Mathematical Modelling of the Pattern of Occurrence of Words in Different Corpora of the Hindi Language

Hemlata Pande & H.S. Dhami

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


To cite this article: Hemlata Pande & H.S. Dhami (2013): Mathematical Modelling of the Pattern of Occurrence of Words in Different Corpora of the Hindi Language*, Journal of Quantitative Linguistics, 20:1, 1-12

To link to this article: http://dx.doi.org/10.1080/09296174.2012.754596
whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
Mathematical Modelling of the Pattern of Occurrence of Words in Different Corpora of the Hindi Language*

Hemlata Pande and H.S. Dhami
Department of Mathematics, Kumaun University, S. S. J. Campus Almora, Uttarakhand, India

ABSTRACT

The incidence of different components of language in natural language texts are not arbitrarily organized but tend to obey particular laws which enable us to explain characteristic features of human language. The present paper is an attempt to analyse and model the pattern of occurrence of words in the Hindi language. Various kinds of corpora have been selected from different sources for the study, and the occurrence of words in these corpora has been observed for a variety of properties such as: frequencies, vocabulary measures, and pattern of initials of words relative to the subsequent matra.

1. INTRODUCTION

The multitude of linguistic utterances around us in the form of conversation, writing and other media can be placed into a theoretical framework that both describes linguistic usage and also predicts the patterns of occurrence of various linguistic features. Bharati et al. (2002) have presented statistical analyses based on measures like unigram frequencies, bigram frequencies, syllable frequencies and so on of 10 machine-readable Indian language corpora. In the context of analysis and modelling for the Hindi language, we also cite the work of Jayaram and Vidya (2006) for discussion of word length frequency distribution in Hindi and some other Indian languages. Jayaram and Vidya (2008) have utilized Zipf’s law for the word frequency distribution of Hindi and three other Indian languages. In earlier work we

*Address correspondence to: Hemlata Pande, Department of Mathematics, Kumaun University, S. S. J. Campus Almora, Almora-263601, Uttarakhand, India. Email: hlpande@rediffmail.com
(Pande & Dhami, 2010) have analysed and modelled the pattern of occurrence of various letters in Hindi texts and in initial position of the words of those texts.

Murthy and Kumar (2006) have noted that Devanagari script is used to write various languages such as Sanskrit, Hindi, Marathi, Konkani and Sindhi, and that mere script identification does not serve the purpose of being “able to identify language irrespective of the script or font being used”. Pingali and Varma (2007) have observed that “script based recognition for Unicode character sets works to an extent, but still ambiguities might exist”, and give examples which show that for some of the Indian languages written in the Devanagari the script-based heuristics may not work. In their opinion a better set of heuristics is required and they have conjectured that statistical language identification techniques can help for ambiguous scripts. Sanderson (2007) has pointed out the importance of the easily determinable characteristic “distribution of letters in a text” in the process of language determination.

In the present paper we discuss the following features associated with the component “word”: word frequency distribution, vocabulary measures, pattern of occurrence of words and the entropy measure of different corpora of the Hindi language. The models for distributions of words in corpora are discussed in Section 3, and Section 4 deals with the comparison of various vocabulary-richness measures. In Section 5 we examine the pattern of occurrence of words in various corpora and determine the entropy of the selected corpora. A comparison of the entropies of the various corpora is also given in this section.

2. DATA SOURCES AND THE DATA FOR THE PRESENT STUDY

We have counted the frequencies of words in the corpus Tanav se Mukti and in five other corpora obtained from various sources:

1. Corpus *Tanav se Mukti* by the author Shivanand (C1).
2. Corpus formed by composition of five stories (C2).
3. Corpus containing 29 short stories (C3).
4. Corpus containing 62 articles (C4).
5. Corpus containing 37 articles (C5).
6. Poetic Corpus containing 180 poems (C6).

