Mathematical Modelling of Occurrence of Letters and Word's Initials in Texts of Hindi Language

Hemlata Pande and H. S. Dhami

Despite the apparent freedom a writer has to create any desired sequence of words, written text tends to follow some very simple set of laws. Frequencies of occurrence of letters in natural language texts are not arbitrarily organized but obey some particular rules which enable us to explain some characteristic features of human language. Present paper is an attempt to analyze the frequencies of letters of Hindi language alphabet on the basis of their occurrence in a text and also for their occurrence as word's initials. We have applied Zipf's law to the letters instead of words for obtaining Zipf's order and rank/frequency profiles of letters for their instances in text and in beginning of words. Different distribution models have been tested for the letter frequencies and finally appropriate parametric models have been obtained for the rank frequency distribution of letters in a corpus and for the rank frequency distribution of first letter of words in the corpus. Validations of the models have been done by comparison of observed frequencies and theoretical frequencies for various texts obtained from different sources, and the variations in the values of parameters have been studied. Comparison work has been accomplished for the entropies in text for the vowel mark in Devanagari script written as matra following the consonant of the most frequent consonants.

Keywords*: rank, frequency, model, entropy, Zipf's order.*

1. Introduction

Linguistics is the scientific study of language. Its aim is to be able to exemplify and explain the multitude of linguistic observations circling around us, in conversation, writing and other media. The underlying goal of the linguist is to try to determine the universals concerning language. That is, what are the common elements of all languages? The linguist then tries to place these elements in a theoretical framework that could describe all languages and could also predict what can not occur in a language. The properties of linguistic elements and their interrelations abide by universal laws which are formulated in a strict mathematical way, in analogy to the laws of well known natural sciences and it leads to the application of some abstract mathematical and statistical tools to the language theory**.** Mathematical linguistics offers a number of tools that allow abstract theories of language and mental computation to be compared and contrasted with one another. An account of work in the field of mathematical methods of linguistics can be had from the book of Partee et al (1990), while compendium of several rules proposed to understand the linguistic structure by which language communicates can be seen in the work of Manning and Schutze (1999). One of the branches of mathematical linguistics is quantitative linguistics. It studies language by means of determining the frequencies of various words, word combinations, and constructions in texts. Currently, quantitative linguistics mainly means statistical linguistics. It provides the methods of making decisions in text processing on the base of previously gathered statistics.

Statistical techniques have brought significant advances in broad-coverage language processing.

Zipf's ideas are the foundation stones of modern quantitative linguistics. His influence is not only restricted to linguistics but also incessantly penetrates other sciences as mentioned by Altmann (2002). Zipf has ranked words, in the descending order of their frequencies of occurrence, in order to find a relation between the frequency of a word and its rank and stated that the frequency of the r^{th} most frequent word type is proportional to *1/r*. Consequently this rank frequency approach was applied to different other components (than word) of language also. In this context we can cite the works of Sigurd, for suggesting a geometric series equation for the distribution of phoneme frequencies and of Good, for suggesting equation to describe the ranked distribution of phoneme and grapheme frequencies, as given in the research paper of Tambovtsev and Martindale (2007); Grzybek and Kelih (2005) for discussing a possible theoretical model for grapheme frequencies of Slavic alphabet and then suggesting that the negative hypergeometric distribution is as an adequate model; Eftekhari (2006) for using Zipf's idea for introducing two new concepts Zipf's dimension and Zipf's order of letters in order to have a fractal geometrical approach for English texts; Tambovtsev and Martindale (2007) for propounding that Yule equation fits the distribution of phoneme frequencies better than other distributions as Zipf, Good etc. Pande and Dhami (2009) have determined the model for grapheme frequencies by modification of Zipf's Mandelbrot distribution.

The process of formation of language models is quite popular in linguistics. A linguistic model is a system of data (features, types, structures, levels, etc.) and rules, which, taken together, can exhibit a "behavior" similar to that of the human brain in understanding and producing speech and texts. According to Bell and Witten (1988)

> Natural-language text is usually the result of an enormously complex process. Many hours- or even years- of human thought can lie behind just a few dozen of words. It is remarkable that some very simple modeling strategies can be applied successfully to such a sophisticated artifact.

Mathematical models for languages had been proposed by Harris (1982) in his famous book entitled *A Grammar of English on Mathematical Principles*. In statistical language modeling, large amounts of text are used to automatically determine the model's parameters. A compendium of different statistical models can be seen in the book of Manning and Schutez (1999). Naranan and Balasubrahmanyan (1998) have described the use of power law relations in modeling of languages. A parsing language model incorporating multiple knowledge sources that is based upon the concept of constraint Dependency Grammars is presented in the paper of Wang and Harper (2002). A neural probabilistic language model has been generated by Bengio et al (2003) while hierarchical probabilistic neural network language model has been generated by Morin and Bengio (2005).Different neural network language models can be seen in the work of Bengio (2008).

The frequency of letters in text has often been studied for use in cryptography and frequency analysis. Modern International Morse code encodes the most frequent letters with the shortest symbols and a similar idea has been used in modern data-compression techniques also. Linotype machines are also based on letter frequencies of English language texts. The reference for these applications can be had from Nation Master- Encyclopedia ("Letter frequencies"). In the area of letter-frequencies, we can cite the works of Solso and King

(1976) for examining frequency and versatility of letters in English language; Bell and Witten (1988) for presenting letter statistic for brown corpus and of Sanderson (2007) for exploration of the fact that distribution of letters in a text can potentially help in the process of language determination.

Since the patterns of letters in language don't happen at random, so letter frequencies can be represented in the form of some rules and the pattern of letter frequencies can help in discrimination of random text from that of natural language text. We have tried to discuss the patterns of Hindi language alphabets on the basis of their occurrence in texts and as word initials with the application of the rank frequency approach. The detailed analyses and the modeling done in the paper is an attempt in the direction of the determination of the linguistic structure of Hindi Language (on the basis of its letters). Most of the previous studies for the letter frequencies have been concreted for English and other languages.

