Sankhyā : The Indian Journal of Statistics
1974, Volume 36, Series B, Pt. 4, pp. 323–347.

A STATISTICAL STUDY OF WORD-LENGTH IN BENGALI PROSE

By N. BHATTACHARYA
Indian Statistical Institute[1]

SUMMARY. This paper presents the distributions of words by length in syllables estimated for 28 works in Bengali prose, mostly fiction, covering roughly the period from 1850 to 1950, and also for *a few* short stories, short essays and poems.

A method of probability sampling of words is devised and the sampling properties of estimators obtained are investigated. Non-probabilistic systematic samples of words are also drawn from many works: Statistical tests show the approximate equivalence of these samples and the probability samples. (This is because the series of word-lengths is nearly random, which will be demonstrated in a subsequent paper.) The technique of interpenetrating subsamples is often used for assessing the sampling errors of estimates.

The word-length distributions reveal historical trends in the average word-length and give dimensional ideas of word-length in different fields of literature. Appreciable and significant differences are sometimes found between similar works by the same author, pointing to the limitations of word-length data for 'literary blood tests'. A classification of syllables is considered for improving upon the number of syllables as a measure of word-length. Finally, the form of the word-length distributions is examined *vis-a-vis* Poisson and lognormal hypotheses.

1. Introduction

Word-length, like sentence-length, is one of the obvious indicators of literary style. Word-length distributions have been used in problems of disputed authorship (Williams, 1956; Brinegar, 1963) and for comparisons between languages, between different fields of writing and between authors writing in different styles (Elderton, 1949; Fucks, 1952, 1955; Herdan, 1956; Oettinger, 1954). Word-length is one of the components of many statistical indices of readability (Flesch, 1946).

The object here is to present some word-length data for a number of works in Bengali prose, covering, roughly speaking, the period from 1850 to 1950. Word-length is measured in syllables. The data reveal the declining trends in the average of word-length, and give dimensional ideas of word-length in different types of works. They also show some considerable and statistically significant differences between similar works by the same author, pointing to the dangers of taking word-length as a characteristic of individual style.

[1]Based on a dissertation for Ph.D. degree submitted to the Indian Statistical Institute in 1965. The author is indebted to R. N. Mukherjee, R. D. Chatterjee, P. N. Bhattacharya and (late) A. K. Sengupta for statistical assistance. He also wishes to thank the referee of Sankhyā for making numerous suggestions which improved the presentation.

Statistical studies on languages have often been based on non-probabilistic samples, if not on complete enumeration (*vide* Yule, 1938; Elderton, 1949; Williams, 1940, 1956). One of the features of the present study is the use of probability samples and of rigorous methods of statistical inference.[2] We first describe the method of probability sampling adopted for drawing the samples of words and investigate the sampling properties of estimates based on such samples. We have also employed non-probabilistic systematic samples of words and demonstrated the validity of such samples as approximations to probability samples. (This is essentially due to the approximate randomness of the series of word-lengths, which will be demonstrated in a subsequent paper.) The technique of independent and interpenetrating subsamples has been often used for assessing sampling errors.

Section 2 is concerned with the method of probability sampling and Section 3 with that of systematic sampling. The word-length data for the prose works—distributions, averages etc.—are presented in Section 4, where the data are discussed with emphasis on historical trends and on variation between works by the same author. A classification of Bengali syllables is considered in Section 5 to throw light on the number of syllables as a measure of word-length. Section 6 examines how well the word-length distributions can be fitted with Poisson(Fucks, 1955) or lognormal (Williams, 1956; Herdan, 1958) laws. Section 7 concludes the paper with a hurried look at some Bengali poetry.

2. The probability samples

For each chosen prose work, some desired number of lines, say 100 or 200, were selected by simple random sampling with replacement (srswr) and all words occurring on all the sample lines together formed the probability sample of words from the work. A hyphenated word occurring partly on a sample line was wholly included (excluded) if it occurred at the end (beginning) of the sample line. Such samples may be regarded as cluster samples, lines acting as clusters. We propose to show in a subsequent communication that the series of word-lengths is not far from random, so that our method of sampling is approximately equivalent to srswr.[3]

Let n_i be the number of words on the i-th randomly selected line from a work ($i = 1, 2, ..., k$), $n_i^{(r)}$ the number, out of these, of r-syllabled words

[2] *Vide* Ross (1950) for a strong criticism of the sampling approach.

[3] Actually, it would have been better to use clusters of several consecutive lines as sampling units.

$(r = 1, 2, \ldots)$ and x_{ij} the length in syllables of the j-th word on the i-th sample line $(j = 1, 2, \ldots, n_i)$. We are mostly concerned with ratio estimates

$$p_r = \sum_i n_i^{(r)} / \sum_i n_i \quad \text{and} \quad \bar{x} = \sum_i \sum_j x_{ij} / \sum_i n_i$$

where $\sum_i$ denotes summation over the k sample lines (clusters).

Such ratio estimates of the form $R = \sum_i z_i / \sum_i y_i$ based on cluster samples are known to be consistent, though generally biased; the bias vanishes if the regression of z on y is a straight line passing through the origin. If $k > 30$ and if further both the sample means $\bar{z}$ and $\bar{y}$ have C.V. less than 0.1 (10 per cent), then one may reasonably assume that R is approximately normally distributed with negligible bias and may also estimate its sampling variance from the expression

$$\hat{V}(R) = \frac{1}{k(k-1)\bar{y}^2} \sum_{i=1}^{k} (z_i - Ry_i)^2 \qquad \ldots \quad (1)$$

(*vide* Cochran, 1963, Chaps. 6 and 9).

