is as under:

$$LT_1 = 5.98636(f_w)^{1.04767} (LT_2)^{-0.11518} \qquad (2)$$

The observed values and the corresponding determined values by the equation (2) for the considered texts have been shown in the following figure (Figure: 1). The deviations from the actual values from the corresponding theoretical values have also been depicted in the figure.

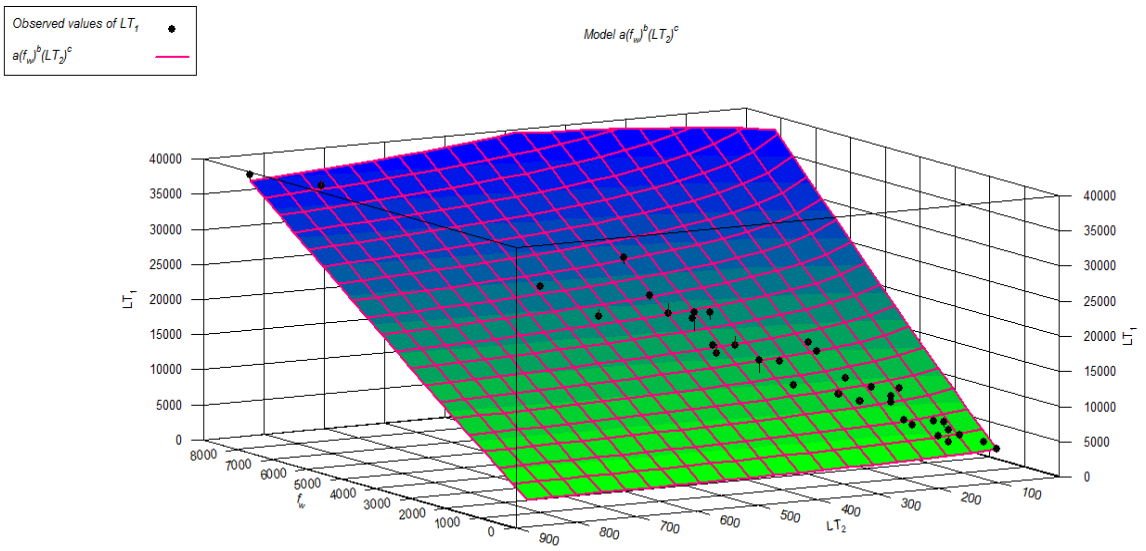

Figure: 1. Observed and corresponding theoretical values of text length (in number of characters)

The values of the coefficient of multiple determination (R^2) and adjusted coefficient of multiple determination (R_a^2)[1] for the relation mentioned in equation (2) have been obtained as 0.988 and 0.987 respectively for the theoretical and observed text lengths for considered texts. As the values of the two coefficients of determination are contiguous to one, therefore the model mentioned in equation (1) can be assumed as the model that represents variation of text length (in number of characters) with respect to the number of words of particular lengths and the text length calculated in number of sentences.

Comparison of style of two authors

After determination of the model for all the considered texts we have applied it for the stories of the two considered authors separately and the obtained results for the fitted model in the case of two authors: for 21 stories of '*Rabindranath Tagore*', and 16 stories of '*Mohan Rakesh*' have been listed in the table below:

[1] From DataFit:

$$R^2 = 1 - \frac{error\ \ sum\ \ of\ \ squares}{total\ \ sum\ \ of\ \ squares}$$

$$R_a^2 = \frac{(n-1)R^2 - k}{n-1-k},\quad k = \text{number of parameters in the model } \&\ n = \text{number of data points}$$

Table: 2. Determined formulae of text length in case of two authors

Relation		Author
$LT_1 = 5.95904(f_w)^{0.88888}(LT_2)^{0.10026}$	I	*Mohan Rakesh*
$LT_1 = 8.07174(f_w)^{0.84879}(LT_2)^{0.12695}$	II	*Rabindranath Tagore*

The equation (2) and the two relations in the Table: 2 demonstrate that the exponents of f_w in the equations have values close to unity.

Now for all the considered texts we have determined the value of LT_1 with the help of both of the equation I and II mentioned in the above table. The determined and observed values of LT_1 for different considered texts of two authors have been depicted with the help of following figure (Figure: 2) and the absolute values of the difference of the text lengths from the calculated values (when determined with the help of relation I and II) have been presented in the Table: 3.

