Information theoretic models in statistical linguistics—Part II: Word frequencies and hierarchical structure in language–statistical tests

S. Naranan and V. K. Balasubrahmanyan

Word frequency distributions in language follow a nearly universal statistical law. We have deduced a modified power law (MPL) function for word frequencies using Shannon's Information Theory (ref. 1). Data on word frequencies from ten diverse discourses are shown to be largely consistent with the MPL function. In the hierarchical structure of written discourse, string lengths—lengths of words, phrases, sentences—follow lognormal distribution. Two models for the lognormal distribution are presented: one based on Information Theory and the other, an evolutionary model, based on the theory of proportionate effect.

1. Introduction

Statistics of word frequencies in language have been studied from the beginning of this century. Following Dewey's[2] classic work of 1923, major statistical analysis of word frequency data was done by Zipf[3] in *The Psychobiology of Language* in 1935. Striking regularities in the frequency distributions of words were noted in diverse literary styles, authors and languages. Zipf[4] extended the statistical studies to other areas of behavioural science—beyond language—in an elaborate work: *Human Behaviour and the Principle of Least Effort*. The same rank frequency law observed for words is seen to be applicable to diverse areas of social behaviour. The law known as Zipf's law is indeed a robust one.

Zipf's law is equivalent to a power law distribution given by

$$W(k) = C k^{-\gamma}, \tag{1.1}$$

where $W(k)$ is the number of different words (word-types) occurring k times in a discourse of words. C and γ are constants with $\gamma \simeq 2.0$.

A model for rank frequency distribution of words, based on Shannon's Information Theory[5], was proposed by Mandelbrot[6]. The model uses the letter or alphabet as the primary symbol and is applicable to written discourse. Recently, we[1] have proposed another model based on Information Theory, using the word as the primary symbol, which is applicable to spoken as well as written languages. We obtain a power law function modified at low frequencies (small k):

$$W(k) = C\, e^{-\mu/k}\, k^{-\gamma} \tag{1.2}$$

with C, μ, γ constants. Deviations from pure power law ($k^{-\gamma}$) observed at low k are accounted by the exponential term. A detailed comparison of this model with Mandelbrot's model is given in ref. 1.

In section 2 we have tested the 'goodness of fit' of equation (1.2) to word frequency data from ten different discourses (section 2.1). With few exceptions the data are seen to be largely consistent with the MPL function giving support to the model. The numerical values of entropy and 'degenerate entropy', a new quantity introduced in ref. 1, are presented in section 2.2.

We deal with the hierarchical structures of language in section 3. Starting from the alphabet, the successive agglomerations are the morphemes, words, phrases, sentences, paragraphs, etc. These structures are defined as strings; each string contains a number of units belonging to the immediate lower hierarchy, defined as string length. These string lengths—e.g. the number of letters per word, the number of words in a phrase, etc.—are known to follow a lognormal distribution. We present some instances of string length distribution (section 3.1) and for the first time provide some data for an Indian language (Tamil). In section 3.2 we discuss two models for the lognormal distribution of string lengths, one based on Information Theory and the other on the theory of proportionate effect. Section 4 contains some general comments and includes a summary.

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.

2. Data on word frequency distributions

Extensive studies on word counts in various languages by different authors and in different types of materials[3,4,6–8] have established a statistical law for the number of word-types $W(k)$ occurring k times (equation (1.1)). Deviations occur at small as well as large k. In fact, the law seems to be of almost universal validity.

Word frequencies are usually represented as frequencies of occurrence of words by rank (r)[3,4]. All the V word-types in a discourse of N words are ranked in decreasing order of occurrence. The word of rank 1 ($r=1$) occurs the maximum number of times (k_m). All words which occur the same number of times are given consecutive ranks arbitrarily. $p(r)$ is the number of occurrences of a word of rank r. The range of r is $1 \le r \le V$. Zipf[3] found that

$$p(r) = A/r \tag{2.1}$$

with the constant $A \simeq N/10$. This is known as Zipf's law. Equation (2.1) does not usually fit the data for low r. Mandelbrot's rank frequency relation[6]

$$p(r) = A(r + r_0)^{-B} \tag{2.2}$$

with r_0, B constants, gives good fit to data at low r. Our information theoretic model gives a modified power law function (equation (1.2)) which accounts for observed deviations from pure power law *at low k*. Since low k corresponds to high rank, equations (2.2) and (1.2) represent deviations from power law at extreme ends of the frequency range of k. Note that equation (1.2) reduces to the empirical (1.1) when $\mu = 0$.

The rank distribution $p(r)$ is unsuitable for statistical tests of 'goodness of fit' to a hypothesis since the $p(r)$ values for $r = 1, 2, \cdots, V$ are not statistically independent. The differential form $W(k)$ of equations (1.2) or (1.1) is better suited for statistical analysis[9].

2.1 *Test of MPL function for word frequencies*

We have used data on word frequencies for ten different samples of discourse to test the validity of the MPL function (equation (1.2)). The samples are: (1) Nouns (*Julius Caesar*, Shakespeare[8]); (2) Nouns (*As You Like it*, Shakespeare[8]); (3) Nouns (Essay on Bacon, Macaulay[10]); (4) Colloquial Chinese (Peiping dialect[3]); (5) Story in Russian (Pushkin[11]); (6) Four plays in Latin (Plautus[3]); (7) American Newspaper English[7]; (8) *Ulysses* (novel by James Joyce[4]); (9) Complete works of Shakespeare[12]; and (10) The Indus text[13]. These are used to represent a variety of languages, styles, types and sample sizes of written material (see also Naranan[9]).

Equation (1.2) can be linearized for multiple regression. The best fit parameters C, μ, γ, their errors, the χ^2 statistic, n_{df} the number of degrees of freedom (number of data bins minus 3) and $P(\chi^2)$ the probability of a deviation exceeding χ^2 are given in Table 1 (columns 6, 7, 8, 9 and 10). [Equation (1.2) is obtained

Table 1. *Parameters of word frequency distribution* $W(k) = C\,e^{-\mu/k}\,k^{-\gamma}\,F(k)$, $F(k) = (k\,e^{1/6k})^{(1-\gamma)/2k}$

Discourse	k_m	V	N	N/V	C (ΔC)	μ ($\Delta\mu$)	γ ($\Delta\gamma$)	χ^2 (n_{df})	$P(\chi^2)$
Julius Caesar[8]	49	964	2849	2.955	1984	1.202	2.320	11.6	0.20
(Nouns)	(125)	(965)	(2919)	(3.025)	(502)	(0.24)	(0.24)	(8)	
As You Like It[8]	59	1239	3421	2.761	2307	1.034	2.350	3 5	0.90
(Nouns)	(113)	(1241)	(3609)	(2.908)	(513)	(0.22)	(0.09)	(81)	
Macaulay[10]	89	2047	7790	3 806	3721	1.230	2.152	16.2	0.50
(Nouns)	(255)	(2048)	(8045)	(3.928)	(553)	(0 15)	(0.06)	(17)	
Chinese[3]	101	3330	10,654	3.199	2647	0.173	2.022	24 3	0.20
	(905)	(3342)	(13,248)	(3.964)	(333)	(0 13)	(0.05)	(19)	
Russian[11]	40	4698	15,611	3.323	8185	1.131	2.170	28 6⁺	0.10
	(?)	(4783)	(28,591)	(5.978)	(842)	(0.11)	(0 04)	(19)	
Latin[3]	61	8366	22,931	2.741	5432	–0.085	2.007	16.9	0.75
	(514)	(8437)	(33,094)	(3 922)	(465)	(0 09)	(0.04)	(22)	
Eldridge[7]	60	5930	20,734	3.496	12,113	1.299	2.239	16 5	0 80
	(4290)	(6001)	(43,989)	(7.330)	(1059)	(0.09)	(0.04)	(22)	
Joyce[4]	50	27,772	71,397	2.571	23,576	0 288	1.945	56 4*	≃0 005
	(?)	(29,899)	(260,430)	(8.710)	(1224)	(0.05)	(0.02)	(21)	
Shakespeare[12]	100	30,688	194,667	6.343	15,364	0.020	1.604	39.2@	0 20
	(?)	(31,534)	(884,647)	(28 05)	(490)	(0.03)	(0.01)	(33)	
Indus Text[13]	381	415	11,328	27.30	173	0.436	1.360	16.8	0.30
	(1395)	(417)	(13,372)	(32.07)	(37)	(0.24)	(0.06)	(15)	