In the present case, the regressions of $n_i^{(r)}$ or $\sum_j x_{ij}$ on n_i resemble straight lines passing through the origin. There is in fact direct evidence to justify the use of the large sample results. First, k is at least about 100 for the probability samples from all the works. Second, so far as the $\bar{x}$'s are concerned, the conditions regarding the C.V.'s of sample means of $\sum_j x_{ij}$ and n_i are satisfied for all the 24 works from which probability samples were drawn. The two C.V.'s were nearly equal, in general, and ranged from 1 per cent to 4 per cent, roughly speaking. As regards the p_r's, the C.V.'s of sample averages of the $n_i^{(r)}$ appeared to be less than 10 per cent for $r = 1, 2, 3$ and perhaps 4, but not for the larger values of r. Hence the large sample properties may be assumed for p_1, p_2, p_3 and perhaps p_4, but not for p_5, p_6, etc.

The sample of k randomly selected lines was split into 4 independent and interpenetrating subsamples (SS) : SS 1 comprised sample lines numbered 1, 2, ..., $k/4$, in the order of selection; SS 2 those numbered $\frac{k}{4}+1, \ldots, \frac{k}{2}$; and so on. The estimates $\bar{x}$ and $p_1, p_2, \ldots$ were obtained separately for each subsample and also for the combined probability sample from each work.[4]

The subsample estimates have the same ratio form, but are based on $\frac{k}{4}$ clusters, and $\frac{k}{4}$ was as low as 25 in certain cases. However, the C.V.'s

[4] Occasionally 8 or 10 subsamples were used for certain analyses for a number of works.

of subsample means of $\sum_j x_{ij}$ and n_i were also less than 10 per cent, so that the subsample x's seem to possess the large sample properties. But the condition regarding C.V.'s was not fulfilled even for $p_1, p_2, \ldots$ except for works for which k was about 200.

That even the subsample estimates are nearly unbiased was seen from the differences between the *simple* averages of the subsample estimates $p_1, p_2, \ldots, p_4$ and $\bar{x}$, and the corresponding combined sample estimates. Since the bias of a ratio estimator based on k observation-pairs is of the order $\frac{1}{k}$, such comparisons reveal the extent of bias of the subsample and the combined estimates (Murthy and Nanjamma, 1959). The differences were found to be very small. For x, the difference is usually less than 0.1 in percentage terms, and the largest difference is only 0.73 per cent (*vide* Table 2).

3. The systematic samples

For many works, non-probabilistic systematic samples were drawn instead of, or in addition to, the probability samples. Numerical rules were followed for the selection : thus, we took the second line from the bottom of every alternate page or the fourth line from top of every third page. More than one such rule was often used for sampling from a given work. All words falling on the selected lines constituted the systematic sample[6] of words. No use was made of any kind of random start. In theory, the use of such samples is open to serious criticism, but they appeared to be equivalent to probability samples, to a close approximation.

For the short essays and stories shown at the end of Table 2, we took every 3rd or 10th line (say) in the systematic sample.

The lines constituting the systematic sample from any work were divided into 4 interpenetrating subsamples (SS). Suppose the sample lines are numbered 1, 2, 3, ... according to the position in the natural reading order. Then SS 1 comprises lines numberd 1, 5, 9, ... ; SS 2, those numbered 2, 6, 10, ... ; and so on. Estimates were prepared separately for the subsamples as well as for the combined sample.

Wherever both types of sample were taken, the systematic and the probablity samples were pooled to get over-all estimates for a given work.

[6]We use the term 'systematic' even though the intervals between successive lines vary to some extent (*vide* Cochran, 1963, p. 206). For the sampling fractions used for most of the works, the use of a fixed interval, say 40, between successive sample lines would have been fairly time-consuming.

Strictly speaking, one cannot think of sampling errors of estimates based on such non-probabilistic samples, but our finding that the series of word-lengths is nearly random encouraged us to take a 'practical' view and assess sampling errors of systematic sample estimates by the divergence among the subsample estimates (Cochran, 1963, Chap. 8). One may imagine that the whole work is divided into a number of strata and each subsample includes one line from each stratum.

The broad agreement between probability samples and systematic samples was evident from the distributions and averages (*vide* Tables 2 and 3) and from fractile graphs (Mahalanobis, 1960) for the distributions based on the two types of samples. We, however, established the validity of the systematic samples in a more objective manner. There were considerable discrepancies between the two types of samples for individual works like '*Rājsimha*'. We needed objective *over-all* tests for deciding whether the frequencies of large and small deviations between the two types of samples are such as could be expected to occur by chance. We also wanted to see whether the sampling errors of the two sets of estimates of $\bar{x}$ or p_r ($r = 1, 2, \ldots$) are nearly equal, apart from differences in the respective sample sizes. The four series of tests carried out for this purpose are summarised in Tables 1(a) and 1(b).

For each work from which a probability sample was drawn, the χ^2-test of homogeneity was applied for comparing the subsamplewise distributions of word-length in syllables. The results are shown in cols. (2)-(4) of Table 1(a). Similar tests for the systematic sample are summarised in cols. (5)-(6) of the same table. The χ^2's in cols. (4) and (6) are mostly non-significant and the P-values fairly spread over the interval (0, 1). But the sum of the χ^2's in col. (4) is nearly significant ($P = 0.08$) and that of the χ^2's in col. (6) is highly significant ($P = 0.009$). So the χ^2's in cols. (4) and (6) seem to have some upward bias.

The third series of homogeneity χ^2-tests, covered in cols. (7)-(8) of Table 1(a), was applied for comparing the word-length distributions from the probability sample and the systematic sample from the same work. No χ^2-value reaches even the 30 per cent level and the sum of the 14 χ^2's has a P-value = 0.953. So there is significant evidence that these χ^2's tended to be on the low side.