Sources have been shown in the Appendix.
We have determined the rank frequency data, number of word tokens and word types (different words used in the corpus, or the vocabulary of the corpus) and number of words occurring once, twice, thrice and so on. Data for the words starting with vowels and with consonants without any *matra*, a Hindi word which designates the vowel mark following the consonant, and with the consonants followed by different *matras* has also been derived. We have assigned the ranks to the words in accordance to their descending order of frequencies as is done by the Zipf distribution after obtaining the word frequency statistics.

### 3. WORD FREQUENCY DISTRIBUTION

In our earlier work (Pande & Dhami, 2010) we discussed various models such as the Zipf, Zipf Mandelbrot, Good, geometric series, Yule, and negative hypergeometric distributions together with the Borodovsky and Gusein Zad and Eftekhari (2006) equations. All these models have been used by researchers to determine frequencies from ranks for different language components. These models (excluding the first two) are used for letters, grapheme and phoneme frequencies of language. The parameter-free models, the model of Good and the Borodovsky and Gusein Zad equation as well as the parametric equations of Eftekhari and negative hypergeometric distribution depend on the highest rank for the data. These do not give good results in cases where the value of the highest rank is large, as in the present case of frequency distribution of words.

We have applied the distributions of Zipf, Zipf Mandelbrot, geometric-series and of Yule to the frequencies and ranks of words for the above-cited Hindi corpora.

Zipf’s law (Zipf, 1949) has been defined in the work of Manning and Schütze (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 this law ‘if the frequency of occurrence of an event is ranked in descending order of frequency, then

\[ F_r \propto \frac{1}{r}, F_r = c, \]

where \( F_r \) is the frequency of event of \( r \)th rank and \( c \) is a constant’. The general form of this formula was given by him in later works as:
where $a$ and $b$ are text parameters.

A modification to the Zipf formula was introduced by Mandelbrot (Mandelbrot, 1966), as cited in Manning and Schütze (1999), by including an additional parameter $c$ in the following formula of rank frequency distribution:

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

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

**Geometric series equation.** General form of the geometric series equation is given by:

$$F_r = ab^r.$$  \hspace{1cm} (3)

The equation for the Yule distribution is:

$$F_r = \frac{a}{r^b} c^r,$$  \hspace{1cm} (4)

where $a$, $b$, $c$ are parameters.

For the determination of the parameters we have used the *datafit* software. We have calculated the theoretical values of the frequencies corresponding to above four distributions and the values of determination coefficient ($R^2$) for these models by the formula

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}}$$  \hspace{1cm} (5)

where the data set has $n$ observed values $y_i$, $i = 1, 2, \ldots, n$ each of which has an associated modelled value $f_i$ and $\bar{y}$ is the mean of the observed data. $SS_{tot}$ and $SS_{err}$ represent the total sum of squares and the sum of squared error (or the residual sum of squares).

$$SS_{tot} = \sum_i \{y_i - \bar{y}\}^2 \text{ and } SS_{err} = \sum_i \{y_i - f_i\}^2$$  \hspace{1cm} (6)

$R^2$ is a statistic that gives information about the goodness of fit of a model for linear models. We have made an attempt to extend its domain to non-linear models which has resulted the investigation of results in the present work. However, in the case of non-linear models the value of above
definition of $R^2$ may be outside the interval $[0, 1]$. We have also compared the values of the squared correlation $R_1^2$ (pseudo-$R^2$) between the observed values $y_i$ and modelled values $f_i$ of the frequencies. For the comparison of models for various corpora we have applied the models for the relative frequencies of various words. The values of $R^2$ and $R_1^2$ for the studied corpora are greatest corresponding to the Zipf Mandelbrot law. For the corpus C1, the values of the $R^2$ and $R_1^2$ are shown in Table 1.

The calculated values of the determination coefficients for the Zipf Mandelbrot distribution are shown in Figure 1.

