# Information theoretic models in statistical linguistics—Part I: A model for word frequencies

*S. Naranan and V. K. Balasubrahmanyan*

*In the structure of language, certain statistical laws occur with striking regularity. The distribution of word frequencies follows Zipf's law, which is nearly universal—for all languages, authors and types of literature. String lengths—lengths of words, phrases, sentences, etc.—are also known to follow a lognormal distribution. In this article we have used a model based on Shannon's Information Theory to obtain a word frequency distribution*

$$W(k) = C e^{-\mu/k} k^{-\gamma},$$

*where $W(k)$ is the number of distinct words occurring exactly $k$ times in a discourse and $C$, $\mu$, $\gamma$ are constants. This is a power law (Zipf's law) modified at low frequencies (small $k$). For most discourses $\gamma = 2.0$ and $0 \leq \mu \leq 1.3$. This model differs from an earlier model by Mandelbrot in many essential ways, and applies to written as well as spoken discourses.*

## 1. Introduction

Language has been recognized as a unique, species-specific achievement of human beings and has been the subject of intense study and speculations by great minds such as Panini[1], Patanjali[2], Bhartruhari[3], Plato and Aristotle[4]. Modern linguistics is generally supposed to have had its beginning in 1786 when William Jones[5], East India Company judge in Bengal, pointed out that the close relationship between common words in languages as disparate as Latin, Greek and Sanskrit may be due to their being descended from a common origin. From these humble beginnings and speculations, the idea of language families (Indo–European, Dravidian, Finno–Ugrian, North American–Indian etc.) has been pursued to the point where multidisciplinary studies by linguists, DNA sequence analysts and archaeologists seem to suggest that all existing languages may have had a common origin in the remote past[6]. The migratory patterns of early humans and their relationship to existing language families[7] have created a lively interest as well as controversy in recent times.

The dual role of language as a codifier of thought and a means of communication has been the object of debate among linguists and philosophers such as Chomsky, Wittgenstein, Fregé and others. This controversy is ongoing with each side being able to present powerful arguments to refute the opposite side, though unable to establish its own premise effectively[8,9].

Chomsky[10,11], whose contributions to linguistics have greatly influenced the style of modern linguistics research, has focused on the innate biological built-in features that enable a child to pick up the complex rules of syntax of any language to which exposure is made during his/her learning phase in early childhood. Chomsky's neglect of 'performance' in real life linguistic communication has been criticized (Narasimhan[12,13]). Performance in linguistic interactions lends itself to the development of computational models, whereas Chomsky's concepts, by their neglect of evolutionary aspects, language acquisition and behaviour, lead only to an asymptotic theory with intangible observables.

Before the Chomsky wave in linguistics, the formulation by Shannon and Weaver[14] of 'Information Theory' led to a great deal of work on the relevance of Shannon's concepts of information and entropy to coding problems in natural languages[15,16]. Even earlier work on word counting by Zipf[17–19], Dewey[20] and others pointed to the relative stability of the frequency distribution of words in discourses. Their empirical work showed that the distribution of words in a discourse is a power law which appeared to be relatively independent of the length of discourse, author

S. Naranan lives at 402 Meghnad, TIFR Housing Complex, Homi Bhabha Road, Bombay 400 005, India and V. K. Balasubrahmanyan lives at "Shanmukhapriya", Raj Mahal Vilas Extension II Stage, Bangalore 560 094, India

or language. This observation has been called Zipf's law.

In this article, we critically review the earlier work of Mandelbrot on the application of information theory to derive Zipf's law. Unlike Mandelbrot, who uses the alphabet as the primary symbol, we adopt the word as the primary symbol. Besides the conventional entropy (Shannon's information), we define another entropy—'degenerate entropy'. This, together with a constraint on the finiteness of vocabulary, plays a crucial role in deducing an optimal word frequency distribution. This is a modified power law function, which is in good agreement with data on word frequencies. Since our model is based on the word as the primary symbol it is valid for written as well as spoken discourse—which is a breakthrough in what has been termed as the 'tyranny of the written word'. The model is described in the next section.

## 2. A model for word frequency distribution based on information theory

In this section we use the principles of Shannon's Information Theory[14, 15] to deduce the word frequency distribution in a discourse: the number of words $W(k)$ that occur exactly $k$ times.

After a brief review of Shannon's Information Theory in section 2.1, we derive expressions for the entropy and cost of encoding (in bits) of a given 'message' or discourse in section 2.2. In section 2.3 an extremum principle is used to derive the optimal word frequency distribution—one that maximizes a quantity called 'degenerate entropy' of the discourse under given constraints on the length of the discourse, the size of the vocabulary used and the entropy of the discourse. The 'goodness of fit' of the distribution so derived, to the actual data on word frequencies is examined[21]. Our model is compared with Mandelbrot's model based on Information Theory and Simon's[22] stochastic model in section 2.4. The section concludes with a discussion and general comments about our model for word frequencies.

### 2.1 *Shannon's Information Theory*

The mathematical theory of communication[14, 15] treats the generation of a message—a string of symbols—as an unpredictable process to some degree. The process is assumed ergodic; in other words, the time average of relative frequencies of different symbols in a single sequence is the same as the relative frequencies averaged over an ensemble of similar sequences. This ensures that any one message is a representative sample of all possible messages generated by the source. The assumption of *ergodicity* of the source, while making the mathematical theory simple, is also realistic.