The tests summarised in Table 1(b) compare the variability of the four subsample averages $\bar{x}_1'$, $\bar{x}_2'$, $\bar{x}_3'$ and $\bar{x}_4'$ based on the systematic sample and the variability of the average from the (combined) probability sample from the work, eliminating the effect of differences in sample size measured by the

TABLE 1(a): RESULTS OF χ^2-TESTS FOR HOMOGENEITY OF DIFFERENT DISTRIBUTIONS OF WORD-LENGTH

work	subsamples of probability samples			4 subsamples of systematic samples		probability and systematic samples	
	no. of SS	d.f.	χ^2	d.f.	χ^2	d.f.	χ^2
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1. *Śhakuntalā*	4	12	18.509				
2. *Sītār Vanavās*	4	15	14.140				
3. *Durgeśhnandinī*	4	12	17.135	15	15.964	5	3.291
4. *Kapālkuṇḍalā*				12	14.809		
5. *Viṣavṛksha*	4	12	12.623	15	12.611	5	1.702
6. *Kṛṣhṇakānter Will*	8	28	39.779	12	8.354	5	5.072
7. *Ānandamaṭh*	8	28	21.874	12	14.015	5	4.832
8. *Devī Chaudhurānī*	8	21	29.088	12	19.126	4	2.165
9. *Rājsimha*	10	36	35.200	12	11.139	5	5.867
10. *Bauṭhākurāṇīr Hāṭ*	8	28	20.443	12	16.196	4	3.573
11. *Rājarṣi*	8	28	23.226	12	20.995	5	3.339
12. *Chokher Bāli*				12	11.713		
13. *Gorā*	4	12	16.335	12	21.724*	4	1.440
14. *Chaturaṅga*	8	28	41.812*	12	13.856	4	1.347
15. *Ghare Bāire*	8	21	26.194				
16. *Śheṣer Kavitā*	4	12	18.621	12	11.052	5	5.674
17. *Yogāyog*				12	7.957		
18. *Chār-Yārī Kathā*	4	12	13.577				
19. *Bīrbaler Hālkhātā*	4	12	7.282				
20. *Pallīsamāj*	4	12	10.153				
21. *Pather Dābī*	4	9	14.462				
22. *Pather Pānchālī*	4	12	13.138	12	18.900	4	2.917
23. *Aparājita*				12	13.800		
24. *Devayān*	4	12	14.703	12	12.473	4	1.436
25. *Dṛṣṭipāt*	4	12	10.581	15	15.122	5	4.207
26. *Janāntik*	4	12	16.808				
27. *Chāchā Kāhinī*	4	9	7.152				
28. *Deśhe Videśhe*	4	12	5.589				
29. Sub-total (1–28)		407	448.484	225	259.808	64	47.012
30. *Sāmya*				15	20.140		
31. *Bankimchandra*				15	34.091**		
32. *Viśhvavidyālay*				12	18.869		
33. *Kābuliwālā*				12	5.200		
34. *Kṣhudhita Pāṣāṇ*				15	12.750		
35. *Laboratory*				12	17.152		
36. Sub-total (30-35)				81	108.202		
37. Total (29+36)		407	448.484	306	368.010	64	47.012

N.B.: (1) Single asterisk (*) denotes significance at 5 per cent level and double asterisk (**) significance at 1 per cent level.

(2) Systematic sampling was slightly different in the two subsets of works (*vide* text).

TABLE 1(b). RESULTS OF χ^2-TEST FOR COMPARING THE VARIABILITY OF THE SUBSAMPLE AVERAGES OF WORD-LENGTH FROM THE SYSTEMATIC SAMPLE WITH THAT OF THE COMBINED SAMPLE AVERAGE FROM THE PROBABILITY SAMPLE, AFTER ADJUSTING FOR DIFFERENCES IN SAMPLE SIZE

work	no. of sample words		χ^2 (3 d.f.)	P = upper tail probability
	prob. sample	syst. sample		
(1)	(2)	(3)	(4)	(5)
1. *Durgeśhnandinī*	577	1782	0.009	>0.90
2. *Viṣavṛkṣha*	611	1852	1.703	0.50–0.70
3. *Kṛṣhṇakānter Will*	1777	740	0.698	0.80–0.90
4. *Ānandamaṭh*	1109	801	8.294	0.02–0.05
5. *Devī Chaudhurānī*	1174	833	5.943	0.10–0.20
6. *Rājsimha*	1423	507	3.275	0.30–0.50
7. *Bauṭhākurāṇīr Hāṭ*	1592	827	6.103	0.10–0.20
8. *Rājarṣi*	1632	680	4.525	0.20–0.30
9. *Gorā*	880	1824	1.077	0.50–0.70
10. *Chaturaṅga*	1458	854	0.903	0.80–0.90
11. *Sheṣer Kavitā*	735	1284	2.724	0.30–0.50
12. *Pather Pānchālī*	922	1630	2.776	0.30–0.50
13. *Devayān*	931	2245	1.491	0.50–0.70
14. *Dṛṣṭipāt*	772	1501	5.712	0.10–0.20
15. Total	—	—	46.313 (42 d.f.)	0.20–0.30

number of words. We assume that the $\bar{x}_i'$'s are independently and normally distributed, and that the sampling variances of such averages from both types of samples are inversely proportional to the sample size with the constant of proportionality the same for the two types of samples. We then see that

$$\chi^2 = \frac{n'}{4n} \sum_{i=1}^{4} (\bar{x}_i' - \bar{x}')^2 / \text{est. } V(\bar{x}) \qquad \ldots \quad (2)$$

would be approximately distributed as χ^2 with 3 d.f. Here n, n' are the sample sizes of the probability and the systematic samples, $\bar{x}$, $\bar{x}'$ the respective combined sample averages, and $V(\bar{x})$ is estimated from eqn. (1) so that it may be taken as nearly exact. The χ^2-values are shown in col. (4) of Table 1(b). The P-values are well-spread over the interval (0, 1) and the total of the χ^2's has a P-value around 30 per cent.