The figure demonstrates that the values of the determination coefficient are very close to 1 for all the corpora. Similarly the values of the squared correlation between predicted and observed frequencies for these are in the range 0.9742–0.9965. Therefore Zipf Mandelbrot law has been taken as the model for the rank frequency distribution of words in Hindi corpora. The values of parameters $a$, $c$, and $b$ for this model for various corpora are given in Table 2.

Since Table 2 shows that the values of the parameter $b$ are very close to 1, we have tested the same, that is, the Zipf Mandelbrot distribution, by

<table>
<thead>
<tr>
<th>Model</th>
<th>$R^2$</th>
<th>$R_1^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zipf distribution</td>
<td>0.9509</td>
<td>0.9573</td>
</tr>
<tr>
<td>Zipf Mandelbrot distribution</td>
<td>0.9742</td>
<td>0.9742</td>
</tr>
<tr>
<td>Geometric-series distribution</td>
<td>0.8830</td>
<td>0.8991</td>
</tr>
<tr>
<td>Yule distribution</td>
<td>0.9699</td>
<td>0.9737</td>
</tr>
</tbody>
</table>

Fig. 1. Values of the determination coefficient for different corpora corresponding to the Zipf Mandelbrot distribution.
taking $b = 1$. This has resulted in the reduction of one parameter from the equation and thus the distribution takes the form:

$$F_r = \frac{a}{(r+c)} = \frac{1}{(p+qr)} = \frac{1}{(p+qr)};$$

where $p$ and $q$ are parameters.

We have applied this model for the rank frequency data of the various corpora. The calculated values of the determination coefficient corresponding to this model are listed in Table 3. In the present case the values of the squared correlation between predicted and observed frequencies are in the range $0.9739–0.9962$.

Results of the determination coefficient in Table 3 are very close to 1, and the values in the case of this model are greater than the corresponding values determined for other models, discussed earlier. Thus we can conclude that when only two parameters are considered the model for the rank frequency distribution of words for the Hindi language corpora should be

<table>
<thead>
<tr>
<th>Corpus</th>
<th>$a$</th>
<th>$c$</th>
<th>$b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>0.1205289</td>
<td>1.9514604</td>
<td>0.9617396</td>
</tr>
<tr>
<td>C2</td>
<td>0.2911290</td>
<td>8.9221350</td>
<td>1.0994770</td>
</tr>
<tr>
<td>C3</td>
<td>0.3032830</td>
<td>8.0375200</td>
<td>1.1173380</td>
</tr>
<tr>
<td>C4</td>
<td>0.2486375</td>
<td>3.6550430</td>
<td>1.1606351</td>
</tr>
<tr>
<td>C5</td>
<td>0.3056000</td>
<td>4.4465500</td>
<td>1.2006200</td>
</tr>
<tr>
<td>C6</td>
<td>0.1282817</td>
<td>3.8978680</td>
<td>0.9970879</td>
</tr>
</tbody>
</table>

Table 2. Values of the parameters for Zipf Mandelbrot distribution for different corpora.

<table>
<thead>
<tr>
<th>Determination coefficient for Zipf Mandelbrot rule</th>
<th>Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.973854</td>
<td>C1</td>
</tr>
<tr>
<td>0.994726</td>
<td>C2</td>
</tr>
<tr>
<td>0.993324</td>
<td>C3</td>
</tr>
<tr>
<td>0.982873</td>
<td>C4</td>
</tr>
<tr>
<td>0.98592</td>
<td>C5</td>
</tr>
<tr>
<td>0.996161</td>
<td>C6</td>
</tr>
</tbody>
</table>
taken as the model given by Equation (7) rather than the Zipf law as presented by Jayaram and Vidya in 2008.

\[ F_r = \frac{1}{p + qr} \]  

(7)

4. VOCABULARY RICHNESS MEASURES

Panas (2007) pointed out that “the index of lexical richness is a fundamental parameter describing the word frequency structure of a text”, and noted that Wimmer and Altmann (1999), in a review article on vocabulary richness, discussed four different approaches to study the indices of vocabulary richness:

- by measuring the indices,
- by curve representation,
- by empirical and theoretical distributions,
- by stochastic processes.