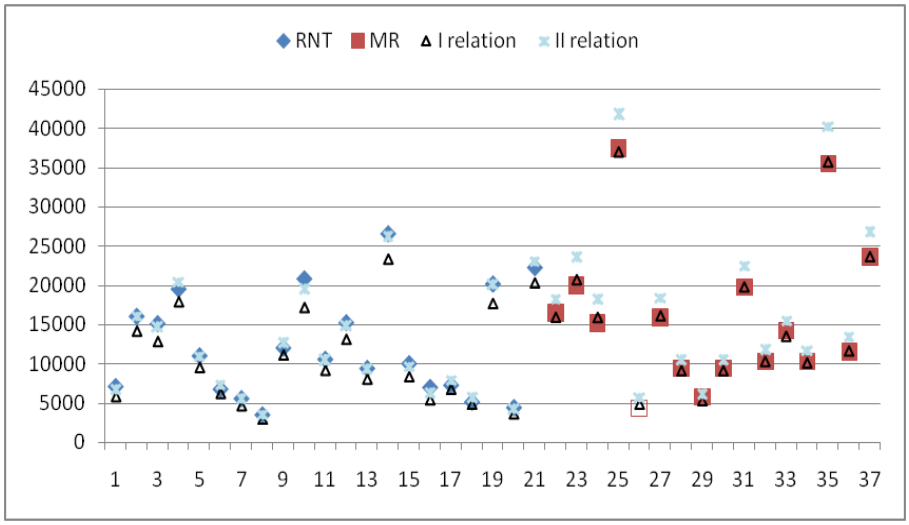

Figure: 2. Empirical vales of text lengths for the stories of two authors and corresponding evaluated values by using two relations I and II

Table: 3. Absolute values of the differences of determined text lengths from observed values of text lengths

Text	Absolute value of the difference of determined text length from actual text length		Text	Absolute value of the difference of determined text length from actual text length		Text	Absolute value of the difference of determined text length from actual text length	
	Relation I	Relation II		Relation I	Relation II		Relation I	Relation II
RNT1	1210.328	297.9261	RNT14	3182.679	277.7905	MR6	231.2083	2415.207
RNT2	1828.899	7.452082	RNT15	1594.4	369.9764	MR7	294.5283	1073.063
RNT3	2214.594	310.2059	RNT16	1518.453	668.3816	MR8	474.6583	353.5586
RNT4	1555.08	861.0156	RNT17	459.4074	608.7714	MR9	294.1204	1075.477
RNT5	1425.228	65.49744	RNT18	200.1161	561.8747	MR10	62.17394	2654.189
RNT6	466.8762	535.3778	RNT19	2435.006	88.36928	MR11	37.63943	1552.593
RNT7	834.6271	35.0197	RNT20	720.171	129.6926	MR12	691.9236	1202.293
RNT8	467.9655	27.07751	RNT21	1917.534	760.5197	MR13	156.5403	1341.563
RNT9	835.207	703.7339	MR1	538.1967	1701.265	MR14	230.9654	4674.776
RNT10	3607.267	1239.298	MR2	755.6471	3607.218	MR15	115.596	1844.931
RNT11	1338.988	103.0728	MR3	765.3171	2994.244	MR16	28.91066	3214.123
RNT12	2070.824	325.4771	MR4	505.5281	4205.516			
RNT13	1293.297	89.84739	MR5	529.9378	1322.781			

In the Figure: 2 and Table: 3, RNT and MR abbreviations have been used for the stories of '*Rabindranath Tagore*' and '*Mohan Rakesh*' respectively. Thus the figure and the table demonstrate

that the values of LT_1 are determined in better way with the help of the relation II in the case of stories of author '*Rabindranath Tagore*' and similarly by relation I in case of *'Mohan Rakesh'*. Thus these two equations can be assumed as supportive for the authorship determination for the stories of the two writers.

Conclusions:

On the bases of the study of data for 37 stories written by two authors we have concluded that the law of variation of text length of the Hindi stories can be formulated as:

Text length in number of characters =

a (total *frequency of words of having length two to four*)^b (*text length in number of sentences*)^c;

where *a*, *b* and *c* are parameters and the value of the exponent *b* is close to unity.