The entropy of a message is a measure of choice for the source of the message and also a measure of the uncertainty of the recipient as to which message will be received, which is resolved on the receipt of the message. According to Shannon, the measure of 'uncertainty' is also a measure of 'quantity of information'. If the message is one among $M$ equally probable messages, the entropy $H$ is $\log_2 M$. If the $n$ different symbols in the message occur with different probabilities $p_i$ $(i=1,2,\ldots,n)$, then

$$H = -\Sigma\, p_i \log p_i \quad \text{bits per symbol.} \qquad (2.1)$$

When all the different symbols occur independent of each other with equal probability $p_i = 1/n$ $(i=1,2,\ldots,n)$,

$$H = n(1/n)\log_2 (1/n) = \log_2 n \text{ bits per symbol} \qquad (2.2)$$

and this is the maximum possible value of $H$ $(H_{max})$.

One way of interpreting the entropy $H$ is that for a given $H$ there are just $2^{NH}$ probable messages, each with $N$ symbols, and the rest of the messages will have vanishing probabilities. When $H = H_{max} = \log_2 n$ the number of possible messages is $2^{NH_{max}} = 2^{N \log n} = n^N$, which is the maximum possible number of messages of $N$ symbols.

In communication theory, the entropy $H$ of a message source in bits per symbol is also the *average number of binary digits per symbol necessary to encode the message*. An efficient coding scheme, for any given set of probabilities $p_i$ $(i=1,2,\ldots,n)$, is given by Huffman[23]. The Huffman code can be proved to be the most efficient code and never requires more than one bit per symbol in excess of $H$.

Equation (2.1) gives a value of entropy higher than the true value when the probability of a symbol depends on the preceding symbol(s). For example, it is well known that in the English language, the probabilities of occurrence of different alphabets are not independent of each other. If the 27 characters (26 letters and space) are encoded assuming that they will occur with the same probability, then $\log_2 27$ or 4.76 bits per character are required. If the different probabilities of occurrence are taken into account—still assuming them to be independent, (e.g. $e$ occurs with probability 0.131 and $z$ with probability 0.00077)—we require 4.03 bits per character. If the primary 'symbol' is not the letter but the *word*, then encoding word by word taking into account the probabilities of word frequencies given by Zipf's[17, 18] rank frequency relation requires 9.72 bits per word or 1.77 bits per character (since the average number of characters per word is 4.5 letters plus 1 space). This is lower than the original estimate of Shannon (11.8 bits per word), which contained an arithmetical slip[24]. The decrease in the number of bits required, is a consequence of the fact that all random (independent) choices of letters do not constitute real words. The correlated probabilities of

occurrence of characters provide clues to the recipient which is termed *redundancy*: an occasional replacement of one character by another by error does not destroy the message because the correct character can be guessed from other characters transmitted correctly. While it is best to remove this redundancy for efficient coding, such a coding scheme is extremely vulnerable to error (caused by noise in transmission): an error will then lead not to a distorted message, but a misleading or false message.

### 2.2 *Entropy and cost of encoding a message/discourse*

The fact that human discourse is both highly structured and unpredictable makes it a very apt candidate for application of the principles of Shannon's Information Theory.

We take the word as the primary symbol. A discourse is a string of $N$ 'word-tokens' (or simply words), with $V$ different or distinct words called 'word-types'. Words that differ phonetically are usually taken as different word-types. We denote the number of word-types occurring exactly $k$ times in a discourse as $W(k)$. The range of $k$ is 1 to a maximum value $k_m$. We define a $k$-word type as a word-type occurring $k$ times. The total number of $k$-word tokens (or simply $k$-words) $n_k$ is given by

$$n_k = k\,W(k)\ (k=1,2,\ldots,k_m),\ \Sigma\, n_k = \Sigma\, k\,W(k) = N \tag{2.3}$$

and

$$\Sigma\, W(k) = V. \tag{2.4}$$

$\Sigma$ represents summation over $1 \le k \le k_m$ hereafter, and every equation involving $W(k)$ actually represents $k_m$ equations with $k=1,2,\ldots,k_m$.

Based on equation (2.3) we can define two different entropies: one for a message in which the $W(k)$ different $k$-words are distinguishable and the other in which they are regarded as equivalent. The former is the standard conventional entropy (Shannon's information); the latter will be called 'degenerate entropy', borrowing a term from statistical physics.

The number of different possible discourses of $N$ words is given by the multinomial distribution

$$M_N = \frac{N!}{(k_1!)^{W(k_1)}\,(k_2!)^{W(k_2)}\ldots(k_m!)^{W(k_m)}}. \tag{2.5}$$

The entropy of the discourse is defined as

$$H_N = K\ \ln\ M_N, \tag{2.6}$$

where $K$ is a numerical scale factor. Using Stirling's approximation for $\ln x!$ when $x$ is large

$$\ln\ x! = x\ \ln\ x - x \tag{2.7}$$

$$\ln\ M_N = \ln\ N! - \Sigma\, W(k)\ \ln\ k!$$

$$= N\ \ln\ N - \Sigma\, k\,W(k)\ \ln\ k. \tag{2.8}$$

From equation (2.6)

$$H_N = K\ [N\ \ln\ N - \Sigma\, k\,W(k)\ \ln\ k]. \tag{2.9}$$

To express $H_N$ in binary digits—bits—we take $K = \log_2 e = 1.443$.

The entropy $H_N$ is a maximum ($= K\,N\ \ln\ N$) when all the $N$ words of the discourse are different [$(W(1)=N$, $W(k)=0$ for all $k \ne 1$]. As already stated $H = H_N/N$ is also the average number of bits per symbol necessary to encode the message.

If we consider that the word-types occurring the same number of times in a discourse are equivalent and hence mutually interchangeable, then the number of different possible discourses is given by

$$M_{DN} = \frac{N!}{n_1!\ n_2!\ n_3!\ldots n_k!}\ , \tag{2.10}$$

in which $n_k$ is the number of $k$-word tokens. The entropy of this discourse, which we call 'degenerate entropy', is

$$H_{DN} = K\ \ln\ M_{DN} \tag{2.11}$$

$$H_{DN} = K\ [N\ \ln\ N - \Sigma\, n_k\ \ln\ n_k]. \tag{2.12}$$

Using equation (2.3)

$$H_{DN} = K\ [N\ \ln\ N - \Sigma\, k\,W(k)\ \ln\ k\,W(k)]. \tag{2.13}$$

$H_{DN}$ is minimum ($=0$) when all the $V$ word-types occur with the same frequency [$W(k)=V$ for $k=N/V$ and $W(k)=0$ for all other $k$].