We have used the five vocabulary richness measures propounded by Yule, Honore, Brunet, Sichel, and Simpson as cited in the research paper of Stamatatos et al. (2000). These measures are given in the following form

\[
\begin{align*}
K &= \frac{10^4}{N^3} \sum_{i=1}^{\infty} \frac{i^2 V_i - N}{N}, \\
R &= \frac{100 \log N}{1 - \frac{V_1}{N}}, \\
W &= N^{V^{-2}}, \\
S &= \frac{V_2}{V}, \\
D &= \sum_{i=1}^{V} \frac{V_i^{i(i-1)}}{N^{(i-1)}},
\end{align*}
\]

(8)

where \( V \) is the vocabulary of a text or corpus, \( N \) is the total number of word tokens in the text, \( V_1 \) (called hapax legomena) and \( V_2 \) (hapax dislegomena) are the number of words occurring once and twice respectively in the text and \( V_i \) is the number of words occurring exactly \( i \) times. Calculated values of Brunet measure \( \times 10^{-2} \) and of Sichel measure, Yule measure \( \times 10^{-2} \), Honore measure \( \times 10^{-3} \), Simpson measure \( \times 10^2 \) are shown in Figure 2.

These two figures show that the values of Yule’s measure and Simpson’s measure are lower for C6, which is the corpus formed by the composition of poems. For C1, C4 and C5 the values of these measures are relatively higher. These corpora are generated from the non-story type text(s): corpus C1 is Tanav se Mukti written by Shivanand, C2 is the corpus formed by the
composition of the 62 articles from the Sampadakeey and corpus C3 is the compendium of 37 articles from the Nazaria section of the Navbharat Times. For the corpora C2 and C3 the values of these measures are higher than the values for the corpus C6 and lower than the values for the C1, C4 and C5. C2 is collection of short stories and C3 is a composition of five stories. Thus it can be concluded that, for the poetic corpus, the values of the measures of Yule and Simpson are lower, for stories these are higher than the corresponding values for the poetic corpus and much higher for the corpora of the Lekh (essay) kind. The value of the Brunet vocabulary measure for all the corpora is nearly equal (in the range from 0.1 to 0.12) and for the corpus C6, it is relatively lower than the value for others (C1-C5). Thus it can be concluded that the verse corpus has a relatively lower Brunet measure than that of non-verse cases.

5. PATTERN AND ENTROPY OF WORDS WITH DIFFERENT INITIALS

The pattern of initials of words in corpora was analysed for the proportion of words initiated with vowels and with consonants. Obtained values of the proportions have been shown in the Table 4.

The table shows that, except in C6, which is the collection of poems, in all the considered corpora 85% to 87% of words start with the consonants and 13% to 15% with vowels. In C6 these percentages are close to 90% and 10% respectively. Thus it can be said that in non-verse Hindi language corpora 85% to 87% of words start with consonants and in case of verse more words start with consonants than in non-verse corpora.
The entropy or degree of uncertainty of the distribution of a random variable $x$ is computed as determined in our earlier work, Pande and Dhami (2010), by the formula:

$$H(x) = -\sum_{x\in\mathcal{X}} p(x) \log_2(x).$$  \hfill (9)

$p(x)$ represents the probability of $x$. Each consonant in a text can have three types of instances as a consonant, as a half consonant (or consonant followed by a halant “◌”) and as a 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 interpreted as occurrence of क followed by matra “◌ः”. Results of frequencies of consonants with different matra for the example text (C1) are shown in Table 5.

After calculating the data for different corpora we have determined the values of the entropy for various corpora with the help of Equation (9), and these are shown in Table 6.