We may now briefly consider the interpretation of these results. The χ^2-tests for homogeneity assume srswr, but both types of samples involve the use of line-clusters, and the lengths of neighbouring words show some positive auto-correlation, which though small, is significant. This is why both types of samples have slightly larger sampling errors than a srswr of equal size, and the same holds for subsamples of these samples. This explains the small upward bias in the χ^2's in cols. (4) and (6) of Table 1(a). Actually, the sub-samples of the systematic samples seem to be just as variable as probability samples of the same size. This is particularly clear from the tests reported in Table 1(b).

The downward bias in the χ^2's in col. (8) of Table 1(a) may be explained in the following (tentative) manner: The series of word-lengths is not perfectly random, but there are relatively homogeneous 'patches', differing from one another in respect of the average of word-length. A subsample of a systematic sample may miss many of the patches altogether, but the combined systematic sample may sample most of the patches. So while the subsamples of the systematic sample may be as reliable as probability samples of the same size, the combined systematic sample may be slightly more reliable than a probability sample of equal size. In other words, the subsamples of the systematic sample may slightly exaggerate the true sampling errors of the combined systematic sample.

4. The word-length distributions

Tables 2 and 3 present the estimates for the prose works. Most of the works are novels of different types. The two works by Vidyasagar are free renderings of classical Sanskrit works. *'Chār-Yāri Kathā'* is a string of four short stories; *'Chāchā-Kāhinī'* is also a collection of short stories. *'Dṛṣṭipāt'* and *'Deśhe Videśhe'* come under *belles lettres*, *'Bīrbaler Hālkhātā'* is a collection of essays. Some of the works represent landmarks in the history of the Bengali language/literature. Thus, the earliest included, *'Śhakuntalā'* (1854) was the first work of art in Bengali prose. Emphasis has been given to the works of Bankimchandra and Tagore, the two greatest makers of Bengali prose. Muztaba Ali and 'Jajabar' were included as representatives of certain trends in recent literature.

2

TABLE AVERAGES AND STANDARD DEVIATIONS OF WORD-LENGTH IN SYLLABLES, ESTIMATED FOR DIFFERENT WORKS IN BENGALI PROSE, SEPARATELY BY TYPE OF SAMPLE AND BY SUBSAMPLES (FOR AVERAGE ONLY)

author	work	type of sample	sample size: no. of lines	sample size: no. of words	average word-length (syllables) by subsamples: SS 1	SS 2	SS 3	SS 4	comb.	s.e. of comb. average	s.d.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Vidyasagar	*Śhakuntalā*	prob.	100	696	2.841	2.475	2.866	2.658	2.704	0.0392	1.103
	Sītār Vanavās	prob.	100	750	2.632	2.665	2.608	2.821	2.695	0.0453	1.218
Bankimchandra	*Durgeśhnandinī*	prob.	100	577	2.682	2.463	2.814	2.392	2.579	0.0599	1.127
		syst.	316	1782	2.593	2.593	2.592	2.585	2.591		1.095
		pooled	416	2359	2.614	2.559	2.644	2.537	2.588	0.0296	1.102
	Kapālkuṇḍalā	syst.	90	493	2.788	2.600	2.662	2.556	2.645		1.229
	Viṣavṛkṣha	prob.	99	611	2.404	2.534	2.531	2.419	2.470	0.0483	1.057
		syst.	300	1852	2.457	2.409	2.500	2.397	2.455		1.071
		pooled	399	2463	2.445	2.485	2.508	2.403	2.459	0.0241	1.068
	Kṛṣhṇakānter Will	prob.	200	1777	2.330	2.281	2.308	2.378	2.340	0.0280	1.010
		syst.	128	749	2.370	2.302	2.415	2.318	2.372		1.080
		pooled	328	2526	2.342	2.316	2.379	2.360	2.350	0.0235	1.031
	Ānandamaṭh	prob.	200	1109	2.395	2.442	2.496	2.440	2.443	0.0313	1.020
		syst.	133	801	2.508	2.510	2.419	2.245	2.438		1.051
		pooled	333	1910	2.441	2.470	2.465	2.352	2.441	0.0238	1.033

TABLE 2. (*contd.*) AVERAGES AND STANDARD DEVIATIONS OF WORD LENGTH IN SYLLABLES, ESTIMATED FOR DIFFERENT WORKS IN BENGALI PROSE, SEPARATELY BY TYPE OF SAMPLE AND BY SUBSAMPLES (FOR AVERAGE ONLY)