2. Occurrence of letters in a text

We have initiated our work by counting the occurrence of letters of Hindi language alphabet in the text "Tanav se mukti" of author Shivanand[1] (source has been given in appendix). Following aspects have been considered in the counting process:

- Independent occurrence of each letter (vowel as well as consonant), as the occurrences of 'अ' and 'य' in 'अध्याय'.
- Each occurrence of consonants followed either by a vowel mark in Devanagari script written as *matra,* as occurrence of क in का, कां, काँ, कः or occurrence in the form of half letters anywhere in the word. For example we may cite the example of occurrence of क in वक्त and as such रविन्द्र = र +व +ि +न +् +द +् +र contains 2 'र', 1 'व', 1 'न' and 1 'द' and ईर्ष्या (ई +र+ ् +ष +्+ य +ा) contains four letters ई, र, ष, य once each.
- Occurrence of आँ has been considered as the occurrence of आ and of अँ, आँ, etc. with अं, आं etc. Occurrences in the form of क़, ख़ etc. have been merged with the occurrences of क, ख and so on.

In our selected text "Tanav se mukti" the total counted values of the frequency of occurrence of each letter have been depicted in the following Table 1. If we represent the total occurrence of all the letters of alphabet in the text by N_1, that is, $N_1 = \sum_k f_k$, k = 1, 2….. , where f_k is the frequency of occurrence of k^{th} letter of the alphabet of Hindi language, then its value for the selected text shall be equal to 41,114.

[1] The text has been taken as an example text by us.

Letter	Frequency	Letter	Frequency	Letter	Frequency	Letter	Frequency
अ	711	क	4468	ढ	92	र	4116
आ	517	ख	341	ण	316	ल	1128
इ	99	ग	679	त	3120	व	1850
ई	167	घ	122	थ	543	श	589
उ	458	ङ	0	द	1260	ष	361
ऊ	17	च	661	ध	490	स	2645
ए	449	छ	194	न	3236	ह	3178
ऐ	23	ज	848	प	1858	क्ष	125
ओ	54	झ	117	फ	84	त्र	143
औ	322	ञ	3	ब	716	ज्ञ	25
अं	16	ट	294	भ	805		
अः	0	ठ	128	म	1921	Total	41,114
ऋ	2	ड	270	य	1553		

Table 1 *Frequencies of occurrence of various letters in the text "Tanav se Mukti"*

We have ranked the letters in descending order of their frequency of occurrence as done by Zipf. The rank and frequency profile of the text *Tanav se Mukti*, for the letters in the text, can be visualized as shown in the following Table 2-

Rank	Frequency	Rank	Frequency	Rank	Frequency	Rank	Frequency
1	4468	13	848	25	361	37	117
2	4116	14	805	26	341	38	99
3	3236	15	716	27	322	39	92
4	3178	16	711	28	316	40	84
5	3120	17	679	29	294	41	54
6	2645	18	661	30	270	42	25
7	1921	19	589	31	194	43	23
8	1858	20	543	32	167	44	17
9	1850	21	517	33	143	45	16
10	1553	22	490	34	128	46	3
11	1260	23	458	35	125	47	2
12	1128	24	449	36	122		

Table 2 *Rank frequency profile of the text "Tanav se Mukti"*

According to Eftekhari (2006) if the letters are ordered in accordance with the ascending order of the frequencies, a monotonic changes can be obtained by the application of the Zipf 's law. Such type of arrangement of letters has been given the name "Zipf 's order". For our text "Tanav se Mukti" the Zipf's order of letters for their occurrence in the text has been obtained as shown in Figure 1.

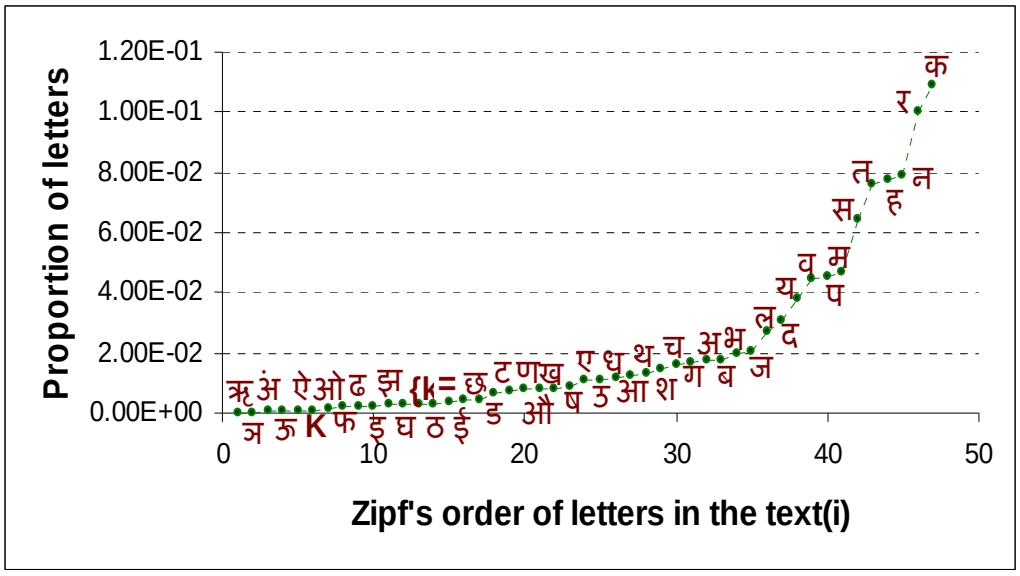

Figure 1 *Zipf's order arrangement of letters and their corresponding proportions in the text "Tanav se Mukti"*

3. Occurrence of letters as word's initials

Results obtained by counting of letters for their occurrence as initial letter of words in the studied text have been tabulated in the following Table 3.