⁺Without the last two bins ($k \le 30$), $\chi^2 = 20\,4$, $n_{df} = 17$, $P(\chi^2) = 0.25$.
*Omitting $W(2)$ and $W(5)$ $\chi^2 = 29.1$, $n_{df} = 19$, $P(\chi^2) \simeq 0\,05$.
@Omitting the last five bins, ($k \leqslant 60$), $\chi^2 = 22.6$, $n_{df} = 28$, $P(\chi^2) = 0.75$.
Numbers in parantheses in columns 2–5 refer to total data.

using Stirling's approximation for large k. Using the more accurate Stirling's approximation, the right-hand side has a multiplying term $F(k)=(k\,e^{1/6k})^{(1-\gamma)2k}$ (ref. 1). This requires that C, γ, μ are determined by using a method of successive approximation. The values in Table 1 are obtained by such a method.] Columns 2,3,4 respectively give the maximum frequency of occurrences of a word (k_m), the size of the vocabulary (V) and the total number of words (N) in the discourse used for fitting the data to equation (1.2). The distributions of word frequencies for the ten discourses are given in Figure 1 *a–d*.

For all the discourses, *the total available data on word frequencies have been used for the analysis.* In eight of the ten discourses (exceptions being Joyce and Shakespeare) the difference between the total number of word-types and the number used for analysis is very small. For Joyce, the excluded word-types constitute 7.1% of the total vocabulary but account for 72.6% of the total number of words. For Shakespeare, the corresponding figures are 2.7% and 78.0% respectively (see Table 1).

For the eight discourses, the χ^2 values are acceptable at 20% significance level or better—very satisfactory for not rejecting a hypothesis. For Joyce $\chi^2=56.4$ for $n_{df}=21$, with just two bins $W(2)$ and $W(5)$ accounting for nearly half of the χ^2 (see Figure 1 *c*). For Shakespeare $\chi^2=39.2$ for $n_{df}=33$. Here, the last five bins ($k>61$) contribute the maximum; when they are excluded $\chi^2=22.6$ for $n_{df}=28$ with $P(\chi^2)=0.75$. Using the C, μ, γ determined for Shakespeare, we calculate that $W(k>61)$, the expected number of word-types occurring >61 times is 2124 whereas the 'observed' number is 1286, clearly indicating that equation (1.2) is not a good fit for the most frequently used words with $k>61$. Similarly for Joyce the 'expected' and 'observed' word types $W(k>50)$ are 619 and 2127 respectively. In the two cases, the deviations are in opposite directions.

The reasons for the deviations could be several. Both Joyce and Shakespeare are exceptional in their usage of words and their works could be atypical of the general English literature in respect of word frequencies. It is a common experience in behavioural sciences that the χ^2 statistic tends to be higher (relative to n_{df}) for larger sample sizes. Joyce and Shakespeare sample sizes are the highest with $N=260$, 430 and 884, 647 respectively (Table 1). It has already been noted that Zipf's law generally fits the observed word frequency data for $N\simeq 10^5$ (ref. 14). It will be interesting to examine if smaller subsets of Joyce and Shakespeare works give better fit to MPL function in terms of χ^2 values.

The texts of the early urban culture of the Indus civilization date back to 2300–1750 BC. The writings on seals comprise 417 different signs (V) and 13,372 legible sign occurrences (N). The text has an unusually high N/V ratio ($\simeq 30$) and low index γ (1.36 ± 0.06) compared to other discourses although the χ^2 (16.8) for $n_{df}=15$ is well within acceptable values (Table 1). Two noteworthy facts about the text are the following[13]: (i) 112 signs ($\simeq 27\%$ of the total) occur only once each; most of them are compounds of two or more other signs and their independent status as different signs is doubtful, (ii) the most frequently occurring signs show considerable graphic variations; all the variants of a sign are regarded as a single sign.

According to Subbarayappa[15] a large fraction of the Indus signs are numeric signs; it is conjectured that the Indus texts are records of commercial transactions of agricultural and other products. Since numeric symbols may have a frequency distribution very different from that of linguistic symbols, the observed index γ may not be typical of linguistic texts. If all numeric signs can be identified and excluded, the index γ will most likely increase towards 2.

We have examined the frequency distribution of digrams in English (Gaines[16]). The number of digrams $V=430$ and sample size $N=10{,}000$. The data for $k<132$ give a very good fit to the MPL with $V=422$, $N=8249$, $C=264\pm14$, $\mu=1.70\pm0.29$, $\gamma=1.35\pm0.07$, $P(\chi^2)=0.08$ and $n_{df}=16$. It is very interesting that V, N/V and γ are very similar to the values for the Indus Text, suggesting that most of the signs in the Indus Text are very likely digrams or compound symbols, supporting Mahadevan[13].