author	work	type of sample	sample size		average word-length (syllables) by subsamples					s.e. of comb. average	s.d.
			no. of lines	no. of words	SS 1	SS 2	SS 3	SS 4	comb.		
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Bankimchandra	*Devī Chaudhurāṇī*	prob.	200	1174	2.353	2.219	2.189	2.378	2.283	0.0298	0.928
		syst.	146	833	2.320	2.251	2.000	2.269	2.227		0.840
		pooled	346	2007	2.339	2.232	2.147	2.333	2.260	0.0228	0.892
	Rājsimha	prob.	250	1423	2.545	2.467	2.434	2.486	2.482	0.0279	1.078
		syst.	90	507	2.715	2.634	2.512	2.520	2.596		1.173
		pooled	340	1930	2.593	2.509	2.455	2.495	2.512	0.0239	1.104
Rabindranath	*Bauṭhākurāṇīr Hāṭ*	prob.	200	1592	2.358	2.300	2.409	2.421	2.386	0.0208	0.970
		syst.	100	827	2.498	2.392	2.266	2.484	2.406		0.906
		pooled	300	2419	2.406	2.371	2.359	2.442	2.393	0.0217	0.909
	Rājarṣi	prob.	200	1632	2.394	2.419	2.432	2.449	2.424	0.0218	0.991
		syst.	88	689	2.454	2.546	2.358	2.513	2.467		0.958
		pooled	288	2321	2.412	2.458	2.410	2.407	2.437	0.0183	0.981
	Chokher Bāli	syst.	156	1318	2.329	2.458	2.306	2.322	2.366		0.910
	Gorā	prob.	100	889	2.291	2.353	2.417	2.268	2.331	0.0362	0.905
		syst.	203	1824	2.360	2.374	2.359	2.283	2.345		0.947
		pooled	303	2713	2.330	2.367	2.377	2.278	2.341	0.0207	0.932
	Ghare Bāire	prob.	200	1901	2.047	2.088	2.141	2.102	2.093	0.0211	0.834
	Chaturaṅga	prob.	200	1458	2.452	2.261	2.254	2.349	2.326	0.0248	0.915
		syst.	113	854	2.347	2.269	2.293	2.274	2.296		0.905
		pooled	313	2312	2.411	2.264	2.268	2.323	2.315	0.0197	0.912

TABLE 2. (*contd.*) AVERAGES AND STANDARD DEVIATIONS OF WORD-LENGTH IN SYLLABLES, ESTIMATED FOR DIFFERENT WORKS IN BENGALI PROSE, SEPARATELY BY TYPE OF SAMPLE AND BY SUBSAMPLES (FOR AVERAGE ONLY)

author	work	type of sample	sample size		average word-length (syllables) by subsamples					s.e. of comb. average	s.d.
			no. of lines	no. of words	SS 1	SS 2	SS 3	SS 4	comb.		
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Rabindranath	*Śheṣer Kavitā*	prob.	100	735	2.122	2.332	2.173	2.125	2.186	0.0382	0.937
		syst.	181	1284	2.267	2.150	2.222	2.160	2.204		0.910
		pooled	281	2019	2.212	2.213	2.204	2.147	2.198	0.0231	0.920
	Yogāyog	syst.	142	1187	2.130	2.137	2.252	2.157	2.168		0.913
Pramatha Choudhury	*Chār-Yārī Kathā*	prob.	100	872	2.049	2.014	2.050	2.124	2.060	0.0273	0.856
	Birbaler Hālkhātā	prob.	100	1041	2.249	2.355	2.314	2.331	2.311	0.0410	1.076
Saratchandra	*Pallisamāj*	prob.	100	890	2.225	2.171	2.205	2.192	2.212	0.0353	0.928
	Pather Dābī	prob.	100	815	2.291	2.300	2.183	2.133	2.228	0.0300	0.890
Bibhutibhusan	*Pather Pānchālī*	prob.	100	922	2.284	2.310	2.206	2.202	2.250	0.0334	0.913
		syst.	172	1630	2.291	2.211	2.302	2.320	2.279		0.901
		pooled	272	2552	2.289	2.246	2.268	2.275	2.260	0.0200	0.906
	Aparājita	syst.	201	1894	2.225	2.248	2.290	2.333	2.273		0.941
	Devayān	prob.	100	931	2.172	2.140	2.059	2.134	2.126	0.0326	0.857
		syst.	244	2245	2.107	2.172	2.131	2.160	2.142		0.851
		pooled	344	3176	2.189	2.162	2.110	2.152	2.138	0.0176	0.853

TABLE 2. (*contd.*) AVERAGES AND STANDARD DEVIATIONS OF WORD-LENGTH IN SYLLABLES, ESTIMATED FOR DIFFERENT WORKS IN BENGALI PROSE, SEPARATELY BY TYPE OF SAMPLE AND BY SUBSAMPLES (FOR AVERAGE ONLY)

author	work	type of sample	sample size		average word-length (syllables) by subsamples					s.e. of comb. average	s.d.
			no. of lines	no. of words	SS 1	SS 2	SS 3	SS 4	comb.		
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Jajabar	*Dṛṣṭipāt*	prob.	100	772	2.394	2.412	2.367	2.382	2.389	0.0395	1.029
		syst.	213	1591	2.400	2.425	2.293	2.472	2.398		1.041
		pooled	313	2363	2.398	2.421	2.318	2.443	2.395	0.0226	1.037
	Janāntik	prob.	100	690	2.368	2.098	2.357	2.364	2.293	0.0421	0.942
Muztaba Ali	*Chāchā Kāhinī*	prob.	100	778	2.222	2.234	2.092	2.206	2.189	0.0353	0.842
	Deśhe Videśhe	prob.	100	791	2.139	2.214	2.218	2.215	2.172	0.0316	0.867
Bankimchandra	*Sāmya*	syst.	114	1010	2.429	2.605	2.686	2.788	2.619		1.252
Rabindranath	*Bankimchandra*	syst.	139	1237	2.767	2.730	2.590	2.642	2.682		1.162
	Viśhwavidyālay	syst.	103	1009	2.339	2.271	2.458	2.296	2.339		0.988
	Kābuliwālā	syst.	86	779	2.439	2.503	2.539	2.523	2.501		1.024
	Kshudhita Pāṣāṇ	syst.	125	1102	2.456	2.601	2.505	2.537	2.524		1.054
	Laboratory	syst.	131	1228	2.192	2.106	2.115	2.108	2.131		0.886