Table 4. Proportion of words started with consonants and vowels in various corpora.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Proportion of words starting with consonants</th>
<th>Proportion of words starting with vowels</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>85.75%</td>
<td>14.25%</td>
</tr>
<tr>
<td>C2</td>
<td>86.87%</td>
<td>13.13%</td>
</tr>
<tr>
<td>C3</td>
<td>85.22%</td>
<td>14.78%</td>
</tr>
<tr>
<td>C4</td>
<td>85.51%</td>
<td>14.49%</td>
</tr>
<tr>
<td>C5</td>
<td>85.56%</td>
<td>14.44%</td>
</tr>
<tr>
<td>C6</td>
<td>89.63%</td>
<td>10.37%</td>
</tr>
</tbody>
</table>

Table 5. Frequencies of the consonants followed by different matra in the corpus Tanav se Mukti.

<table>
<thead>
<tr>
<th>Consecutive matra of consonants</th>
<th>None (occurrence as whole letter)</th>
<th>◌/◌ः</th>
<th>◌ः</th>
<th>◌ँ</th>
<th>◌ः</th>
<th>◌ः</th>
<th>◌ः</th>
<th>◌ः</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>4381</td>
<td>1580</td>
<td>1074</td>
<td>953</td>
<td>670</td>
<td>284</td>
<td>103</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Consecutive matra of consonants</th>
<th>◌</th>
<th>◌</th>
<th>◌</th>
<th>◌</th>
<th>◌</th>
<th>◌</th>
<th>◌</th>
<th>◌</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>1575</td>
<td>1232</td>
<td>1384</td>
<td>60</td>
<td>895</td>
<td>211</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
The table shows that the values of the entropy for all the considered corpora, whether these are verse or non verse, is nearly 3.1 and is in the range from 3.05 to 3.15.

6. CONCLUSION

We can conclude that

- The model for governing the rank frequency distribution in Hindi language corpora is the law of Zipf Mandelbrot:

\[ F_r = \frac{a}{(r + c)^b}, \]

where \( a, b, c \) are parameters and \( F_r \) is the frequency of word of rank \( r \).

- In the case of only two parameters, the appropriate model can be expressed as:

\[ F_r = \frac{1}{p + qr}, \]

where \( p \) and \( q \) are parameters.

- For the poetic corpus the values of the measures of Yule and Simpson are lower, for stories these are higher than the corresponding values for poetic corpus and for the corpus of Lekh (essay) type these values are much higher. The verse corpus has relatively lower Brunet measure than that of non-verse case.

- Out of all the words in prose corpora 85–87% words initiate with the consonants. In the case of verse this proportion of words on average is larger than the non-verse case.
• The entropy for the pattern of the consonants occurring in the initial position of words of the corpora is within the specific range 3.05–3.15, which gives an important and easily determinable property of Hindi corpora. This fact can be used for the language identification task to discriminate Hindi language from some other languages written in the Devanagari script.

ACKNOWLEDGEMENT

The authors are grateful to the University Grants Commission (UGC), New Delhi, India, for providing financial assistance in the form of a Post-doctoral fellowship to the first author. The research has been sponsored by the UGC under the ‘UGC Dr. D. S. Kothari Post Doctoral fellowship scheme’.

REFERENCES


APPENDIX

Sources of different texts.

(These texts have also been used for the study by Pande and Dhami, 2010).

<table>
<thead>
<tr>
<th>Text</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>128 articles from 30 Oct. 08 to 5 Dec. 08, as 62 articles (Corpus 3) from the ‘Sampadakeey’ section 37 articles (Corpus 4) from ‘Nazaria’ section and 29 short stories (Corpus 2) from the ‘Katha-Sagar’ section</td>
<td>From ‘Navbharat Times’ <a href="http://navbharattimes.indiatimes.com/editorial/2279782.cms">http://navbharattimes.indiatimes.com/editorial/2279782.cms</a></td>
</tr>
</tbody>
</table>