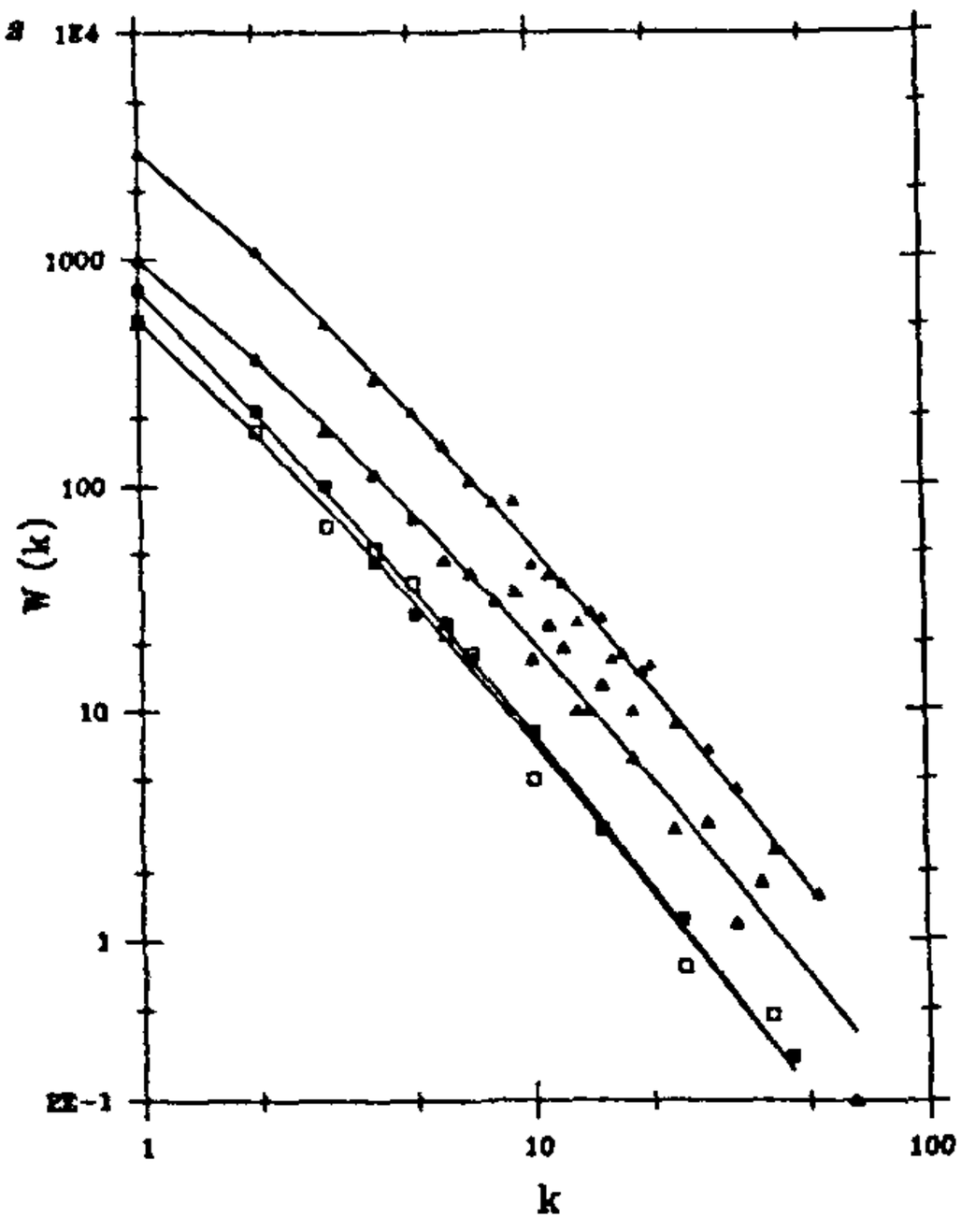
From Table 1 we note the following:

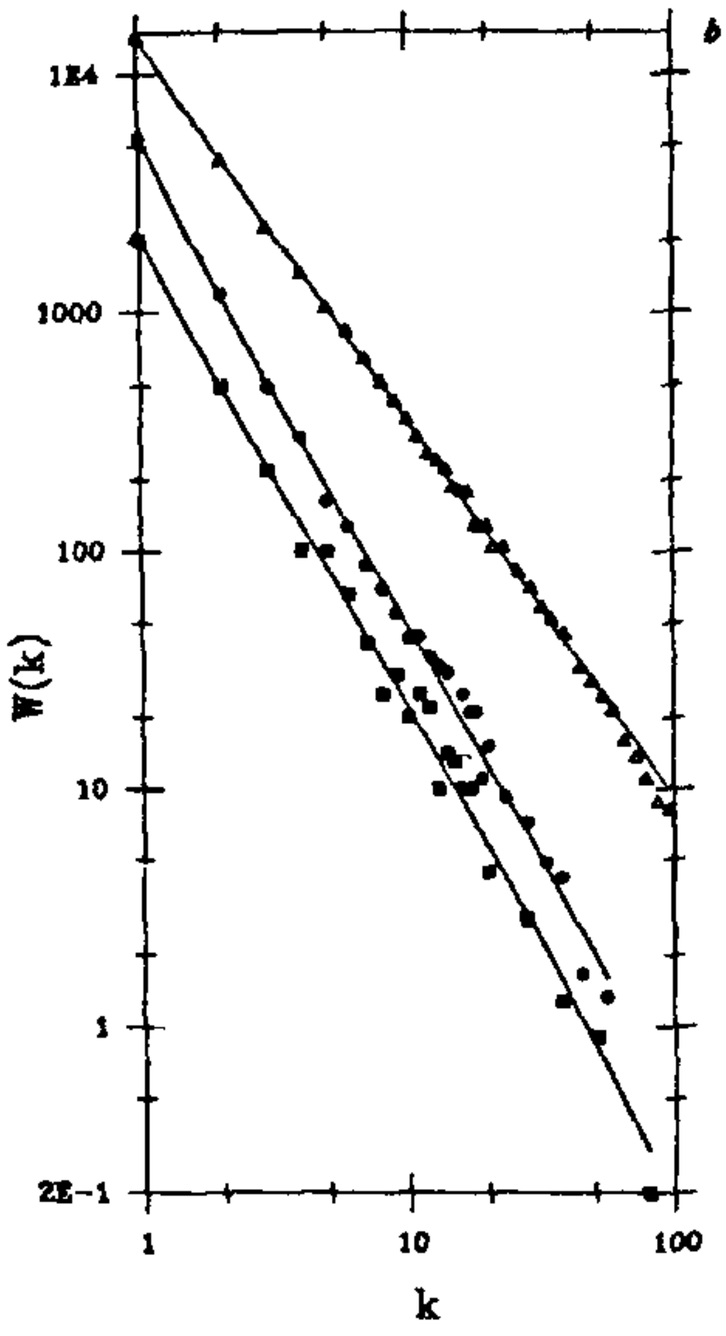
(1) In eight of the ten discourses, γ varies from 1.95 ± 0.02 (Joyce) to 2.35 ± 0.10 (*As You Like It*), not significantly different from 2.0. For Shakespeare $\gamma=1.60\pm0.01$, significantly different from 2.0 and the lowest γ (1.36 ± 0.06) is for the Indus Text.

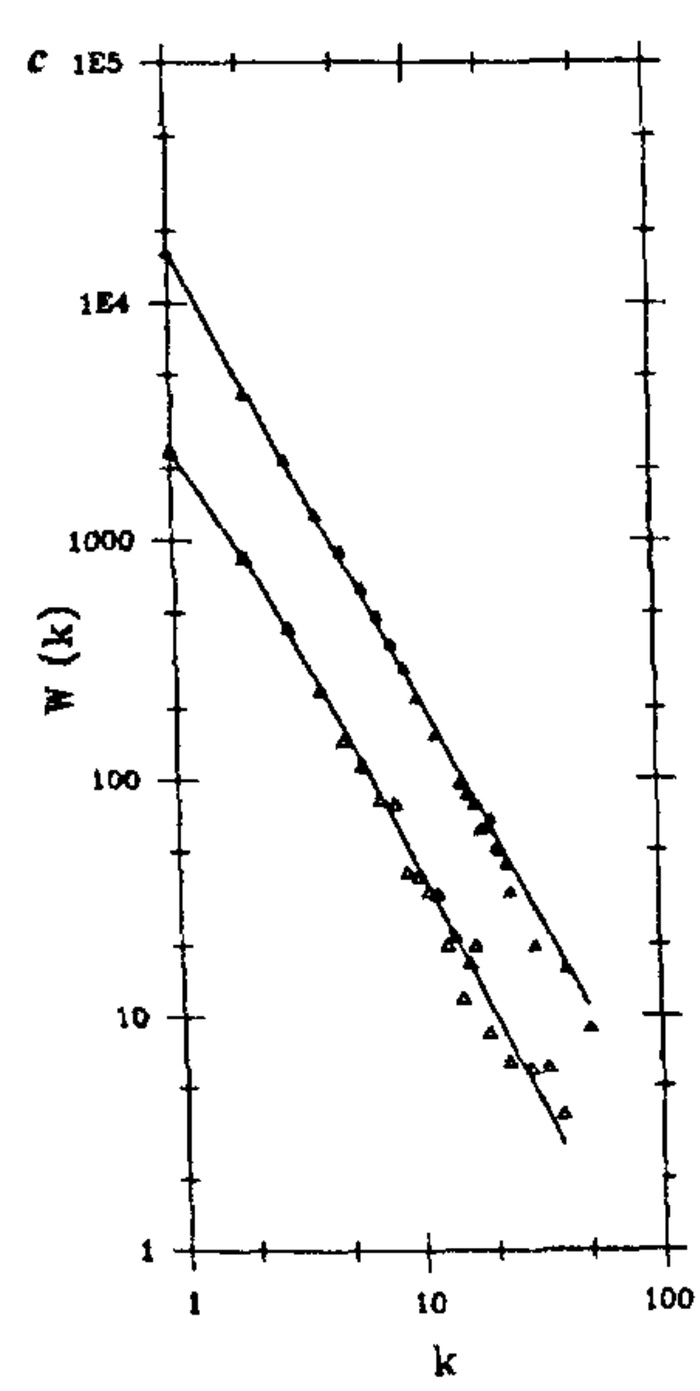
(2) The value of μ ranges from -0.09 ± 0.09, consistent with 0 (Latin) to 1.30 ± 0.10 (Eldridge). The most significant positive μ values are for Eldridge, Pushkin and Nouns (Macaulay, Shakespeare) with $1.0<\mu<1.3$. $\mu>0$ signifies fewer words occurring rarely (small k) than implied by a power law, whereas $\mu<0$ implies the opposite.