TABLE 3: DISTRIBUTION OF WORDS BY LENGTH IN SYLLABLES, ESTIMATED FOR DIFFERENT WORKS IN BENGALI PROSE, SEPARATELY BY TYPE OF SAMPLE

author	work	type of sample	no. of sample words	percentage of words by length in syllables								
				1	2	3	4	5	6	7	8	9*
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Vidyasagar	*Śhakuntalā*	prob.	690	11.64	34.63	33.76	13.65	4.45	1.73	0.14		
	Sītār Vanavās	prob.	750	15.07	31.33	34.27	11.47	5.47	1.60	0.53	0.13	0.13(10)
Bankimchandra	*Durgeśhnandinī*	prob.	577	15.42	35.53	33.11	10.05	4.16	1.21	0.35	0.17	
		syst.	1782	13.58	37.60	32.21	11.82	3.54	0.79	0.56	0.11	
		pooled	2359	14.03	37.09	32.43	11.23	3.69	0.89	0.51	0.13	
	Kapālkuṇḍalā	syst.	493	15.82	36.11	27.99	10.95	6.69	2.03	0.20	0.20	
	Viṣavṛkṣha	prob.	611	16.20	40.59	28.81	10.15	3.27	0.65	0.33		
		syst.	1852	15.77	43.20	27.16	9.29	3.40	0.76	0.32	0.05	0.05
		pooled	2463	15.87	42.55	27.57	9.60	3.37	0.73	0.32	0.04	0.04
	Kṛṣhṇakānter Will	prob.	1777	19.19	42.77	27.01	7.77	2.76	0.28	0.17		0.06
		syst.	749	20.69	38.99	28.57	7.48	3.34	0.40	0.40	0.13	
		pooled	2526	19.64	41.65	27.47	7.68	2.93	0.32	0.24	0.04	0.04
	Ānandamaṭh	prob.	1109	16.23	40.67	30.03	10.10	1.98	0.63	0.30		
		syst.	801	16.85	42.32	27.59	8.86	3.12	0.62	0.25	0.37	
		pooled	1910	16.49	41.36	29.01	9.58	2.46	0.63	0.31	0.16	
	Devī Chaudhurāṇī	prob.	1174	18.99	44.12	28.88	6.22	1.45	0.20			0.09
		syst.	833	21.01	44.30	27.73	5.16	1.68		0.12		
		pooled	2007	19.83	44.20	28.40	5.78	1.54	0.15	0.05		0.05
	Rājsimha	prob.	1423	16.44	40.20	28.11	10.61	3.51	0.84	0.28		
		syst.	507	14.20	38.86	29.98	10.85	3.75	1.58	0.39	0.20	0.20
		pooled	1930	15.85	39.84	28.60	10.67	3.58	1.04	0.31	0.05	0.05
Rabindranath	*Bauṭhākurāṇir Hāṭ*	prob.	1592	16.90	41.71	31.09	6.85	3.08	0.38			
		syst.	827	16.81	39.06	34.70	6.17	2.66	0.60			
		pooled	2419	16.87	40.80	32.33	6.61	2.94	0.45			

* The numbers inside brackets show the actual lengths of the words in syllables in case the length exceeds 9.

TABLE 3 (*contd.*): DISTRIBUTION OF WORDS BY LENGTH IN SYLLABLES, ESTIMATED FOR DIFFERENT WORKS IN BENGALI PROSE, SEPARATELY BY TYPE OF SAMPLE

author	work	type of sample	no. of sample words	percentage of words by length in syllables								
				1	2	3	4	5	6	7	8	9*
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Rabindranath	*Rājarṣi*	prob.	1632	14.05	43.32	30.39	8.27	2.27	0.55	0.18		0.06(10)
		syst.	689	13.35	41.80	33.24	9.00	1.74	0.73	0.15		
		pooled	2321	14.48	42.87	31.24	8.49	2.11	0.60	0.17		0.04(10)
	Chokher Bāli	syst.	1318	15.25	44.60	30.42	7.89	1.37	0.30	0.08		
	Gorā	prob.	889	15.52	47.59	27.78	6.64	2.36	0.11			
		syst.	1624	15.84	47.42	26.37	7.62	2.36	0.27	0.05	0.05	
		pooled	2713	15.74	47.48	26.83	7.30	2.30	0.22	0.04	0.04	
	Chaturaṅga	prob.	1458	16.32	45.75	29.63	6.10	1.85	0.14	0.21		
		syst.	854	17.68	45.55	28.34	6.56	1.64	0.23			
		pooled	2312	16.83	45.67	29.15	6.27	1.77	0.17	0.13		
	Ghare Bāire	prob.	1901	20.67	57.23	15.94	4.79	1.10	0.21	0.05		
	Sheser Kavitā	prob.	735	19.05	54.28	18.91	5.44	2.04	0.14			0.14(11)
		syst.	1284	18.22	53.97	19.86	6.07	1.09	0.62	0.08	0.08	
		pooled	2019	18.52	54.09	19.51	5.84	1.44	0.45	0.05	0.05	0.05(11)
	Yogāyog	syst.	1187	19.63	55.52	16.01	6.66	1.68	0.42	0.08		
Pramatha Choudhury	*Chār-Yārī Kathā*	prob.	872	22.71	56.88	14.22	4.82	0.82	0.46	0.11		
	Birbaler Hālkhātā	prob.	1041	21.13	43.81	23.62	6.92	3.07	0.67	0.38		0.10
Saratchandra	*Pallīsamāj*	prob.	890	20.23	48.20	25.06	3.93	2.14	0.22	0.11	0.11	
	Pather Dābī	prob.	815	20.37	44.79	28.10	5.52	0.98	0.12	0.12		
Bibhutibhusan	*Pather Pānchālī*	prob.	922	18.55	49.13	23.54	6.29	2.40				
		syst.	1630	16.93	49.02	25.58	6.56	1.53	0.31	0.06		
		pooled	2552	17.52	49.06	24.84	6.47	1.88	0.20	0.04		
	Aparājita	syst.	1894	18.32	47.52	25.02	5.65	2.10	0.32	0.05		0.05(10)
	Devayān	prob.	931	20.30	56.07	16.00	6.12	1.40	0.11			
		syst.	2245	19.33	55.90	17.55	5.88	1.11	0.18	0.04		
		pooled	3176	19.62	55.95	17.10	5.95	1.20	0.16	0.03		

The numbers inside brackets show the actual lengths of the words in syllables in case the length exceeds 9.