Letter	frequency	letter	frequency	letter	frequency	letter	frequency
अ	707	क	2708	ढ	8	र	372
आ	495	ख	64	ण	0	ल	391
इ	86	ग	155	त	419	व	555
ई	56	घ	70	थ	23	श	191
उ	457	ङ	0	द	672	ष	2
ऊ	16	च	295	ध	133	स	1581
ए	195	छ	79	न	550	ह	2153
ऐ	23	ज	599	प	1020	क्ष	62
ओ	20	झ	23	फ	52	त्र	4
औ	322	ञ	0	ब	527	ज्ञ	9
अं	16	ट	21	भ	501		
अः	0	ठ	13	म	963	Total	16,799
ऋ	2	ड	31	य	158		

Table 3. *Frequencies of occurrence of various words' initials in the example text*

Here if N_2 represents the total word tokens in the text (where words formed by numerals 1, 2 etc. have been excluded), then $N_2 = 16{,}799$.

A table similar to the table 2 can be constructed for the rank frequency profile of the text corresponding to word's initials. The Zipf's order can be demonstrated with the help of Figure 2.

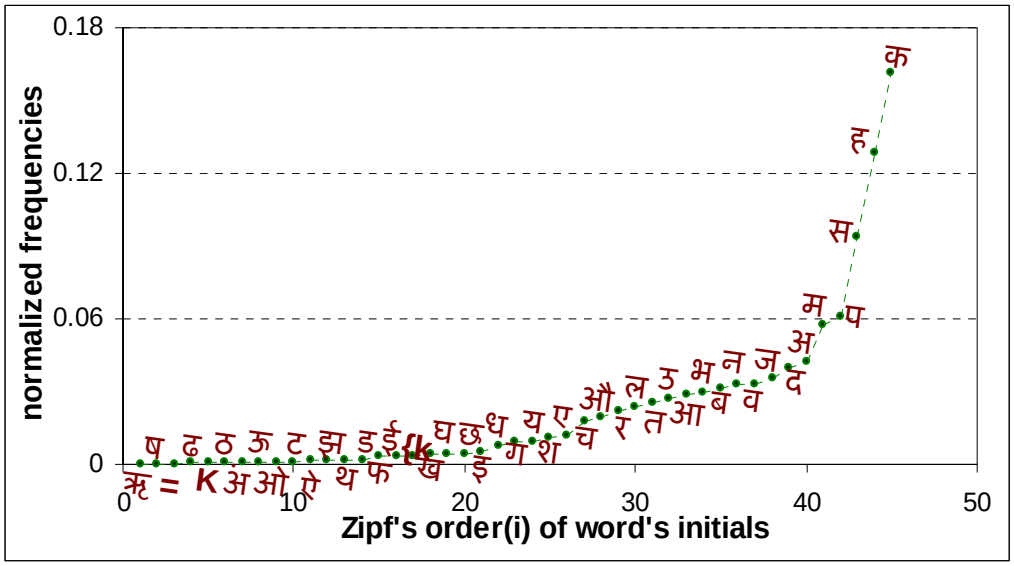

Figure 2. *Zipf's order arrangement of initials of words of the text and their normalized frequencies, obtained from the Table 3*

3. Mathematical models for frequencies of letters

After obtaining the rank frequency profiles of texts for different letters and for word's initials we have tried to formulate these in the form of mathematical equations. For this we have tested following models available in literature for rank frequency distributions:

Distributions specified for rank frequency distribution of word frequencies, Zipf's distribution and Zipf Mandelbrot distribution as given in the book by Manning and Schutze (1999); various rank frequency distributions used for linguistic components as- Good, Geometric series distribution, Borodovsky and Gusein Zade equation as cited in the research paper of Tambovtsev and Martindale (2007) and Yule equation as concluded for the phoneme frequencies by the authors; negative hypergeometric distribution and determination of frequencies by Zipf's order.

1. Zipf distribution:

Zipf's law has been defined in the work of Manning and Schutze (1999) as a relation between the frequency of occurrence of an event and its rank when the events are ranked with respect to the frequency of occurrence (the most frequent one first). According to Zipf's law 'if the frequency of occurrence of an event ranked in descending order of frequency, then

(1) $F_r \propto \frac{1}{r}$ *or* $F_r r = c,$

c is a constant, where F_r is the frequency of event of r^{th} rank'. General form of this formula was given by him in the later works as

(2) $F_r = \frac{a}{r^b}$,

where a and *b* are parameters of text.

2. Zipf Mandelbrot distribution:

The modification to Zipf formula has been introduced by Mandelbrot as cited in the book of Manning and Schutze (1999) by including an additional parameter *c* to express the formula of rank frequency distribution in the form

(3) $F_r = \frac{a}{(r+c)^b}$,

where a , *b* and *c* are parameters.

3. Good distribution:

Good has suggested an equation for ranked distribution of phoneme and grapheme frequencies as-

(4) $F_r = \frac{1}{n} \sum_{i=r}^{n} \frac{1}{i}$,

where *n* is number of symbols and F_r is normalized frequency of event of rank *r*.

4. Geometric series distribution:

Sigurd has suggested a geometric-series equation for the ranked distribution of phoneme frequencies in the form:

(5) $F_r = ak^{r-1}$,

where a is the frequency of the most frequent phoneme, *k* is the parameter to be estimated, and *r* is rank. To reduce the dependency of this equation on most frequent phoneme, the researcher has modified the equation as

(6) $F_r = \frac{(1-k)k^{r-1}}{1-k^n}$,

where *n* is the number of symbols, and *k* is the parameter to be estimated and F_r , the normalized frequency corresponding to rank *r*.

Thus the general form of the geometric series equation is:

(7) $F_r = ab^r$,

where a and *b* are parameters and F_r be the frequency for rank *r*.