The effect of the correction factor $F(k)$ for equation (1.2) is to increase slightly μ and γ and reduce χ^2. Without the correction, the range of μ is $-0.4<\mu<0.9$ and γ is lower by $\simeq 0.1$. It is noteworthy that the reduction in χ^2 is significant especially for discourses with relatively large χ^2. For example, for Shakespeare, without the $F(k)$ term $\chi^2=54.2$ much higher than 39.2 given in Table 1. This can be regarded as an additional support for our model of word frequencies[1].

We therefore conclude that the MPL function does indeed describe the word frequency data adequately with $\gamma\simeq 2.0$ and $0<\mu<1.3$ with a few exceptions noted regarding deviations for large frequencies.







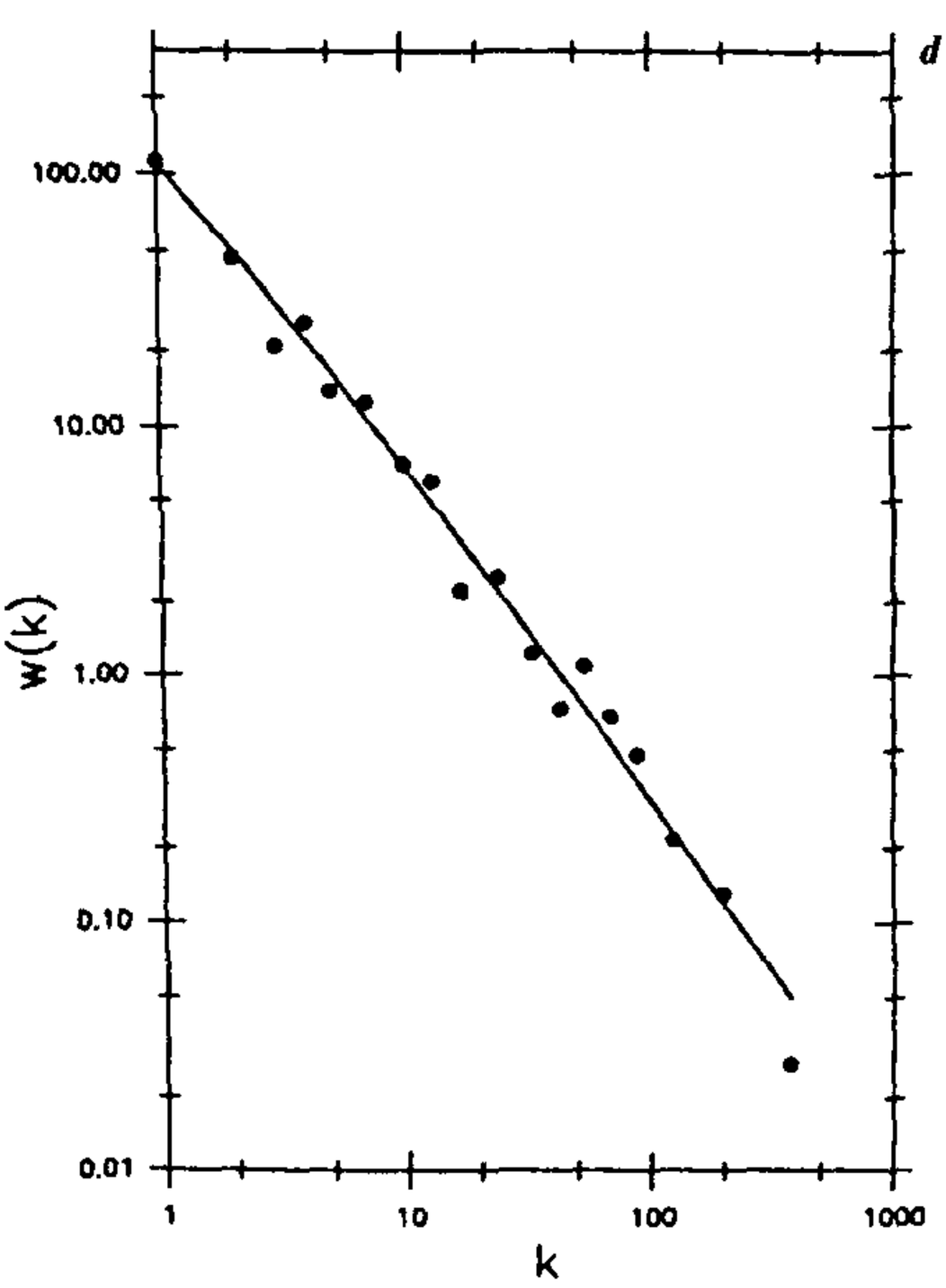


Figure 1. Word frequency distributions for ten discourses. $W(k)$ is the number of word-types occurring exactly k times. ***a.*** ▲ Eldridge, △ Macaulay, ■ *As You Like It*, □ *Julius Caesar*; ***b.*** ● Latin, ■ Chinese, △ Shakespeare; ***c.*** ▲ Joyce, △ Russian; ***d.*** The Indus Text. (see Table 1 and section 2.1.)

2.2 *Entropy and cost of encoding the discourse*

We have proposed an information theoretic model for word frequencies[1]. The essential features of the model are as follows. For a discourse of N words with V word-types and entropy H (Shannon's information), using words as primary symbols, we define a quantity called 'degenerate entropy' H_D. This is the entropy for a 'degenerate' discourse in which word-types occurring with the same probability (same k) are regarded as indistinguishable. The optimal word frequency distribution is the one that maximizes the entropy H, given N, V, and H_D. It is given by

$$W(k)=C e^{-\mu/k} k^{-\gamma}, \qquad (1.2)$$

where $W(k)$ is the number of word-types occurring k times. The quantities N, V, H and H_D are defined as

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

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

$$H=K[N \ln N-\Sigma k W(k) \ln k]/N$$

$$H_D=K[N \ln N-\Sigma k W(k) \ln k W(k)]/N. \qquad (2.3)$$

Here, the summation Σ is over the range $1 \leq k \leq k_m$. The scale factor $K=\log_2 e=1.443$. H and H_D are in units of bits per word.