TABLE 3 (*contd.*): DISTRIBUTION OF WORDS BY LENGTH IN SYLLABLES, ESTIMATED FOR DIFFERENT WORKS IN BENGALI PROSE, SEPARATELY BY TYPE OF SAMPLE

author	work	type of sample	no. of sample words	percentage of words by length in syllables								
				1	2	3	4	5	6	7	8	9*
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Jajabar	*Dṛṣṭipāt*	prob.	772	17.23	44.69	24.35	10.36	2.98		0.26	0.13	
		syst.	1591	16.40	46.20	23.95	9.37	2.95	1.01	0.06	0.06	
		pooled	2363	16.67	45.70	24.08	9.69	2.96	0.68	0.13	0.08	
	Janāntik	prob.	690	15.94	52.75	20.87	7.68	2.03	0.73			
Muztaba Ali	*Chāchā Kāhinī*	prob.	778	16.71	56.29	20.31	5.14	1.29	0.13	0.13		
	Deśhe Videśhe	prob.	791	18.46	55.62	18.08	6.19	1.52		0.13		
Bankimchandra	*Sāmya*	syst.	1010	15.64	39.40	24.06	13.27	5.15	1.49	0.69	0.10	0.20(9,10)
Rabindranath	*Bankimchandra*	syst.	1237	11.48	39.29	29.43	11.90	5.90	1.29	0.40	0.08	0.08
	Viśhvavidyālay	syst.	1009	16.15	49.75	22.50	8.23	2.48	0.79	0.10		
	Kābuliwālā	syst.	779	13.61	41.34	32.09	9.11	2.31	1.28	0.26		
	Kṣhudhita Pāṣāṇ	syst.	1192	11.91	43.78	31.04	9.06	2.77	0.92	0.09	0.34	0.09(10)
	Laboratory	syst.	1228	20.00	55.29	17.02	5.21	1.38	0.33	0.16		

* The numbers inside brackets show the actual lengths of the words in syllables in case the length exceeds 9.

In the last six rows in both tables, we cover three short essays and three short stories.

Words were taken *as printed*, demarcated from one another by spaces, and no attempt was made to count compounds of two words, say, as two words, instead of one. Between works variation in average word-length is partly due to variation in the proportion of compounds. As is well-known, compounds were more frequent in the elevated Sanskritised style of Vidyasagar and Bankimchandra (early phase).

Counting of syllables was based on the standard pronunciation of literary Bengali, which means the modes prevailing in learned circles in and around Calcutta (Chatterjee, 1921). Sometimes an *a* sound ('a' as in English 'fall') seemed to be optional. Such cases were few in the prose works and the older mode of pronouncing it was adopted there.*

The following diphthongs were treated as similar to single vowel sounds in that they form the core of single syllables : ei, eu, aee, aeo, ai ae, ao, au, oe, oo, oe, oi, ou, ui. The remaining diphthongs, viz., ie, ia, io, iu, ea, eo, oa, oa, ue, ua, uo were each considered as two distinct vowels. All triphthongs and higher combinations were split into different syllables on the basis of the rules adopted for diphthongs (*vide* Chatterjee, 1921, pp. 16-17).

In Table 2, the standard errors of the averages were computed for the combined probability samples, using eqn. (1) given earlier : The standard error of the pooled average based on the probability and the systematic samples was obtained by multiplying the s.e. of the average from the probability sample by $\sqrt{\frac{n}{n+n'}}$, where n, n' are the number of words in the probability and the systematic samples respectively.

Historical trends in word-length : The estimates, especially the averages $\bar{x}$, corroborate what is generally known about the historical changes in literary Bengali. Bengali fiction started with $\bar{x}$ around 2.7 in the works of Vidyasagar written in chaste, Sanskritized style (*sādhu bhāsā*), but the average declined sharply during Bankimchandra's period, even though Bankimchandra generally used the chaste style throughout. A striking figure is the average 2.26 for '*Devī Chaudhurānī*' (1884); here the style is almost colloquial, excepting for

*Such cases were more frequent in Bengali poetry, and in each case we had to ascertain which mode of pronunciation was the more appropriate. For poems in '*payār*' and other meters where the vocal drawl is predominant, pronouncing the 'a' sound seemed to be desirable. In any case, our data on word-length and syllable-type in Bengali poetry are partly subjective because of this.

verbs and pronouns, in the conversational passages. There was some further decline in x during Tagore's period, first when the colloquial style began to be used *in the conversational matter* —e.g., in *'Corā'* (1910)—and then when the said style was used throughout, beginning with *'Ghare Bāire'* (1916) where $x = 2.09 \pm 0.02$.

The few figures for essays and short stories also tell the same story. Everywhere the older chaste style employing longer words and compounds has been replaced by the colloquial style using shorter words.

Word-length in different types of works : The works of Vidyasagar and the early works of Bankimchandra show x around 2.6 or 2.7, but in 20th century Bengali fiction the effective range is from 2.1 to 2.4. Historical novels seem to have somewhat higher averages. In the subsequent communication on the randomness of word-length series, we propose to show that words used in conversational passages are shorter, on the average, than words used elsewhere. So the over-all average tends to be lower if the weightage of conversational matter is relatively high, Actually, variation in x between different works can be partly explained by the unequal weightage of conversational matter. For essays containing no conversational passage, the effective range of x seems to be 2.3 to 2.7.