5. Equation given by Borodovsky and Gusein Zade:

Gusein-Zade and Borodovsky have suggested a parameter-free equation, by assuming that F_r depends on $\log\ (r)$, rather than that $\log\ (F_r)$ as assumed in Zipf's equation , as-

(8) $F_r = \frac{1}{n}(\log(n+1) - \log r)$,

where *n* is the number of symbols. They have presented evidence from distributions of letter frequencies of four languages.

The references of the works cited at serial numbers 3, 4 and 5 are available in the research paper of Tambovtsev and Martindale (2007).

6. Yule distribution:

Tambovtsev and Martindale (2007) proposed Yule equation for phoneme frequencies, and have showed that the Yule equation fits the distribution of phoneme frequencies better than the Zipf equation or equations proposed by Sigurd , Borodovsky and Gusein-Zade. The equation has been given as-

(9) $F_r = \frac{a}{r^b}c^r$,

where *a, b, c* are parameters.

7. Negative hypergeometric distribution:

Grzybek and Kelih (2005) explored that grapheme frequencies of Slovean alphabets in text follow negative hypergeometric distribution. Grzybek (2007) has also examined the regularity of parameters of negative hypergeometric distribution for German letter frequencies and concluded that parameter behaviour follow clear rules as well. The negative hypergeometric distribution has been characterized by them in the form of the recurrence relation:

(10) $$F_r = \frac{(b+r-1)(n-r+1)}{r(a-b+n-r)} F_{r-1}$$

and the distribution has been given as:

(11) $$F_r = \frac{\binom{b+r-2}{r-1}\binom{a-b+n-r}{n-r+1}}{\binom{a+n-1}{n}},$$

with r =1, 2, ..…... $n+1$; where $a > b > 0$ are parameters and $n \in \{1,2,3\ldots\ldots\ldots\ldots\}$is the inventory size and F_r is the normalized frequency corresponding to the rank r.

Besides these distributions, we have tested the applicability of the following equation, proposed by Eftekhari (2006), for expressing the relation between Zipf's order of letters of English texts and their frequencies

(12) $$F_r = a / i^b ,$$

where a and b are parameters and i is the Zipf order of letter of rank r.

Application of the distributions discussed above in the form of equations (2), (3), (4), (7), (8), (9), (11) and (12) to the data of frequencies of different letters, ranks of letters and their Zipf orders for the text "Tanav se Mukti" has enabled us to calculate the theoretical values of the frequencies and the values of determination coefficient (R^2) for each model, defined in the following manner-

If a data set has n observed values y_i i = 1, 2, ……..n each of which has an associated modelled value f_i and if $\bar{y}$ is the mean of the observed data then the coefficient of determination is obtained by the formula-

(13) $$R^2 = 1 - \frac{SS_{err}}{SS_{tot}},$$

where

$SS_{tot} = \sum_i (y_i - \bar{y})^2$ is the total sum of squares, and $S_{err} = (y_i - f_i)^2$ is the sum of squared error or the residual sum of squares. R^2 is a statistic that gives information about the goodness of fit of a model.

The calculated values of determination coefficients of the fitted models for the rank frequency distribution of letters, for their occurrences in the text "Tanav se Mukti" and in the starting position of the words in the text, have been shown in the Table 4. We have used the software 'Datafit' for determination of parameters for all the equations except equation (3).

This software was producing numerical errors for equation (3) and as such we were left with no option but to apply the least square method for linear relation:

(14) $\log(F_r) = \log(a) - b\log(r+c) = A + B\log(r+c)$,

by first setting c = 0 and then increasing it by small steps until the goodness of fit stops improving.

Distribution	For occurrence of letters in the text		For occurrence of letters in the word's initials	
	Determination coefficient	Parameters	Determination coefficient	Parameters
Zipf	0.836	a = 5683.239, b = 0.685	0.932	a = 3051.361, b = 0.815
Zipf Mandelbrot	0.985	$a = 10^{7995144.491}$, $b = 1142162.974$ $c = 10^7$,	0.925	$a = 10^{5531903.571}$ $b = 816137.0218$ $c = 6\times10^6$,
Good	0.900	&	0.825	&
Geometric series	0.9891	a = 5007.781, b = 0.889	0.937	a = 2806.074, b = 0.843
Borodovsky & Gusein-Zade equation	0.866	&	0.772	&
Yule	0.9894	a = 5072.373, b = 0.0382, c = 0.8958	0.981	a = 2982.119, b = 0.459, c = 0.936
Negative hypergeometric	0.939	a = 3.186, b = 0.6872	0.976	a = 2.2503, b = 0.337
$F_r = a/i^b$, i is Zipf order of letter of rank r	0.978	$a = 2.113 \times 10^{-4}$, $b = -4.365$	0.892	$a = 3.663\times10^{-5}$, b = -4.673

Table 4 *Distributions, parameters and calculated determination coefficients for the rank frequency data of letters and words' initials of the text "Tanav se Mukti"*

The table depicts that the highest determination coefficient in both cases has been obtained for the Yule distribution (0.9894 and 0.981 respectively). Therefore on the basis of the results for the data of the example text (Tanav se Mukti), it can be assumed that the frequencies of occurrence of letters and word's initials in Hindi text follow Yule's distribution. Corresponding theoretical and empirical frequencies for this text for different ranks have

been shown in the Figure 3. In the next section we shall check whether this distribution is followed by the frequencies of occurrence in other texts also.