We have used the actual word frequency data to calculate the entropy H and the degenerate entropy H_D for eight discourses. Joyce and Shakespeare are excluded because, as already noted, the available data on word frequencies are incomplete. H and H_D are given in columns 3 and 4 of Table 2. The maximum value of H, H_{max} ($=\log_2 N$) is given in column 2. The difference $H_{max}-H$ is a measure of the information content manifested in the discourse due to the non-uniform distribution of word frequencies. This leads to a quantitative measure of 'redundancy'[17]

$$R_H=1-(H/H_{max}), \qquad (2.4)$$

which is given in column 5.

The following remarks apply to the seven discourses in Table 2, excluding the Indus Text, which we have seen to be atypical. H ranges from 8.28 (Julius Caesar) to 11.32 (Latin) with mean 9.88. H_D varies from 4.31 (Julius Caesar) to 5.03 (Macaulay) with mean 4.71. The difference between H and H_D is the difference in the costs of encoding the discourse and the corresponding 'degenerate discourse'. Note that the variation in H_D is less than the variation in H: within $\pm 16\%$ for H and within $\pm 7\%$ for H_D. The redundancy R_H ranges from 0.218 to 0.280 with a mean 0.254, the variation being within $\pm 12.5\%$ of the mean.

Table 2. Entropy (H) and degenerate entropy (H_D) parameters

Discourse	H_{max}	H	H_D	R_H
Julius Caesar[8]	11.51	8.28	4.31	0.280
As You Like It[8]	11.82	8 57	4 45	0.274
Macaulay[10]	12.97	9.36	5.03	0.278
Chinese[3]	13.38	9 89	4 93	0.261
Russian[11]	13.93	10.71	4 65	0.231
Latin[3]	*14.48*	*11.32*	4 68	0 218
Eldridge[7]	14.34	11.00	4.93	0 233
Indus Text[13]	13.71	6.54	6 01	0 523
Mean*		9.88	4.71	0 254

*Excluding Indus Text.

3. Hierarchies of language symbols—some statistical laws

So far we have considered only the frequency distribution of words in a discourse, and seen that a power law relation modified for low frequencies, such as equation (1.2) describes the data adequately.

Statistical studies have also been done on the language symbols at different hierarchical levels—starting from letters and moving up to syllables (phonemes), words, phrases, sentences, paragraphs, etc. The distribution relevant for these is the *lognormal distribution*[18]. A variable x is lognormally distributed if $z=\ln x$ is normally distributed. The probability density function is given by

$$d\Lambda(x)=(1/\sigma\sqrt{2\pi})(1/x)$$
$$\exp[-(\ln x-\mu)^2/2\sigma^2]\, dx\,(x>0) \qquad (3.1)$$

with μ and σ completely specifying the distribution. They are the mean and standard deviation of the *normal variate* $z=\ln x$.

A large number of statistical studies of hierarchical language symbols have been summarized by Dolby[19]: "... words are formed by variable length strings of alphabets; phrases are then constructed as strings of words, sentences as strings of phrases, paragraphs as strings of sentences and so on. This process of successive agglomeration occurs with great statistical accuracy. (1) The frequency distribution of string length is well approximated at every level by the lognormal distribution. (2) The mean (α) of the lognormal distribution is constant when the measurements are made in terms of next lower level; e.g. the mean number of sentences per paragraph is the same as the mean number of phrases per sentence etc. The constant will vary from one type of text to another but appears to be $\simeq 2e$ (5.4) for non-fiction library materials. (3) The variance (β^2) of the lognormal distribution is also a constant... the coefficient of variation ($\eta=\beta/\alpha$) is constant on the original scale of measurement. It is also a function of the type of text studied, but appears to be 0.25 to 0.30."

The mean α and the standard deviation β of the lognormal variate x are related to μ and σ, the mean and standard deviation of the normal variate $z(=\ln x)$:

$$\alpha = \exp[(\mu + (\sigma^2/2)], \beta = \alpha[\exp(\sigma^2) - 1]^{1/2}$$
$$\eta = \beta/\alpha = [\exp(\sigma^2) - 1]^{1/2}. \quad (3.2)$$

3.1 *Test of lognormal distribution of string lengths*

We test the hypothesis of lognormal distribution following Aitchison and Brown[18]. The distribution

$$\Lambda(x|\mu, \sigma^2) = Pr[X < x] \quad (3.3)$$

corresponds to a normal (Gaussian) distribution function

$$N(z|\mu, \sigma^2) = Pr[Z < z] \quad (3.4)$$

with $z = \ln x$. The standardized normal distribution with mean 0 and standard deviation 1 is

$$N(y|0,1) = [1/\sigma\sqrt{(2\pi)}]\int_0^y \exp(-y^2/2)\, dy \quad (3.5)$$

where the standardized normal variable y is given by

$$y = (z - \mu)/\sigma \text{ or } z = \sigma y + \mu. \quad (3.6)$$

From the data, for a given value of z $(=\ln x)$, the corresponding y can be obtained from standard Tables of the normal distribution. For a lognormal distribution a plot of $\ln x$ vs. y is a straight line with slope σ and intercept μ on the ordinate axis. For dealing with discrete values of x, we adopt a prescription of J. L. Williams described by Aitchison and Brown[18]. The estimates of μ and σ given in Table 3 (see below) are obtained by linear regression.

We describe a few specific examples of string length distributions. The number of letters in a word is known to be lognormally distributed[20]. Herdan[21] finds the word length distribution of 738 most commonly occurring words in telephone conversations is lognormal with $\mu = 1.62$ and $\sigma = 0.39$. The word length distribution in a dictionary as well as discourse is lognormal. Kaeding's data on the distribution of number of syllables in German words quoted by Zipf[3] cover 20 million syllables in a discourse of 10,910,777 words. (For the sheer size of the sample used, this perhaps holds the record in linguistic studies.) We have estimated the lognormal parameters μ, σ, the mean α, standard deviation β and the coefficient of variation η for these samples (Figure 2 *a* and Table 3). The sentence length distributions (Williams[22]) for samples of 600 sentences each from the writings of G. K. Chesterton, H. G. Wells and G. B. Shaw fit lognormal distribution. The parameters estimated by Williams are also included in Table 3.

We are not aware of any data on string length distributions in Indian languages. For a start, we have obtained the word length and sentence length distributions for two discourses in Tamil. The results are given in Figure 2 *b* and Table 3. The lognormal fits are satisfactory. In the case of word length distribution ('B' in Figure 2 *b*), the last three points which deviate from the line account for only 2% of the data. Note that μ and σ (1.43 and 4.52) for word length distribution in Tamil are somewhat higher than corresponding values for English (1.18 and 3.57). (The mean word length for English is usually quoted as 4.5 (ref. 14) similar to the value for Tamil. The lower value 3.57 in Table 3 refers to the 750 most frequently used English words.) The

Table 3. Lognormal distribution $\Lambda(X|\mu, \sigma^2)$ of string lengths

Language	String (text)	X (String length)	N	μ	σ	α	β	η	Ref.
English	Word (Dictionary)	Number of letters	387	1 96	0 34	7.48	2 60	0.35	20
English	Word* (Discourse)		75,624	1.18	0.44	3.57	1.66	0.46	20
German	Word (Discourse)	Number of syllables	10,906,235	0.54	0.48	1 91	0 96	0.50	3
English	Sentence (Chesterton)	Number of words	600	3.16	0 46	26.1	12.7	0.49	22
English	Sentence (Wells)		600	3.02	0 55	23.7	14.0	0 59	22
English	Sentence (Shaw)		600	3.20	0.67	30.7	23.0	0.75	22
Tamil	Word (Discourse)†	Number of letters	885	1 43	0.40	4.52	1 88	0 42	This work
Tamil	Sentence (Discourse)@	Number of words	799	2.04	0.56	9.00	5 49	0 61	This work

For comments on estimates of errors of μ, σ, α, β see text.
N = Number of strings in the sample.
*Dewey's data quoted in ref. 20 for 750 most frequently used words.
†Text. *Arunachala Mahimai* (a biography of Ramana) by Bharanidharan, vol. 2, Ch. 1, 2.
@ Text: *Venkatam mudal kumari varai* (on temples of South India) by Bhaskara Thondaiman, Ch. 1.