We note from equations (2.9) and (2.13)

$$H_N - H_{DN} = K\,\Sigma\, k\,W(k)\ \ln\ W(k) > 0. \tag{2.14}$$

### 2.3 *Derivation of the optimal word frequency distribution*

From the previous section we have

$$\Sigma\, W'(k) = V'$$

$$\Sigma\, k\,W'(k) = N$$

$$N\ \ln\ N - \Sigma\, k\,W'(k)\ \ln\ k = H_N \tag{2.15}$$

$$N\ \ln\ N - \Sigma\, k\,W'(k)\ \ln\ k\,W'(k) = H_{DN}.$$

We have set the scale factor $K=1$ for $H_N$ and $H_{DN}$ for the present.

The first two equations represent the constraints on the sizes of the vocabulary and discourse; the third ($H_N$) is the entropy or information of the discourse. The 'degenerate entropy', $H_{DN}$, is a new concept introduced here, which will play a crucial role in deducing the optimal word frequency distribution.

The entropy of a discourse is also the 'cost'—in number of bits—in encoding the discourse. The only difference between $H_N$ and $H_{DN}$, from an information theoretic point of view, is that whereas every different $k$-word is assigned a different code in $H_N$, all the $k$-words [$W(k)$ of them] are assigned a single code in $H_{DN}$. Semantically the $W(k)$ different $k$-words are different, whereas for the purposes of word frequency distribution —essentially a word-counting process—all the $W(k)$ words are equivalent. For instance, the Huffman coding algorithm will assign a *different* binary code (of the same length) to each $k$-word, but the $W(k)$ different binary codes will be arbitrarily assigned to the $W(k)$ different $k$-words, since all $k$-words have the same probability of occurrence in a discourse. Similarly, in assigning ranks to words, according to their frequency, while every $k$-word will have a different rank, the actual rank assignment for the $W(K)$ words is arbitrary[18, 21].

We hypothesize that *the optimal word frequency distribution is the one that maximizes the entropy for given values of V, N and 'degenerate entropy' $H_{DN}$*. $H_{DN}$ may be regarded as the cost of encoding a 'degenerate discourse' in which word-types of the same probability of occurrence are considered indistinguishable. We will return later to a discussion of the significance of extremum principles in general in section 2.5.

Using the method of Lagrange multipliers[25], the function to be maximized is

$$H_N = N \ln N - \Sigma k W(k) \ln k \tag{2.16}$$

a function of $k_m$ variables $W(1), W(2), \ldots, W(k_m)$, subject to the three constraints:

$$\Phi_1 = \Sigma W(k) - V = 0, \tag{2.17}$$

$$\Phi_2 = \Sigma k W(k) - N = 0, \tag{2.18}$$

$$\Phi_3 = \Sigma k W(k) \ln k W(k) + H_{DN} - N \ln N = 0, \tag{2.19}$$

$\Phi_1, \Phi_2, \Phi_3$ are also functions of $W(1), W(2), \ldots, W(k_m)$.

We form the function

$$L = L[W(1), \ldots, W(k_m)] = H_N + \Lambda_1 \Phi_1 + \Lambda_2 \Phi_2 + \Lambda_3 \Phi_3, \tag{2.20}$$

where $\Lambda_1$, $\Lambda_2$, $\Lambda_3$ are constants. $H_{DN}$ and $L$ acquire stationary values (maximum or minimum) at the same set of values $W(k)$ ($k = 1, 2, \ldots, k_m$). Note that terms with $W(k) = 0$ do not contribute to $\Phi_1$, $\Phi_2$, $\Phi_3$. The stationary values are determined by

$$\partial L / \partial W(k) = 0. \tag{2.21}$$

From equations (2.16) to (2.19)

$$\partial \Phi_1 / \partial W(k) = 1, \quad \partial \Phi_2 / \partial W(k) = k$$

$$\partial \Phi_3 / \partial W(k) = k [1 + \ln k + \ln W(k)],$$
$$\partial H_{DN} / \partial W(k) = -k \ln k. \tag{2.22}$$

From equations (2.20) and (2.22)

$$\partial L / \partial W(k) = k [\Lambda_3 \ln W(k) + (\Lambda_3 - 1) \ln k + (\Lambda_1 / k) + \Lambda_2 + \Lambda_3]. \tag{2.23}$$

For stationary values of $W(k)$ ($k = 1, 2 \ldots k_m$) equation (2.21) gives

$$\ln W(k) = -\gamma \ln k - (\mu / k) + \ln C, \tag{2.24}$$

where the constants $\gamma$, $\mu$ and $C$ are given by

$$\gamma = (\Lambda_3 - 1)/\Lambda_3, \quad \mu = \Lambda_1 / \Lambda_3, \; C = \exp[-(\Lambda_2 + \Lambda_3)/\Lambda_3]. \tag{2.25}$$

Rewriting equation (2.24) as

$$W(k) = C \, e^{-\mu/k} k^{-\gamma}, (k = 1, 2, \ldots, k_m), \; W(k) \neq 0 \tag{2.26}$$

we obtain the required optimal word frequency distribution.