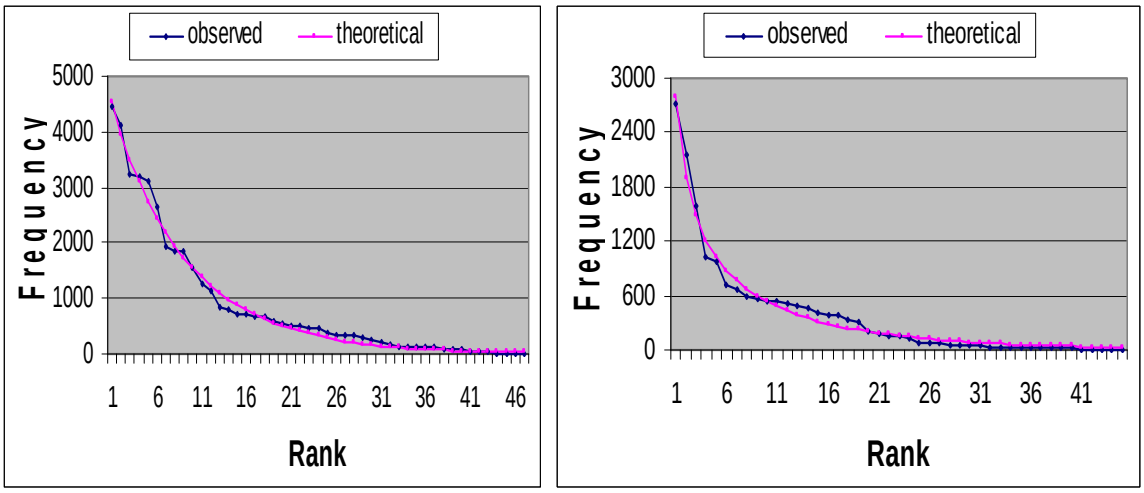

Figure 3. *Observed and theoretical (corresponding to Yule distribution) rank-frequency data for the occurrence of letters and words' initials. The first figure for letters while the second for initials of words.*

5. Validation of models for different texts

We have counted the frequencies of letters and have applied the distributions, discussed in the previous section, for various kinds of texts obtained from different sources[2] in the following forms:

1. Text composed of 5 stories. (Say, Text 1)
2. Text containing 29 short stories. (Text 2)
3. Text containing 62 articles. (Text 3)
4. Text containing 37 articles. (Text 4)
5. Poetic text containing 180 poems (text 5), and
6. Text formed as mixture of all the considered texts, from text 1 to text 5 above and the text Tanav se Mukti. (Text 6)

The calculated values of the determination coefficients and corresponding distributions (for the models or distributions for which the values of the determination coefficients are greater than 0.98 in case of letter frequencies in the texts and greater than 0.95 for the letter frequencies in the word's initials) and the Zipf's orders for the occurrence of letters in text and for occurrence as word's initials have been given in the following four tables:

[2] Sources of the texts have been given in the Appendix.

Text	Total counted letters(N_1)	Determination coefficient & distribution	Determination coefficient & distribution	Determination coefficient & distribution
Text1	18,162	0.990(Yule)	0.989(Geometric series)	0.981(Zipf Mandelbrot)
Text 2	22,375	0.996(Yule)	0.989(Geometric series)	0.988(Zipf Mandelbrot)
Text 3	70,622	0.992(Yule)	0.981(Negative hypergeometric)	0.980(Geometric series)
Text 4	82,803	0.992(Yule)	0.984(Geometric series)	0.983(Zipf Mandelbrot)
Text 5	90,908	0.991(Yule)	0.990(Geometric series)	&
Text 6	3,25,983	0.992(Yule)	0.990(Geometric series)	0.984(Zipf Mandelbrot)

Table 5 *Calculated values of determination coefficients (which are >0.98) for various fitted distributions corresponding to the rank frequency data of letters in texts*

Text	Zipf's order of letters for their instances in text
Text 1	ज्ञञऋअंक्षऐऊत्रढषणओइफघझठछईधऔशएटअआडखउभथचजवगदयबपलतमनसहरक
Text 2	ऋऊऐअंओज्ञढघक्षत्रझठफषछणईऔटइधडखअभचथशआएउगबजवदयलपमतसहनरक
Text 3	ज्ञऊअंझऐओढक्षघत्रठणषछथईधऔफखडउआइशचभअटएगबवजदपयलमतनहसरक
Text 4	ऋऊज्ञअंझढऐघठक्षओछत्रणषफईखडऔथधउइआटचशभअएगबजवदपयलमतसनहरक
Text 5	ञ ङऋज्ञऐऊअंओत्रक्षढऔषघफठइईणझछएटडउथआखशधअभचबगवदजयपलसतमनहकर
Text 6	ङञऋज्ञऊअंऐढओक्षघत्रठझषणछफईऔइडखधथटउआएशअचभगबजवदयपलमतसनहरक

Table 6 *Zipf's order of letters for their instance in various texts*

Text	Total counted words(N_2)	Determination coefficient & distribution	Determination coefficient & distribution	Determination coefficient & distribution
Text 1	8,298	0.978(Yule)	0.973(Negative hypergeometric)	&
Text 2	9,831	0.960(Yule)	0.953(Negative hypergeometric)	&
Text 3	29,623	0.975(Yule)	0.963(Negative hypergeometric)	0.962(Zipf)
Text 4	34,404	0.984(Yule)	0.979(Negative hypergeometric)	0.963(Zipf)
Text 5	40,418	0.987(Yule)	0.984(Negative hypergeometric)	0.968(Geometric series)
Text 6	1,39,373	0.985(Yule)	0.982(Negative hypergeometric)	&

Table 7 *Calculated values of determination coefficients (which are >0.95) for various fitted distributions corresponding to the rank frequency data of words' initials*

Text	Zipf's order of letters for their instances in the beginning of words in the texts
Text1	ज्ञत्रईढऋक्षटअंऊठडऐझछओधशएयफघइखचऔगअआभवरलथउतजदनबपहसमक
Text2	ऊऋत्रढओईज्ञटऐठडक्षझछअंधघफखचशऔएइयथभगअतआजवलरउदनबपहमसक
Text3	त्रईऊक्षढझठओअंछधटडऐघफथशखएचगऔयभआतइउरवअदलनबजपमसहक
Text4	त्रऋज्ञषझऊईक्षठढअंटडघओछधऐफशखथएचगऔइयभउतआलदवरअनबजपमसहक
Text5	षत्रऋईज्ञऐओऊठटढक्षडअंझघऔएछफइधशखथयचउगभआलवअरदतबनजपसहमक
Text6	षत्रऋज्ञऊईढठक्षओटअंझडऐघछधफखशएथऔइचयगभउआलतवरअदनबजपमसहक