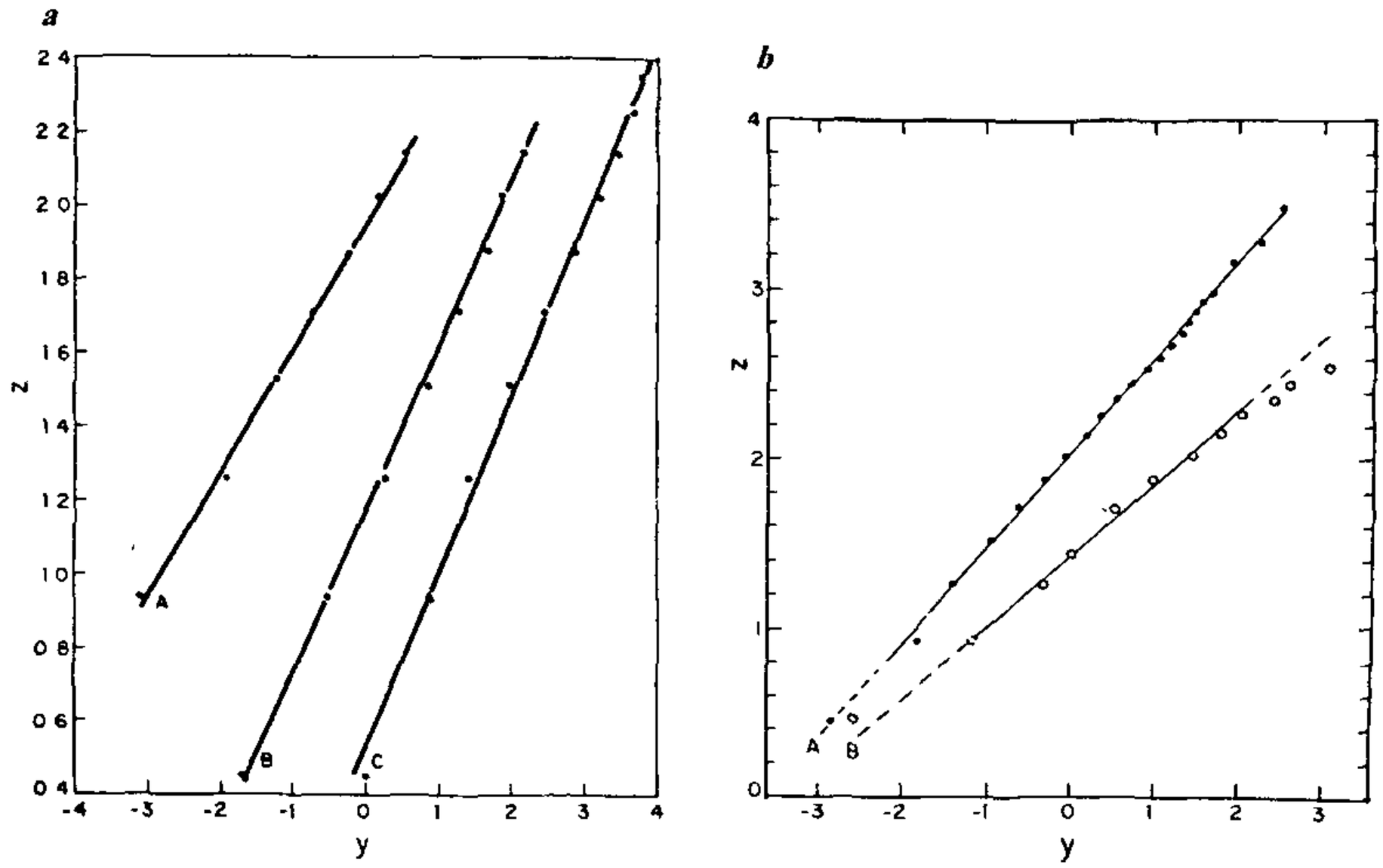


Figure 2. Graphical test of lognormal distribution of string length x, y is defined by equation (3.6) and $z = \ln x$. ***a***, A: x = *number of letters in word, text: dictionary (English)*; B: x = number of letters in word, text: discourse (English); C: x = number of syllables in word, text: discourse (German). ***b***, A: x = number of words in a sentence, text: discourse (Tamil); B: x = number of letters in word, text: discourse (Tamil).

parameter σ is however similar in both (0.40 for Tamil and 0.44 for English). For sentence length distribution in Tamil $\mu = 2.0$ and $\alpha = 9.0$ much less than the corresponding values for Chesterton, Wells and Shaw ($\mu = 3.0 - 3.2$, $\alpha = 26 - 31$). Again σ for Tamil (0.56) is well within the range for English authors ($\sigma = 0.46$–0.67). It would be interesting and perhaps more relevant to compare the sentence length distribution in Tamil with similar one for modern English literature. The coefficient of variation η, which depends only on σ (equation (3.2)), is the same for Tamil and English discourses.

Assuming that the lognormal distribution is a good description of the data, one can estimate μ and σ and the errors $\Delta\mu$ and $\Delta\sigma$ using the maximum likelihood (ML) method[23]. These estimates are given by

$$\mu_{ML} = \langle \ln x \rangle, \Delta\mu = \mu_{ML}/\sqrt{N} \tag{3.7}$$

$$\sigma_{ML}{}^2 = \langle \ln^2 x \rangle - \langle \ln x \rangle^2, \Delta\sigma_{ML} = \sigma_{ML}/\sqrt{(2N)}, \tag{3.8}$$

where N is the sample size. From the sample sizes in Table 3, it is seen that the fractional errors $\Delta\mu_{ML}/\mu_{ML}$ and $\Delta\sigma_{ML}/\sigma_{ML}$ are at the most a few per cent.

3.2 *Models for lognormal distribution of string lengths in discourses*

There is so far no satisfactory theory for the lognormal distribution of string lengths in language. We present two models — the first based on a hypothesis invoking Information Theory and, the second, a more general theory based on the theory of proportionate effect.

3.2.1 *Model based on Information Theory*

The principal idea for the model derives from observations by Herdan[21] on the distribution of *number of letters in a word* (m) *in a dictionary* as well as a *discourse*. (In a discourse the words occur with a frequency $W(k)$, whereas in a dictionary, each word-type occurs only once.) Herdan found that both the distributions are lognormal. Further, the same is true for the distribution of the number of phonemes in a word. Herdan's explanation is based on two facts:

(1) The moment of a lognormal variate is also a lognormal variate. In particular the jth moment of a lognormal distribution $\Lambda(x|\mu, \sigma^2)$ is a lognormal distribution $\Lambda_j(x|\mu', \sigma'^2)$, where

$$\mu' = \mu + j\,\sigma^2, \sigma' = \sigma. \tag{3.9}$$

(2) The probability of occurrence of a word with m letters in a discourse

$$p(m) = a\,m^{-t} \tag{3.10}$$

with a and t constants. From data $t \simeq 2.4$.