Differentiating equation (2.23)

$$\partial^2 L / \partial W(k)^2 = k \, \Lambda_3 / W(k) = k / [W(k) (1 - \gamma)], \quad (k = 1, 2, \ldots, k_m)$$

$$\partial^2 L / \partial W(i) \, \partial W(j) = 0 \; (i = j, \; i, j = 1, 2, \ldots, k_m). \tag{2.27}$$

If $W(k) \neq 0$ and $\gamma > 1$, all second derivatives of $L$ exist and

$$\partial^2 L / \partial W(k)^2 < 0 \quad (k = 1, 2, \ldots, k_m). \tag{2.28}$$

This implies that $L$ and therefore $H_{DN}$ *has a strict maximum* at the stationary values defined by equation (2.26). For most discourses, it is found that $\gamma > 1$ (ref. 21).

It can be shown that for given values of $V$, $N$ and $H_N$, equation (2.26) will imply *a strict maximum for the 'degenerate entropy' $H_{DN}$*.

The constants $C$, $\mu$, $\gamma$ can be determined by substituting equation (2.26) in equations (2.17), (2.18) and (2.19) and solving them for the three constants. A better procedure is to estimate the best values of $C$, $\mu$, $\gamma$ by a least square analysis of all the available data on word frequencies $W(k)$ ($k = 1, 2, \ldots, k_m$)[21, 26].

Equation (2.26) is a power law function ($k^{-\gamma}$) modified at low $k$ values by the exponential term $e^{-\mu/k}$. We refer to it as the modified power law (MPL). It is deduced using Stirling's approximation for $\ln x!$ valid for large $x$ (equation 2.7). Since $k$ and $W(k)$ take values as low as 1, equation (2.7) is not valid for $\ln k!$ and $\ln W(k)!$. A more exact form of MPL (Appendix 1) is

derived taking into account this fact. Equation (2.26) becomes

$$W(k)=C\ \mathrm{e}^{-\mu/k}k^{-\gamma}F(k), \tag{2.29}$$

where

$$F(k)=(k\ \mathrm{e}^{1/6k})^{(1-\gamma)/2k}. \tag{2.30}$$

It is the equation (2.29) that is used for statistical tests of word frequency data in ref. 21. It is found that equation (2.29) fits the data better than equation (2.26).

### 2.4 *Comparison with Mandelbrot's and Simon's models for word frequencies*

Extensive studies on word counts in various languages by different authors and in different types of discourses[16–18, 27, 28] have established a statistical law for the number of word-types $W(k)$ occurring $k$ times

$$W(k)=C\ k^{-\gamma}, \tag{2.31}$$

where $C$ is a constant and $\gamma\simeq 2.0$. Usually word frequencies are given by rank[17, 18] (see also ref. 21). All the different word-types are ranked in decreasing order of occurrence. $p(r)$ is the number of occurrences of a word of rank $r$. Zipf found that

$$p(r)=A/r, \tag{2.32}$$

where the constant $A\simeq N/10$, $N$ being the size of the discourse.

The first application of Shannon's Information Theory to linguistics — more specifically for modelling a theory of word frequencies in psycholinguistics — was made by Mandelbrot in a pioneering work in 1954. For a lucid presentation of the theory, see Mandelbrot[16] (see also Mandelbrot[29]).

First, Mandelbrot's rank frequency relation is a modified Zipf's law (equation 2.32)

$$p(r)=A(r+r_0)^{-B} \tag{2.33}$$

which fits most of the rank frequency data. $r_0$ and $B$ are constants. $r_0$ is small and accounts for the observed deviation of $p(r)$ from a pure power law ($r^{-B}$) for small $r$.

Second, Mandelbrot takes the letters and space (to demarcate words) as the $M$ primary symbols of a discourse. *Assuming all the symbols occur with equal probability*, a relation is obtained between the number of letters in a word ($m$) and its rank ($r$):

$$r\propto M^m=\mathrm{e}^{m\ln M}\ . \tag{2.34}$$

Third, the informationally optimal system of word frequencies is regarded as the one that carries the largest amount of Shannon's information, for a given *cost or mean value of letters per word*, and the given number of words in the discourse ($N$). Here information is the entropy

$$H=-\Sigma\, p(r)\log_2 p(r). \tag{2.35}$$

The optimum relation gives

$$p(r)\propto \mathrm{e}^{-\beta m(r)} \tag{2.36}$$

with $\beta$ a constant. Equations (2.36) and (2.34) together lead to

$$p(r)=A\,r^{-B}. \tag{2.37}$$

Fourth, by further refining the concept of cost — to include unequal letter costs — Mandelbrot obtains equation (2.33) which reduces to equation (2.37) when $r_0=0$.

We note several points of departure in our model (sections 2.2, 2.3) when compared with Mandelbrot's model. We will discuss them one by one.

(1) In Mandelbrot's simplified model all the letters are assumed to have the same probabilities of occurrence which is obviously untrue. Also, the number of word-types is assumed infinite whereas in real life, the speaker or writer has indeed a limited vocabulary. Mandelbrot's refined model takes into account unequal letter frequencies and finiteness of word-types. However the distribution of the words by the number of letters they contain is an exponential function whereas the observed distribution is lognormal[21].

In our model the primary symbol is the word. By choosing the 'word' as the primary symbol, our model becomes *equally applicable to written and spoken discourses*, whereas any model based on letter primary symbols will obviously apply only to written discourse.

(2) In Mandelbrot's model, for given size of the discourse ($N$) and mean 'cost' or value of the letters per word, the entropy is maximized. In our model, the MPL function maximizes the entropy for given $N$, $V'$ and degenerate entropy $H_{DN}$. Thus, we have added a constraint (the size of the vocabulary $V'$) and replaced the cost function by the degenerate entropy, which is essentially a cost function for encoding a discourse in which *all word-types with the same probability of occurrence are treated as equivalent*.