Table 8 *Zipf's order of letters for their instance as the first letter of the words in various texts*

Tables 5 and 7 depict that the value of the determination coefficient in both cases is best for the Yule distribution for all the considered texts and its least value obtained for various texts is 0.989 in case of frequencies of letters and 0.96 in the case of words' initials, which is adequate to assume it as a model for the frequencies. From Table 6, letters - **क, त, न, म, र, स, ह** always take any position out of last seven positions in the corresponding Zipf's order list and similarly from the Table 8, **क, प, म, स, ह** get any one position out of the last five positions in the Zipf's order list for the occurrence of letters as words' initials for all the texts. Thus these seven letters and five word's initials can be assumed as seven letters of highest Zipf 's orders or seven most frequent letters and five word's initials of highest Zipf 's order or five most frequent word's initials respectively.

The variations in the values of the parameters of the Yule's distribution, as given by equation (9), for various considered texts have been depicted in the Figure 4,where the first figure represents the parameters of the model for letter frequencies and the second one for the letter frequencies in the word's initials. As the value of parameter '*a*' increases with the size of text, for comparison we have drawn the values of a/N_1 and a/N_2, where N_1 denotes total counted letters in the text and N_2 total counted word tokens in the texts.

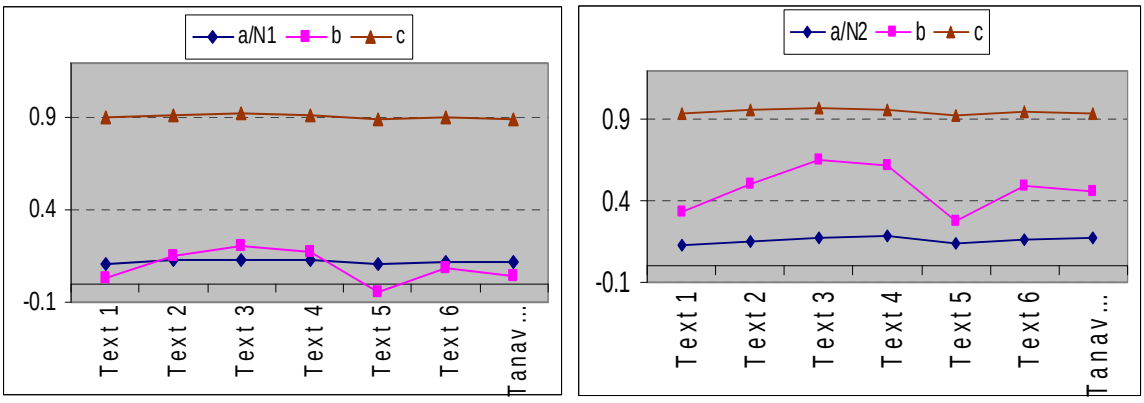

Figure 4 *Parameters of the Yule distribution for different texts for rank frequency distribution of letters and for distribution of the word initials*

From the figure, it is clear that although the parameters of the model for the distributions of letters in the text and in the starting position of words of the texts lay in slight different ranges but their manner of change for the two is almost the same.

As discussed above 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. The percentage of the total letters of the texts formed by these seven letters and percentage of the total words started by these five word's initials have been calculated and summarized in the following table:

	Tanav se mukti	Text 1	Text 2	Text 3	Text 4	Text 5	Text 6
% of total letters formed by 7 most frequent letters	55.2	53.1	53.4	53.9	54.98	52.6	53.9
% of total word-tokens for the words started by any one of the 5 most frequent word's initials	50.2	42.4	42.6	46.3	47.9	46.2	46.6

Table 9 *% of total letters and % of total word tokens of the texts formed by 7 most frequent letters and 5 most frequent word's initials*

Table 9 demonstrates that out of all used letters (excluding *matras*) of Hindi language alphabet in a text more than 50% are one of क, त, न, म, र, स, ह and out of all used words more than 40% words have their initial letter one of क, प, म, स, ह. In the next section we shall compare the entropies of these seven most frequent letters in texts and five most frequent word initials.

Here we have taken an initiative in order to improve the values of the determination coefficients for the rank frequency distribution of letters in the case of word's initials, for whom the value of determination coefficient for the texts text 1 and text 2 is not greater than 0.98, in the form of replacement of r by $(r \pm \delta)$ in the Yule equation $F_r = \frac{a}{r^b} c^r$, to take the form as

$$(15) \qquad F_r = \frac{a}{(r \pm \delta)^b} c^{r \pm \delta} = \frac{k}{(r \pm \delta)^{b_1}} c_1{}^r,$$

where c_1, b_1 and k are parameters and δ is a small number less than 1. After applying this relation we have calculated the values of the determination coefficients:

0.983 for text 1 by $(r - \delta)$ and δ=0.9 ,

0.986 for text 2 by $(r - \delta)$ and δ=.95,

0.99 for text 3 by $(r - \delta)$ and δ=.95 ,

0.989 for text 4 by $(r - \delta)$ and δ=.8 and

0.989 for text 6 by $(r-\delta)$ and δ =.9 and for the text 5 and the text "Tanav se Mukti" there is not any considerable improvement. Thus the model for the rank frequency distribution of letters for their occurrence in the beginning of words of a text can be formulated as-

$$F_r = \frac{a}{(r-\delta)^b} c^r, \quad (16)$$

c, *b* and *a* are parameters and δ is either zero or a number between 0 and 1.