If $\Lambda(m|\mu,\sigma^2)$ is the lognormal distribution in a dictionary, then it has to be weighted with the function $p(m)$ to obtain the distribution in a discourse. Since $p(m) \propto m^{-t}$, the distribution will be a lognormal with $j = -t = -2.4$

$$\mu' = \mu - 2.4\,\sigma^2,\ \sigma' = \sigma. \tag{3.11}$$

Herdan finds that the observed μ, μ', σ, σ' satisfy equation (3.11) remarkably well.

Equation (3.10) is a result of the well-known fact that short words are most frequent. For example, the 50 most frequently used words in English—accounting for a large fraction of discourses—are all monosyllabic. This is a principle of 'least effort' which is widely applied in all coding schemes—from the Morse code (where the most commonly used letter 'e' is represented by a single dot), stenography (where the most commonly occurring word 'the' is again represented by a dot) to the Huffman code (ref. 14) in which the most frequently used symbols are assigned the smallest number of bits.

If we assume that all string lengths are distributed according to equation (3.10)—i.e. the number of phrases with w words, number of sentences with p phrases, etc. are given by an equation similar to (3.10)—then it follows that a lognormal distribution in a 'dictionary' (of phrases, sentences, etc.) will transform as a lognormal distribution in the discourse as well. As we move up in the hierarchy of symbols, dictionary and discourse are not very different and the exponent t in equation (3.10) will tend to 0, making $\mu' = \mu$ and $\sigma = \sigma'$.

Now, it only remains to explain why a lognormal distribution applies for a dictionary. It is an interesting fact that both the lognormal variable and the entropy involve logarithmic functions, and the following model is suggested.

The entropy of a symbol i occurring with probability p_i is the logarithm of the number of different possible occurrences of the symbol and is $\log_2(1/p_i)$. Averaging over all the different symbols (n)

$$H = -\sum_{i=1}^{i=n} p_i \log_2 p_i \quad \text{bits per symbol} \tag{3.12}$$

the standard expression for the entropy H. The entropy of a word of m letters, $H(m)$ is

$$H(m) = -\log_2 p(m). \tag{3.13}$$

Using equation (3.10)

$$H(m) = -\log_2 a + t \log_2 m \tag{3.14}$$

implying that $H(m)$ is linearly related to $\log m$.

In general, a word is formed according to the general rules applicable to the language. Often longer words are derived from other words (such as 'motherhood'). The number of rules and restrictions due to semantics, syntax, etc. in forming words are many; so we can consider a word as governed by the random collection of the information-constituting symbols (alphabets or phonemes) which, by the Central Limit Theorem, would constitute an 'information string' with a Gaussian (normal) distribution. If we stipulate that the ubiquitous normal distribution applies to $H(m)$, the entropy or information content of the word, then it *follows that m is lognormally distributed.* From equation (3.14)

$$\log_2 m = [H(m) + \log_2 a]/t. \tag{3.15}$$

Specifically, if $H(m)$ is normal with mean h and variance s, then $\log_2 m$ is normally distributed with mean h' and variance s' given by

$$h' = (h + \log_2 a)/t,\ s' = s/t^2. \tag{3.16}$$

The total entropy of the discourse is

$$H = \sum_{m=1}^{m=n} p(m)\,H(m).$$

Substituting eqution (3.14) and using $\Sigma\, p(m) = 1$,

$$H = -\log_2 a + t\,\langle \log_2 m \rangle. \tag{3.17}$$

Since the maximum likelihood estimate of μ is $\langle \ln m \rangle$ (equation (3.7)) the entropy H is linearly related to μ.

As already noted, the above model applies for all string lengths, if they are distributed according to an equation similar to (3.10).

3.2.2 *Theory of proportionate effect*

An alternate model of lognormal distribution of string lengths is provided by the 'theory of proportionate effect' (see Aitchison and Brown[18]). In a discourse, strings grow in size by aggregation. A string of some initial length X_0 grows to length X_n in n steps and at the jth step

$$X_j - X_{j-1} = \varepsilon_j X_{j-1}, \tag{3.18}$$

where ε_j's are mutually independent random numbers. The change at any step is proportional to the value of the variable at that step. Then

$$\sum_{j=1}^{n} (X_j - X_{j-1})/X_{j-1} = \sum_{j=1}^{n} \varepsilon_j. \tag{3.19}$$

For infinitesimal steps, the left-hand side is replaced by the integral

$$\int dX/X = \ln X_n - \ln X_0 \tag{3.20}$$

giving

$$\ln X_n = \ln X_0 + \varepsilon_1 + \varepsilon_2 + \ldots \varepsilon_n. \tag{3.21}$$

By the Central Limit Theorem, $\ln X_n$ is asymptotically

distributed normally and hence X_n is lognormally distributed.

This model is more general than the first model (section 3.2.1) and can be applied to a discourse directly without involving the dictionary. But it implicitly assumes that step increases in length are proportional to the 'current' length. On the other hand, the information theoretic model makes a plausible assumption that the entropy or information of a string follows the canonical normal distribution; coupled with the observed power law relation between string frequencies and their length (equation (3.10)), this assumption leads to a lognormal distribution of string lengths.

The two models described *together* suggest a rationale for equation (3.10) which embodies the principle of least effort in some sense. It can be shown that it follows from two hypotheses: (a) the entropy of a string is normally distributed, and (b) strings grow by the proportionate effect.

4. Discussion and summary

We have proposed a new model for the distribution of word frequencies based on Shannon's Information Theory[1]. It is different from earlier models in two essential ways: (a) it uses the 'word' as the primary symbol, (b) it defines a cost function related to the 'degenerate entropy' of the discourse. The entropy of the discourse (H) is maximized under given constraints of the size of the vocabulary (V), the size of the discourse (N) and the degenerate entropy (H_D). The optimal word frequency distribution is

$$W(k) = C\, e^{-\mu/k} k^{-\gamma}, \qquad (1.2)$$

where $W(k)$ is the number of word-types occurring exactly k times. C, μ, γ are constants. Equation (1.2) is a power law modified at low frequencies (small k).

(1) The MPL function (equation (1.2)) has been tested for data from ten different samples of discourse covering a wide range of languages, authors, style and literature (Table 1, Figure 1). The parameter μ, which determines the behaviour of equation (1.2) for small k, ranges from 0 to 1.3. $\mu > 0$ implies a restricted vocabulary and $\mu < 0$, a prolific vocabulary. The index $\gamma \simeq 2.0$ with two prominent exceptions (Shakespeare and Indus text). μ and γ are quantifiers of author's vocabulary with μ being more variable than γ. Only two samples (Shakespeare, Joyce) show significant deviations from the MPL function, especially for large k. While the low γ for Shakespeare (1.60 ± 0.01) could be attributed to the author's style, the low value for the Indus text ($\gamma = 1.36 \pm 0.06$) could be due to several unknown factors about the signs themselves, e.g. compound symbols, graphic variants and possible large admixture of numeric symbols with linguistic signs.