It should be noted that Mandelbrot is ambiguous about the meaning of cost of a word. While his model implies that it is the number of letters it contains, Mandelbrot writes 'this is unfortunately impossible, and the best that one can do is to look for the cost within the recoding of discourse by the higher nervous system of the receiver of the message and, perhaps, even that of the emitter'. Brillouin[30] has derived equation (2.33) without recourse to coding in the brain and the associated cost, but with the assumption of 'good channel matching' of cost to word length.

(3) The constraint on the number of word-types ($V'$) is realistic since a writer or a speaker has a limited

vocabulary at his disposal, usually a severely restricted subset of a universal dictionary spanning all possible discourses of language. It leads to the exponential term in the derived word frequency distribution (equation 2.26). When the constraint is removed, $\Lambda_1 = 0$ in equation (2.20), $\mu = \Lambda_1$ becomes 0 (equation 2.25) and the exponential term becomes 1. Then equation (2.26) reduces to a pure power law. Thus, the well-known departures of word frequency distributions from pure power law ($k^{-\gamma}$) for small $k$ (ref. 21) are accounted in terms of the extent of constraint on the size of the vocabulary. This is natural, since $\mu > 0$ implies a lack of *enough words in the vocabulary for a one-time* or infrequent use whereas $\mu < 0$ implies the opposite. $\mu = 0$ implies essentially no effective constraint of vocabulary size. All the discourses analysed in ref. 21 show $0 \leq \mu \leq 1.3$.

(4) Unlike Mandelbrot, we have chosen to deal directly with the word frequency distribution $W(k)$, the number of word-types occurring exactly $k$ times, rather than the rank distribution. The rank distribution $p(r)$ is unsuitable for rigorous statistical tests since the probabilities of different ranks are not statistically independent[17, 21, 26]. The difficulty naturally extends to the actual evaluation of entropy, for example, from equation (2.35). In our view, a differential formulation using directly the word frequencies has much to commend from the statistical as well as theoretical points of view.

(5) Mandelbrot's modified Zipf's law (equation 2.33) accounts for deviations from a pure power law for *low ranks, i.e. few words occurring very frequently* (such as 'the', 'I', 'of', etc.) whereas, in equation (2.26), the exponential term accounts for substantial deviations from pure power law in the *large number of words occurring very infrequently* (*small k*).

It is interesting to note however, that equation (2.26) can also be approximated by a function similar to equation (2.33)

$$W(k) = C'(k + k_0)^{-\gamma'}. \tag{2.38}$$

The two equations are the same when $|\mu|/\gamma \ll 1$ (ref. 26). *It should be emphasized however that equations (2.33)* and (2.38) account for departures from power law at the two extreme ends of the frequency range.

The rank frequency distribution $p(r)$ and the word frequency distribution $W(k)$ are related:

$$r = \sum_{k=p(r)}^{k=k_m} W(k). \tag{2.39}$$

Using the above, Mandelbrot's equation (2.33) can be shown to correspond to a pure power law with index $\alpha = (1+B)/B$. $B = 1$ corresponds to $\alpha = 2$. Hence any rank distribution given by equation (2.33) for small $r$, is equivalent to a pure power law relation for large $k$.

Simon has deduced a power law function for word frequencies (equation 2.31) from a stochastic model of the evolution of a discourse. It is based on two main assumptions: (1) when a discourse is being written and has reached a length of $m$ words, the probability that the next word is a word that has already occurred exactly $i$ times is proportional to $i\,p(i,m)$, where $p(i,m)$ is the number of different words that have occurred $i$ times in the first $m$ words. (2) There is a constant probability $\delta$ that the $(m+1)$th word is a 'new' word, a word that has not occurred in the first $m$ words. By making $\delta$ dependent on $m$, the exponent $\gamma$ can be made to depend on $i$. Unlike Mandelbrot's model, Simon's model uses the word as the primary symbol and the assumptions made appear to be tailored to produce a power law distribution of word frequencies. Our model, while adopting Mandelbrot's basic ideas of applying Information Theory to word frequencies, adopts the word as a primary symbol as in Simon's model, without the rather arbitrary assumptions of that model.

## 2.5 *Discussion and comments*

We summarize the essential features of our information theoretic model of word frequencies. For a discourse of $N$ words with $V$ word-types and entropy $H_N$ (Shannon's information), using words as primary symbols, we define a 'degenerate entropy' $H_{DN}$ which is the entropy of a 'degenerate discourse' in which all the word-types that have the same probability of occurrence are considered equivalent. The optimal word frequency distribution is regarded as the one that maximizes $H_N$ given $N$, $V$ and $H_{DN}$ of the discourse. Using Lagrange multipliers it is found

$$W(k) = C\,\mathrm{e}^{-\mu/k}\,k^{-\gamma}, \tag{2.26}$$

where $W(k)$ is the number of word-types occurring $k$ times.

The constants $C$, $\mu$, $\gamma$ are determined respectively by the constraints on $N$, $V$ and $H_{DN}$ (see equation 2.25). As already stated in section 2.4 the exponential term, crucial for explaining the frequencies for small $k$, emerges from the constraint on $V$. The exponent $\gamma = (\Lambda_3 - 1)/\Lambda_3$ is solely determined by the constraint on the degenerate entropy (since $\Lambda_3$ is the multiplier for $\Phi_3$). Finally $C$—essentially a normalization constant—is determined by the size of the discourse $N$.