6. Entropies of most frequent consonants

The entropy or the degree of uncertainty of the distribution of a random variable *x* is computed as:

$$H(x) = -\sum_{x \in X} p(x) \log_2 p(x), \quad (17)$$

where *p(x)* represents the probability of *x*. It measures the amount of information in a random variable and is invaluable in NLP, speech recognition, and computational linguistics. The entropy is bounded having lowest value equal to zero which implies that the entropy of a distribution that have no uncertainty at all while its greatest value is *log n*, where *n* is the total number of values that *x* can take which is possible in the case of the uniform distribution (or equiprobable distribution) over all possible values of *x*.

We have so far investigated that all five most frequent word initials and seven most frequent letters for a text are consonants and each consonant in a text can have three types of instances- as a consonant, as half consonant (or consonant followed by a *halant* '्') and as consonant followed by any *matra* (ा/ॉ, ि, ी, ु, ू, ृ, े, ै, ो, ौ, ं, ः). Thus each consonant can occur in 14 different ways in text, where the occurrences of क in कां, काँ (क + ा + ं, क + ा + ँ) etc. have been supposed as occurrence of क followed by *matra* 'ा'. In order to calculate the entropy for the pattern of occurrence of most frequent consonants, we have analyzed occurrences of each of seven most frequent letters in different texts and the pattern of the occurrence of five most frequent initials. The obtained pattern for the occurrence of seven most frequent letters in the text "Tanav se Mukti" has been shown in the table 10. If the random variable *X* for a particular consonant is taken as 'the *matra* following that consonant' then *X* = { none, ा/ॉ, ि, ी, ु, ू, ृ, े, ै, ो, ौ, ्, ं/ँ, ः }, where occurrence of a consonant as a half letter has been assumed as the occurrence followed by '्'. The values of entropies (for the example text , "Tanav se Mukti") obtained with the help of the equation (17) for क, र, ह, न, स, त, म from the Table 10 for the subsequently followed *matras* (discussed above) are respectively 2.695, 2.008, 2.669, 2.474, 2.607, 2.808, and 2.401. Similarly entropies of most frequent word initials **क, प, म, स, ह** are respectively 2.811, 1.992, 2.427, 2.666 and 2.114.

Pattern of occurrence \ Letters	क	र	ह	न	स	त	म
Whole letter	1688	2550	577	1278	1053	827	785
With ा/ॉ	710	285	223	650	225	709	362
With ि	243	133	159	164	59	404	99
With ी	326	102	414	106	124	163	22
With ु	82	90	94	127	141	111	49
With ू	8	16	13	4	12	8	27
With ृ	44	2	14	0	3	2	18
With े	535	241	80	516	418	338	444
With ै	7	3	1113	3	3	7	13
With ो	493	219	465	43	13	137	19
With ौ	19	0	3	10	15	1	3
With ् or as half letter	303	454	7	323	425	385	65
With ं/ँ	3	19	16	3	154	7	14
With ः	7	2	0	9	0	21	1
Total	4468	4116	3178	3236	2645	3120	1921

Table 10 *Frequencies for patterns of occurrences of most frequent letters for the consecutive matra in the text "Tanav se Mukti"*

The calculated values of entropies for all the considered texts have been shown in the following figure:

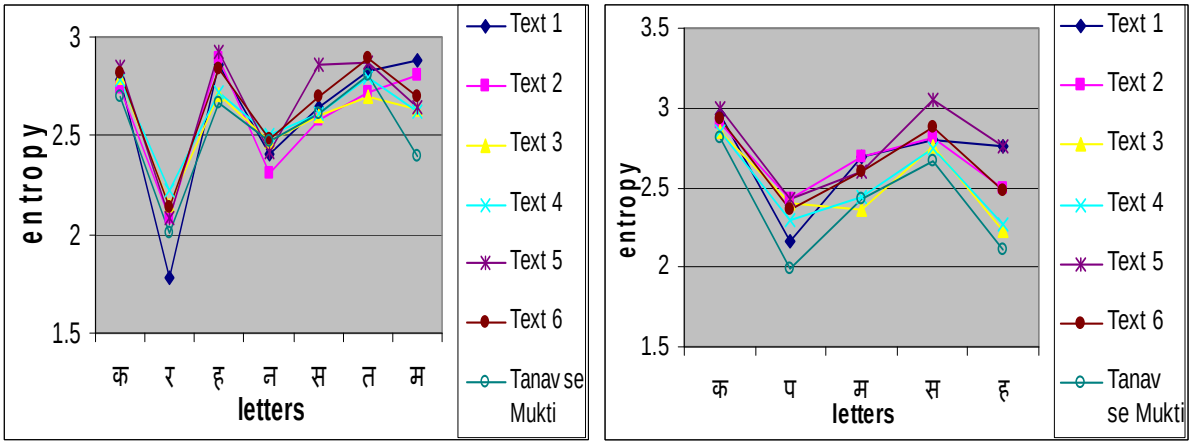

Figure 5. *Entropies of most frequent letters and most frequent word-initials for their pattern of occurrence*

On the basis of this study applied to seven texts, we can say that the letter र is much certain (less entropic) than 6 other most frequent letters in all the considered texts. It is followed by न except in the text "Tanav se Mukti" in which म has less entropy than न. In case of word's initials the occurrence pattern is that प is much certain in the texts- "Tanav se Mukti", Text formed by 5 stories, the text formed by 29 short stories, the text formed by poems and in the text formed as the mixture of all the texts. In the texts formed by 62 articles ह is less entropic followed by म and in the text formed by 37 articles ह is much certain followed by प and the entropies of these two are nearly equal thus प can be considered as much certain word's initial than other most frequent initials with regard to its pattern of occurrence.