(2) H and H_D for eight discourses (Table 2) show small variation from sample to sample. $\langle H \rangle = 9.88$ and $\langle H_D \rangle = 4.71$ bits per word. The latter corresponds to 0.86 bits per letter for the English language, compatible with the 'experimental' value (0.6–1.3) of Shannon[1]. This could be a fortuitous coincidence.

(3) In the hierarchical structure of language, string lengths—such as the number of letters per word, number of words per sentence—are known to conform to a lognormal distribution. Actual lognormal fits to some already existing data are demonstrated (Figure 2*a*, Table 3) and for the first time new data are given for an Indian language (Tamil, Figure 2*b*, Table 3).

(4) Two different models for the lognormal distribution of string lengths are presented: (a) In the information theoretic model, it is postulated that the entropy of a string is normally distributed. This, coupled to the fact that string frequencies (by their length) are given by a power law, leads to a lognormal distribution of string lengths (section 3.2.1). (b) In the theory of proportionate effect, it is assumed that strings grow or evolve in a stochastic process with an infinitesimal increase at every step being proportional to the string length at that step. This leads—by the Central Limit Theorem—to a lognormal distribution of string lengths (section 3.2.2).

Lognormal distributions are encountered frequently in physical, biological and behavioural sciences. For methods of testing lognormal hypothesis, parameter estimation and an extensive bibliography, see Crow and Shimuzu[24]. For an illuminating study of 'long-tailed' distributions—especially the lognormal—in condensed matter physics, hydrodynamics and astronomy, see Zeldovich, Ruzmaikin and Sokoloff[25]. The authors refer to the phenomenon responsible for lognormal distribution as 'intermittency' and the theory is closely related to the theory of proportionate effect. The connection between the lognormal and the power law (also referred to as '$1/f$ noise) distributions—both long-tailed distributions—is discussed by Montroll and Shlesinger[26] and it is applied for studies of surface growth and directed polymers by Yi-Cheng Zhang[27].

1. Naranan, S. and Balasubrahmanyan, V. K., *Curr. Sci.*, 1992, **63**, 261–269.
2. Dewey, G., *Relativ Frequency of English Speech Sounds*, Harvard University Press, Cambridge, 1923.
3. Zipf, G. F., *The Psychobiology of Language*, Houghton Mifflin Co., New York, 1935.
4. Zipf, G. K., *Human Behaviour and the Principle of Least Effort*, Addison-Wesley, Reading, 1949.
5. Shannon, C. E., *Bell Syst. Tech. J.*, 1948, **27**, 379 & 623.
6. Mandelbrot, B., in *Readings in Mathematical Social Sciences* (ed. Lazarsfield, P. F. and Henry, N. W.), M. I. T. Press, Cambridge, 1966 and references therein.
7. Eldridge, R. C., *Six Thousand Common English Words*, The Clements Press, Buffalo, 1911.
8. Bennett, P. E., in *Statistics and Style* (ed. Dolezel, L. and Bailey, R. W.), American Elsevier Publishing Co. Inc, New York, 1969.

9 Naranan, S., *J. Sci. Ind. Res.*, (to appear).
10. Yule, G. U., *A Statistical Study of Literary Vocabulary*, Cambridge University Press, Cambridge, 1944.
11. Herdan, G., *Quantitative Linguistics*, MacMillan Press, London, 1958.
12. Efron, B. and Thisted, R., *Biometrika*, 1976, **63**, 435.
13. Mahadevan, I., *The Indus Script, Texts, Concordance and Tables*, Archaelogical Survey of India, New Delhi, 1977.
14. Pierce, J. R., *Symbols, Signals and Noise: The Nature and Process of Communication*, Harper & Brothers, New York, 1961.
15. Subbarayappa, B. V., *Q. J. Mythic Soc.*, 1987, **78**, and 1988, **79**.
16. Gaines, H. F., *Cryptanalysis*, Dover Publications, New York,
17. Welsh, D., *Codes and Cryptography*, Clarendon Press, Oxford, 1989.
18. Aitchison, J. and Brown, J. A. C., *The Lognormal Distribution*, Cambridge University Press, Cambridge, 1957.
19. Dolby, J. A., *J. Document.*, 1971, **27**, 136.
20. Bell, D. A., *Information Theory and its Engineering Applications*, Sir Isaac Pitman & Sons Ltd., London, 1956.
21. Herdan, G., *Biometrika*, 1958, **45**, 222.
22. Williams, C. B., *Biometrika*, 1940, **31**, 356.
23. Kendall, M. G. and Stuart, T., *The Advanced Theory of Statistics*, Charles Griffin & Co, London, Vols. I & II, 1961.
24. Crow, E. L. and Shimuzu, K. (eds), *Lognormal Distributions*, Marcel Dekker Inc., New York, 1988.
25. Zeldovich, Ya B., Ruzmaikin, A. A. and Sokoloff, D. D., *The Almighty Chance* (Ch 8), World Scientific, Singapore, New Jersey, London, Hong Kong, 1990.
26. Montroll, E. W. and Shlesinger, M. F., *Proc. Natl. Acad. Sci. USA*, 1982, **79**, 3380.
27. Yi-Cheng Zhang, in *Fractals: Physical Origin and Properties* (ed. Pietronero, L.), Plenum Press, New York, 1990.

ACKNOWLEDGEMENTS. We are grateful to A. V. John for computational help. We also thank Smt. Mythili Rangarao for the reference to Indus Text (ref.13).

Glucocorticoids: The anti-inflammatory agents

K. K. Mishra and H. P. Pandey

Four decades have passed since the discovery of anti-inflammatory effects of glucocorticoids, yet the function of these compounds has remained an enigma and eluded the scientific community. However, glucocorticoids exert profound suppressive effects at almost every step of inflammation and they have a significant therapeutic role in medical practice. This article is an attempt to give a generalized account of glucocorticoid action at a molecular level.

One of the most important effects of glucocorticoids was discovered almost by chance in the late forties when it was observed by Hench *et al.*[1] that administration of cortisone reduced the severity of disease in patients suffering from rheumatoid arthritis. This discovery led to the Nobel prize for medicine in 1950, and called global attention to the anti-inflammatory effects of glucocorticoids. Since then, four decades have passed yet the anti-inflammatory effects of glucocorticoids are still not fully understood and are ruled out by some as pharmacological side-effects[2], produced by overdoses of hormone. Virtually it was Hench who in 1929 noticed that the condition of his patients with rheumatoid arthritis improved if they became pregnant or jaundiced. He thought that it might be due to a hormone from the adrenal cortex but he had to wait till 1949 to test his hypothesis when he with his colleagues synthesized cortisone. Administration of cortisone brought about rapid relief of the symptoms of rheumatoid arthritis[3]. For this remarkable achievement, Hench and his associates, Kendall and Reichstein, were jointly awarded the Nobel prize.

K. K. Mishra and H. P. Pandey are in the Department of Biochemistry, Banaras Hindu University, Varanasi 221 005, India.

Inflammation and its mediators

Inflammation, stated to be an essential prelude to healing, is the response of living tissues to injury. It is characterized by redness, heat, swelling, pain and loss of function. Redness and heat are the manifestations of increased circulation resulting from vasodilation. Swelling results from collection of protein-rich exudates because capillaries and venules become leaky to protein due to vasodilation. Chemical products formed after injury produce pain. When microorganisms breach local defences at skin and mucosal surface, systemic reactions are set off to destroy the foreign invaders, which result in inflammation. Inflammation mainly stems from the effects of mediators involved in the body's defence mechanism[4,5]. Immediately after injury, the white blood cells rush to the site of injury to protect the