From $H_N$ and $H_{DN}$, we obtain the entropies $H$ and $H_D$ as bits per symbol

$$H = H_N/N \quad \text{and} \quad H_D = H_{DN}/N. \tag{2.40}$$

We note that both $H$ and $H_D$ vary over a narrow range from sample to sample (Table 2 of ref. 21)—within $\pm 16\%$ for $H$ and $\pm 7\%$ for $H_D$. The mean $H$ is 9.88 bits

per word. Shannon's estimate using Zipf's rank frequency relation is 9.72 bits per word. Shannon has also determined the entropy of English words—experimentally—by measuring a person's ability to guess the 'next' letter following $n$ preceding letters (see also Pierce[31]). Shannon's estimate lies between 0.6 and 1.3 bits per letter for large $n$ ($\simeq 100$). Our estimate of mean $H_D$ for representative samples of a wide range of discourses is 4.71 or 0.86 bits per letter (taking the mean number of letters per word as 5.5, including space). This value is in the middle of the range given by Shannon. This apparent near-equality of the degenerate entropy and Shannon's empirical measurement of the entropy is interesting, but it could also be a fortuitous coincidence.

Our model applies equally to written and spoken words and therefore implies similar word frequency distributions for both types of discourse. This is in fact true, as demonstrated by Zipf[18], who has analysed speech not only of normal adults but also verbalization of children of different ages and schizophrenic speech.

The generality of the model tempts one to claim that the model most likely applies to other primary symbols also, such as syllables (below the words in hierarchy) or phrases (above the words in hierarchy). Dewey's data[20] of syllable frequencies are restricted to 1370 syllables occurring $>10$ times in a total sample size of 143,000 syllables. These syllables constitute only 31% of the total of different syllables (4400). The data are consistent with $\gamma \simeq 2.0$.

We have developed a model for alphabet and phoneme frequency distributions on lines similar to the one described in section 2.2 for word frequencies[32]. The distribution, appropriate for these primary symbols, is the rank distribution since the number of symbols is small (25 to 40) and fixed for a given language. Available data for English and several other Indian languages are in good agreement with the rank distribution predicted by the model.

For a wide variety of discourses the index $\gamma \simeq 2.0$ (see, for example, Table 1 of ref. 21) although low $\gamma$ values are also occasionally found (e.g. Shakespeare and the Indus Text). A plausible rationale for $\gamma \simeq 2$ has been provided by Naranan[26], as follows.

A new variable $\lambda (= N/k)$ is defined as the *average length of gap* or interval between two consecutive occurrences of a $k$-word type. $\lambda$ is the same for all $k$-word types of which there are $W(k)$. The gap distribution $G(\lambda)$ is the well-known gamma distribution function[33] (Appendix 2). It can be shown that for small $\mu$ ($\ll 1$) and $\gamma > 1$, $G(\lambda)$ becomes independent of $N$ only when $\gamma = 2$. Then $G(\lambda)$ becomes a uniform distribution. Thus the $W(k)$ distribution from Information Theory supplemented by the requirement of invariance of the gap distribution with the size of the discourse, leads to $\gamma = 2$.

We conclude with some general remarks on extremum principles. In mechanics an equilibrium state of a mechanical system is an extremum (maximum/minimum); the actual physical state realized is the one such that in comparison with all other possible states, it is stable against small changes of position (static case) or of orbits (dynamic case). In optics according to Fermat's principle, light travels from one point to another through an arbitrary system of mirrors and refracting glasses in such a way that the optical path is an extremum (see, for example, Feynman *et al.*[34]). Following the optical analogy, one can define functions in mechanics that take extreme values in the case of the actually realized physical state. The extremum principles, reflecting philosophical ideas, are formulated in modern axiomatic terms in variational calculus. Variational principles thus unify many diverse fields, provide powerful methods of calculation and often lead to new theoretical insights.

It is not apparent how an extremum principle—such as maximum entropy of a discourse—applies to psycholinguistics or in general in behavioural sciences. But one can argue that any abstract 'system' such as language, a free creation of the human mind, although without material existence, has well-defined relations between elements composing them and hence bears a resemblance to physical systems. Following the physical analogy it is then not surprising that extremum principles, as those defining equilibrium behaviour, can also apply to linguistic behaviour.

## 3. Summary and conclusions

(1) Using Shannon's Information Theory, we have derived a distribution function for word frequencies in a discourse: the number of word-types $W(k)$ occurring exactly $k$ times

$$W(k) = C\, e^{-\mu/k} k^{-\gamma}. \tag{2.26}$$

It is obtained as the optimal distribution that maximizes the entropy of a discourse under given constraints on the sizes of the vocabulary, of the discourse and the degenerate entropy of the discourse (sections 2.1, 2.2, 2.3). Unlike Mandelbrot's model based on Information Theory, this model uses the 'word' as the primary symbol and a cost function that is related to the degenerate entropy. Equation (2.26), a modified power law distribution (MPL), is applicable to written as well as spoken discourse, unlike Mandelbrot's model (section 2.4, 2.5).

(2) The MPL distribution fits the observed data on word frequencies with few exceptions. Ten discourses representing a wide spectrum of languages, authors, styles and types of discourse have been examined[21,30]. Most discourses conform to an index $\gamma \sim 2.0$, with some

prominent exceptions. Parameter $\mu$, which accounts for deviations from pure power law ($k^{-\gamma}$) at low frequencies, arises from the constraint on the vocabulary. Its value ranges from 0 to 1.3. Apart from $\gamma$, the parameter $\mu$ is also an important measure of the author's vocabulary, since most contribution to the total vocabulary comes from low frequencies (small $k$). Negative $\mu$ values reflect a prolific vocabulary whereas positive $\mu$ implies restricted vocabulary.

It is emphasized that rank frequency distribution for words—commonly used in the statistical and theoretical studies—is unsatisfactory for statistical tests of goodness of fit to hypotheses. A differential distribution like the MPL is best suited for such tests.

(3) Entropy ($H$) and the degenerate entropy ($H_D$)—in bits per word—for the discourses show a narrow range of values with $\langle H\rangle = 9.88$ and $\langle H_D\rangle = 4.71$. The latter value corresponds to a 'cost' of 0.86 bits per letter for English—within the range of Shannon's 'experimental' values (section 2.5).