The fact propounded by Bagavandas and Manimannan in 2008 that the style is the relation between a writer and writer's work is also certified by us with the help of the determination of the laws of variation of text length in the form of two equations and presentation of their better pertinence in the case of data of texts of corresponding author. This new approach determined by us can be tested and applied in the case of the writings of different authors and different languages and hence has significance for the authorship attribution process.

Acknowledgement:

Authors are grateful to the University Grants Commission (UGC), New Delhi, INDIA for providing financial assistance in the form of Postdoctoral fellowship [F.4-2/2006(BSR)/13-770/2012(BSR)] to the first author. The research has been sponsored by the UGC under the 'UGC Dr. D. S. Kothari Postdoctoral fellowship scheme'.

References:

1. Antic G.; Stadlober E.; Grzybek P.; Kelih E. (2006). Word length and frequency distributions in different text genres. In M. Spiliopoulou et al. (Eds.): From Data and Information Analysis to Knowledge Engineering. Springer, Heidelberg. pp. 310-317.
2. Bagavandas, M. and Manimannan, G. (2008). Statistics and Literary Analysis. In P. Mohanty and R. Köhler(Eds.). Readings in quantitative linguistics. Indian Institute of Language Studies, Delhi. pp. 45-58.
3. Buk S. and Rovenchak, A. A. (2008). Menzerath-Altmann law for syntactic structures in Ukrainian. Glottotheory, vol. 1, No. 1, pp. 10-17.
4. Butler, C. (1985). Computers in Linguistics. Blackwell, Oxford.
5. Dash, N. S. (2008). The Techniques of Text Corpus Processing. In P. Mohanty and R. Köhler(Eds.). Readings in quantitative linguistics. Indian Institute of Language Studies, Delhi. pp. 81-112.
6. Fengxiang, F. (2007). A corpus based quantitative study on the change of TTR, word length and sentence length of the English language. In P. Grzybek and R. Köhler (Eds.) Exact Methods in the Study of Language and Text. Mouton de Gruyter Berlin – New York. pp. 123-130.
7. Grzybek, P. ,Kelih E. and Stadlober E. (2008). The relation between word length and sentence length: an intra - systemic perspective in the core data structure. Glottometrics, vol. 16, pp. 111-121.
8. Kelih, E., Antic, G., Grzybek, P. and Stadlober, E. (2005). Classification of Author and/or Genre? The Impact of Word Length. In: C. Weihs, and W. Gaul (Eds.), Classification – The Ubiquitous Challenge. Springer, Berlin, pp. 498–505.
9. Kelih, E., Grzybek, P., Antic, G. and Stadlober, E. (2006). Quantitative Text Typology: The Impact of Sentence Length. In: M. Spiliopoulou et al. (Eds.): From Data and Information Analysis to Knowledge Engineering.Springer, Berlin, pp. 382–389.
10. Köhler , R. and Altmann, G. (2008). Aims and Scope of Quantitative Linguistics. In P. Mohanty and R. Köhler(Eds.). Readings in quantitative linguistics. Indian Institute of Language Studies, Delhi. pp. 1-32.
11. Pande, H. and Dhami, H. S. (2013). Analysis for the significance of statistical word-length features in genre discrimination of Hindi texts. IOSR Journal of Mathematics 8(1). pp. 5-10.
12. Pande, H. and Dhami, H. S. (2015). Analysis and Mathematical Modelling of the Pattern of Occurrence of Various Devanāgari Letter Symbols according to the Phonological Inventory of Indic Script in Hindi Language. Journal of Quantitative Linguistics, 22:1, pp. 22-43.
13. Pande, H. and Dhami, H. S. (2016). Determination of distribution of sentence length frequencies for Hindi language texts and utilization of sentence length frequency profile for authorship attribution. Journal of Quantitative Linguistics 23(1). (to appear).
14. Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B. D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L. and Vidya, M.N. (2009). Word Frequency Studies. Berlin, New York: de Gruyter.
15. Pruscha, H.(1998). Statistical models for vocabulary and text length with an application to the NT corpus. Literary and Linguistic Computing, vol. 13(4), pp. 195-198.
16. Zorbaz, K. Z. (2007). An evaluation on the word-sentence lengths and readability levels of tales in turkish textbooks. Journal of Theory and Practice in Education 3 (1), pp. 87-101