7. Conclusions

We can conclude that:

- क, त, न, म, र, स, ह are seven most frequent letters of Hindi language and these form more than 50% of texts in terms of letters of Hindi language alphabet.
- क, प, म, स, ह are five most frequent word's initials and more than 40% of words of the text begin with any one of these letters.
- Frequencies of letters in a text and in combinations of texts follow Yule distribution $F_r = \frac{a}{r^b} c^r$, where a, b and c are parameters.
- Frequencies of letters as word's initials follow a Yule distribution $F_r = \frac{a}{r^b} c^r$, which can be more properly expressed by a distribution of type $F_r = \frac{a}{(r-\delta)^b} c^r$ where δ is either zero or lies between o and 1, and a, b and c are parameters.
- The letter र is much certain than other 6 most frequent letters and प is much certain word's initial than other 4 most frequent word's initials in the form of their patterns of occurrence.

Acknowledgement

The authors are grateful to the Council of Scientific & Industrial Research (CSIR), New Delhi, for providing financial assistance to carry out the research work on this interesting field of applied Mathematics in the form of a senior research fellowship to the first author.

References

ALTMANN, Gabriel. 2002. Zipfian Linguistics. In *Glottometrics 3,* pp. 19-26.

BELL, Timothy C., WITTEN, Ian H. 1988. Source models for natural language. <*http://hdl.handle.net/1880/46172* >.

BENGIO, Yoshua., DUCHARME, Rejean, VINCENT, Pascal and JAUVIN, Christian. 2003. A Neural Probabilistic Language Model. In *Machine Learning Research 3,* pp.1137-1155.

BENGIO, Yoshua. 2008. Neural Net Language Models. In *Scholarpedia,* 3(1), pp.3881.

EFTEKHARI, Ali. 2006. Fractal Geometry of Texts, an Initial Application to the Works of Shakespeare. In *Journal of Quantitative Linguistics,* 13(2-3), pp. 177-193.

GRZYBEK, Peter., KELIH, Emmerich. 2005. 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.

GRZYBEK, Peter. 2007. On the Systematic and System Based Study of Grapheme Frequencies: a Re-analysis of German Letter Frequencies. In *Glottometrics 15,* pp. 82-91.

HARRIS, Zelig. 1982. *A Grammar of English on Mathematical Principles*, John Wiley & Sons, New York, 1982.

"Letter Frequencies" *Nation Master- Encyclopedia.* Retrived 17 Oct. 2008 from
< http://www.nationmaster.com/encyclopedia/Letter-frequencies >

MANNING, Christopher. D., SCHUTZE, Hinrich. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, Mass. ; London, Eng.: MIT Press, 1999.

MORIN, F. and BENGIO, Yoshua. 2005. Hierarchical Probabilistic Neural Network Language Model. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics,* Barbados, Jan 6-8, 2005.

NARANAN, S. and BALASUBRAHMANYAN, V. K. 1998. Models of Power Law Relations in Linguistics and Information Science. In *Journal of Quantitative Linguistics,* 5, pp. 35-61.

PANDE, Hemlata. and DHAMI, H. S. 2009. Generation of a model for grapheme frequencies and its refinement and validation by group theoretic aspects. In *Journal of Quantitative Linguistics,* 16(4), pp. 307-326.

PARTEE, Barbara H., WALL, Robert. E., MEULEN, Alice ter. 1990. *Mathematical Methods to Linguistics*. Dordrecht: Kluwer, 1990.

SANDERSON, Robert. 2007. COMP527: Data Mining, Retrieved 19 Jan. 2008 from
<*www.csc.liv.ac.uk/~azaroth/courses/current/comp527/lectures/comp527-28.pdf* >

SOLSO, Robert L., KING, J. F. 1976. Frequency and versatility of letters in the English language. In *Bahavior research methods and instrumentation,* 8, pp. 283-286.

TAMBOVTSEV, Yuri, MARTINDALE, Colin. 2007. Phoneme Frequencies Follow a Yule Distribution. In *SKASE Journal of Theoretical Linguistics,* 4(2), pp. 1-11.

WANG, Wen, HARPER, Mary P. 2002. The SuperARV Language Model: Investigating the Effectiveness of Tightly Integrating Multiple Knowledge Sources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Philadelphia, July 2002, pp. 238-247.

Appendix

Sources of different texts

'Tanav se Mukti' by Shivanand	From the books by c-dac Noida. http://mobilelibrary.cdacnoida.com/Books/
5 stories: 'Cheelein' by Bhisham Sahni, 'Kabuliwala' by Rabindranath Tagore, 'Haar Ki Jeet' by Sudershan 'Vairagya' and 'Vardan' by Munshi Premchand	Cheelein Kabuliwala and Haar Ki Jeet from Bharat-Darshan, a Hindi literary magazine http://www.bharatdarshan.co.nz/ and Premchand's stories from the books by c-dac Noida. http://mobilelibrary.cdacnoida.com/Books/
128 articles from 30 Oct. 08 to 5 Dec. 08,as-62 articles(text 3) from the 'Sampadakeey' section 37 articles(text 4) from 'Nazaria' section and 29 short stories(text 2) from the 'Katha-Sagar' section	From 'Navbharat Times' http://navbharattimes.indiatimes.com/editorial/2279782.cms
180 poems: 55 poems from 'navkusum' section and 125 poems from 'yugvani' section	From 'Kaavyaalaya' The House of Hindi Poetry http://manaskriti.com/kaavyaalaya/yugvani.asp & http://manaskriti.com/kaavyaalaya/navkusum.asp

Hemlata Pande
Dept. of Mathematics,
University of Kumaun,
S. S. J. Campus Almora,
Almora (Uttarakhand)-263601 INDIA
E-mail: pande_hemlata1011@yahoo.com

H.S. Dhami
Dept. of Mathematics,
University of Kumaun,
S. S. J. Campus Almora,
Almora (Uttarakhand)-263601 INDIA
E-mail: drhsdhami@yahoo.com

In *SKASE Journal of Theoretical Linguistics* [online]. 2010, vol. 7, no. 2 [cit. 2010-06-30]. Available on web page <http://www.skase.sk/Volumes/JTL16/pdf_doc/02.pdf>. ISSN 1339-782X.