(4) Information Theory does not determine the numerical value of the index $\gamma$. But a rationale for its being $\simeq 2.0$ in most discourses is provided by the distribution of average gap of $k$-word types, which is proportional to $1/k$ (a $k$-word type is a word-type occurring $k$ times). The 'average gap distribution' is invariant with respect to the size of the discourse only when $\gamma = 2$. The average gap distribution is itself a gamma function (section 2.5).

(5) In the hierarchical structure of language starting from letters, going up to words, phrases, sentences, etc., it is known that the frequency distribution of string lengths at every level is lognormal. This is sought to be explained by two different models[21]: (a) an information theoretic model and (b) an evolutionary model based on the theory of proportionate effect.

Power law relations of the form equation (2.31) occur frequently in information science (see Naranan[35–37]). Recently Naranan[26] has shown that a distribution given by equation (2.26) and the lognormal distribution occur in the system of natural numbers, providing some striking analogies between language and numbers.

We suggest further statistical studies on word frequencies on the following lines:

(a) It is desirable to test the validity of the MPL function for a large number of discourses, especially large ones. In particular, the spread of $\mu$ values and the correlation, if any, between $\mu$ and the vocabulary size or discourse size will be illuminating. $\mu$ is also a quantifier of author's style besides $\gamma$, which seems to show relatively less variation from discourse to discourse.

(b) There are indications that the MPL function may not fit the word frequency data at high frequencies, especially for large discourses (e.g. Shakespeare, Joyce[21]). This may require another parameter (besides $\mu$, $\gamma$) in the distribution function to fit the data for the entire range of word frequencies.

(c) Word-types are defined as words that sound and are spelt differently. This convention obviously overestimates the vocabulary of authors. It will be interesting to study the distribution of lexical units in a discourse and compare it with the word-type distribution.

(d) Word frequency studies and studies of hierarchical structures in language need to be extended to speech in many different languages, by children and adults of different ages and different ethnic and cultural backgrounds. This field of study, popular in forties and fifties, seems to have not received the attention it deserves in recent years. With the immense data handling capacity of modern computers, it should be easy to obtain the necessary statistical data.

(e) Finally, statistical data of the kind discussed in this article are practically non-existent for the Indian languages. There is thus enormous scope for statistical studies of structure of language symbols and strings in Indian languages, which are bound to be illuminating in the general context of linguistic studies and, in particular, for identifying the 'universals' of language.

**Acknowledgements.** We are grateful to A. V. John for computational help. We also thank N. M. Malwad and A. Ratnakar for their help in literature search for this study.

## Appendix 1.

## Derivation of optimal word frequency distribution using Stirling's approximation

In the derivation of equations for $H_{DN}$ and $H_N$ (equation 2.15) the approximation used for ln $x!$ is

$$\ln x! = x \ln x - x \qquad (2.7)$$

valid only for large $x$. The more appropriate approximation[25] is

$$x! = (2\pi x)^{1/2}(x/e)^x e^{1/12x} \qquad (A1.1)$$

$$\ln x! = (x+0.5)\ln x - x + 0.9189 + (1/12x). \qquad (A1.2)$$

Even for $x=1$, equation (A1.1) gives 1.0022 and for larger $x$ the error is much less. Using equation (A1.2) instead of equation (2.7) modifies equation (2.9) for $H_N$ as follows:

$$H_N = -\alpha V + N \ln N - \Sigma(k+0.5)\,W(k)\ln k - (1/12)\Sigma[W(k)/k], \qquad (A1.3)$$

where $\alpha = 0.9189$. Regarding $H_{DN}$, we note that it involves ln $k\,W(k)!$. Whereas $k$ and $W(k)$ can take values as low as 1, it is seen from most data on word frequencies that $k\,W(k)$ remains $\gg 1$ (ref. 21). Therefore $H_{DN}$ is accurately represented by equation (2.13). Working through the equations (2.17) to (2.26) with the $H_N$ given by equation (A1.3) instead of equation (2.9), the following changes are noted:

$$\partial H_N/\partial W(k) = -k \ln k - 0.5 \ln k - (1/12k) \quad \text{(A1.4)}$$

$$\partial L/\partial W(k) = k[\Lambda_3 \ln W(k) + (\Lambda_3 - 1) \ln k + (\Lambda_1/k) + (\Lambda_2 + \Lambda_3) - (\ln k/2k) - (1/12k^2)] \quad \text{(A1.5)}$$

$$\ln W(k) = -\gamma \ln k - (\mu/k) + \ln C + (1-\gamma)[\ln k + (1/6k)]/2k \quad \text{(A1.6)}$$

giving

$$W(k) = C e^{-\mu/k} k^{-\gamma} F(k) \quad (k = 1, 2, \ldots, k_m),\ W(k) = 0 \quad \text{(A1.7)}$$

where

$$F(k) = (k e^{1/6k})^{(1-\gamma)/2k} \quad \text{(2.30)}$$

$F(k)$ is a correction factor for the MPL given by equation (2.26).

## Appendix 2.

## Modified power law function and gamma function

The modified power law function

$$W(k) = C\, e^{-\mu/k} k^{-\gamma} \quad \text{(2.26)}$$

corresponds to a probability density function

$$P(k) = A\, e^{-\mu/k} k^{-\gamma} \quad \text{(A2.1)}$$

$$\int P(k)\, dk = 1. \quad \text{(A2.2)}$$

The integral converges for $\mu > 0$. The $r$th moment of the function about the origin $m_r$ exists only if $\gamma > r+1$ (ref. 33, p. 89).

$$m_r = \mu^r \Gamma(\gamma - r - 1)/\Gamma(\gamma - 1) \quad (r < \gamma - 1). \quad \text{(A2.3)}$$

In particular the mean $m_1$ exists only if $\gamma > 2$. However, in word frequency distributions the range of $k$ is $1 < k < k_m$ and a mean is defined for all $\gamma$ and also $\mu < 0$. By defining a new variable $\lambda = N/k$, equation (A2.1) becomes

$$g(\lambda) = (\mu/N)^{(\gamma-1)} e^{-\lambda(\mu/N)} \lambda^{-(\gamma-2)}/\Gamma^{(\gamma-1)} \quad \text{(A2.4)}$$

which is the gamma density function

$$f(\lambda) = a^{\alpha} e^{-a\lambda} \lambda^{(\alpha-1)}/\Gamma(\alpha) \quad \text{(A2.5)}$$

with $a = \mu/N$ and $\alpha = \gamma - 1$. $f(\lambda)$ is defined for $0 \leq \lambda \leq \infty$ and $\alpha > 0$. $f(\lambda)$ converges for $\alpha > 0$, i.e. $\gamma > 1$. Moments of all orders exist for $f(\lambda)$. In particular

$$m_1 = \alpha/a,\ m_2 = \alpha(\alpha+1)/a^2.$$

Note that the range of the variable $\lambda$ is $N/k_m \leq \lambda < N$. The range of $N/k_m$ is 10 to 100 in the present case.


1. Panini, *Ashtadyai of Panini*, Kashi Sanskrit Series, Benares, 1952.
2. Patanjali, *Vyakarana Mahabhashya of Patanjali* (ed. Joshi, S. D.), University of Poona, 1968.
3. Bhartruhari, *Vakyapadiya of Bhartruhari with Vriti* (Trans. by Subramania Iyer, K. A.), Deccan College, Poona, 1966.
4. Plato, *Great Dialogues of Plato* (Trans. by Rouse, W. H. D., eds. Warmington, E. H. and Rouse, P. G.), New American Library, 1956.
5. Jones, W., *Asiatik Researches*, 1788, **1**, 422 (Reprinted as vol. 1 of *Trans. R. Asiatic Soc. Bengal*).
6. Cavalli-Sforza, L. L., Piazza, A., Menozzi, P. and Mountain, J., *Proc. Natl. Acad. Sci. USA*, 1988, **85**, 6002.
7. Piazza, A., *Language Change and Biological Evolution*, (ed. Cavalli-Sforza) Stanford University Press, 1991.
8. Harrison, B., *An Introduction to the Philosophy of Language*, Macmillan Press Ltd., London, 1979.
9. Moore, T. and Carling, C., *Understanding Language — Towards a Post-Chomskian Linguistics*, Macmillan Press, London, 1982.
10. Chomsky, N., *Language and Mind*, Harcourt Brace, New York, 1972.
11. Chomsky, N., *Aspects of Theory of Syntax* MIT Press, Cambridge, 1965.
12. Narasimhan, R., in *Advances in Speech, Hearing and Language Processing*, vol. II, 1992, (to be published).
13. Narasimhan, R., *Modelling Language Behaviour*, Springer-Verlag, Heidelberg, 1981.
14. Shannon, C. E., *Bell Syst. Tech. J.*, 1948, **27**, 379 and 623.
15. Shannon, C. E. and Weaver, W., *The Mathematical Theory of Communication*, University of Illinois, Urbana, 1949.
16. Mandelbrot, B., in *Readings in Mathematical Social Sciences* (eds. Lazarsfield, P. F. and Henry, N. W.), MIT Press, Cambridge, 1966 and references therein.
17. Zipf, G. K., *The Psycho-Biology of Language*, Houghton Mifflin Co., New York, 1935, MIT Press, Cambridge.
18. Zipf, G. K., *Human Behaviour and the Principle of Least Effort*, Addison-Wesley, Reading, 1949.
19. Zipf, G. K., *Science*, 1942, **96**, 344.
20. Dewey, G., *Relativ Frequency of English Speech Sounds*, Harvard University Press, Cambridge, 1923.
21. Naranan, S. and Balasubrahmanyan, V. K., *Curr. Sci.*, 1992, **63**, (to be published).
22. Simon, H. A., *Biometrika*, 1955, **42**, 425.
23. Huffman, D. A., *Proc. Inst. Radio Engineers*, 1952, **40**, 1098.
24. Yavuz, D., *IEEE Trans. Inf. Theory*, 1974, **20**, 650.
25. Sokolonikoff, I. A. and Redheffer, R. M., *Mathematics of Physics and Engineering*, McGraw-Hill Inc., New York, 1966.
26. Naranan, S., *J. Sci. Ind. Res.*, (to be published).
27. Eldridge, R. C., *Six Thousand Common English Words*, The Clements Press, Buffalo, 1911.
28. Yule, G. U., *A Statistical Study of Literary Vocabulary*, Cambridge University Press, Cambridge, 1944.
29. Mandelbrot, B., *The Fractal Geometry of Nature*, W. H. Freeman & Co., San Francisco, 1983.
30. Brillouin, L., *Science and Information Theory*, Academic Press, New York, 2nd edn, 1962.
31. Pierce, J. R., *Symbols, Signals and Noise: The Nature and Process of Communication*, Harper & Brothers, New York, 1961.
32. Naranan, S. and Balasubrahmanyan, V. K., (to be published).
33. Kendall, M. G. and Stuart, T., *The Advanced Theory of Statistics*, Charles Griffin & Co., London, vols. I & II, 1961.
34. Feynman, R. P., Leighton, R. B. and Sands, M., *The Feynman Lectures on Physics*, Addison-Wesley Publishing Co. Inc., Reading, Mass., USA, 1963, vol. I.
35. Naranan, S., *Nature*, 1970, **227**, 631.
36. Naranan, S., *J. Doc.*, 1971, **27**, 83.
37. Naranan, S., *Scientometrics*, 1989, **17**, 211