High values of $\bar{x}$ usually indicate the chaste elevated style with a high proportion of *'tatsama'* words (Sanskrit words in unmodified form) and compounds, while a low x is generally associated with the colloquial style with a high proportion of *'tadbhava'* (i.e., Prakrit) words. Whether the verbs and pronouns have the chaste or the colloquial form is of little direct consequence. The average is really low when the colloquial style is used throughout, and not merely in the conversational passages. This happens for works written as thoughts or speeches of the leading character(s).

It appears that any non-trivial work in Bengali will have x in the neighbourhood of 2, at least.

Within author differences : Not only Bankimchandra and Tagore, but others also (e.g. Bibhutibhusan) show appreciable and statistically significant variation in x between different works written by them. This is a major finding, although in a negative sense. Some statistical investigations on western languages have created the impression that statistical style measures, based on word-length, sentence-length, size and diversity of vocabulary, etc., can be used for characterizing *individual* style (Yule, 1938, 1944 ; Fucks, 1952 ; Williams, 1956). But the situation seems to be different for Bengali prose. This may be partly because Bengali prose was changing fast between 1850 and 1925 (broadly speaking) which was its formative period.

Studies on Plato's works and also Shakespeare's show that an author's style can vary with his age. One can also expect that an author will vary his style when writing in different fields of literature. But there are instances in Tables 2 and 3 where the word-length distribution varies erratically between similar works written by an author at not too distant dates. One may, for example, compare '*Visavrksha*' and '*Krshnakānter Will*', or '*Ānandamath*' and '*Devī Chaudhurānī*', or '*Chaturanga*' and '*Ghare Bāire*', or '*Pather Pānchālī*', '*Aparājita* and '*Devayān*'..

5. A CLASSIFICATION OF BENGALI SYLLABLES

Bengali syllables vary sufficiently in respect of length to make the number of syllables an inadequate measure of word-length. One may recognize two relatively homogeneous types among Bengali syllables if one is interested in their length. These types are defined below :

type		*definition*[7]	*illustration*
A (i.e. short)		open syllables without diphthongs	o, mā, khā, srā
B (i.e. long)	B_1	closed syllables	an, nun, snān, bāng, āik
	B_2	open syllables with diphthongs	ai, māe, strai

For certain purposes, type B syllables were further subdivided into types B_1 and B_2.

Generally speaking, type B syllables are longer than type A syllables. For purposes of metric analysis, type B syllables are sometimes assumed to take two *mora* or instants for pronunciation, as against one required by type A syllables (Chatterjee, 1945, pp. 377-8). Thus, instead of saying that the average word-length for a Bengali work is 2.1 syllables, one might say that the average word has (say) 1.4 syllables of type A and 0.7 syllables of type B.

Table 4 shows the percentages of type A syllables estimated for a number of works from the samples described earlier. Large sample properties of ratio estimates may be safely assumed for all these percentages. Most of the percentages lie in the range from 62 to 72, and although some of the differences are statistically significant, the overwhelming impression is one of stability.

[7]Open (vowel ending) and closed (consonant-ending) syllables are defined in Chatterjee (1945, pp. 25, 35).

So the distinction between 'long" and 'short' syllables may be ignored in comparing average word-length in syllables in different works in Bengali prose.

TABLE 4. PERCENTAGES OF SHORT OR TYPE A SYLLABLES ESTIMATED FOR A NUMBER OF WORKS IN BENGALI PROSE

work	type of sample	no. of sample		percentage of type A syllables by subsamples				
		words	syllables	1	2	3	4	comb.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Śhakuntalā	prob.	600	1882	68.3	67.4	71.5	69.0	69.1
**Sītār Vanavās*	,,	750	2021	66.3	67.2	67.3	65.4	66.6
Durgeśhnandinī	,,	577	1488	67.8	64.6	66.4	68.3	66.8
Viṣavṛkṣha	,,	611	1509	71.1	68.1	69.9	68.3	69.3
Gorā	,,	889	2072	70.9	69.4	70.9	66.8	69.4
Śheṣer Kavitā	,,	735	1607	65.4	68.0	68.4	68.0	67.4
Chār-Yārī Kathā	,,	872	1796	66.7	65.0	66.2	64.4	65.6
**Bīrbaler Hālkhātā*	,,	1041	2406	61.8	61.8	60.5	63.4	61.9
**Pallīsamāj*	,,	890	1909	72.4	70.7	70.0	69.3	70.6
Pather Dābī	,,	815	1810	70.4	69.8	69.4	68.2	69.5
Pather Pānchālī	,,	922	2075	73.6	71.2	69.3	69.7	70.9
Devayān	,,	931	1979	67.6	66.7	70.7	71.4	69.1
Dṛṣṭipāt	,,	772	1844	62.4	65.8	64.5	60.2	63.3
**Janāntik*	,,	600	1582	64.8	62.2	67.7	60.2	63.7
Chūchā-Kāhinī	,,	778	1703	66.7	65.7	65.2	63.8	65.4
Deśhe Videśhe	,,	701	1718	68.8	64.0	62.8	63.8	64.9
**Sāmya*	synt.	1010	2645	66.0	66.0	66.7	64.3	65.8
**Bankimchandra*	,,	1237	3318	63.6	63.0	62.1	62.4	62.8
**Viśhwavidyālay*	,,	1009	2300	66.8	63.9	64.7	62.1	64.4
**Kābuliwālā*	,,	779	1948	68.7	74.0	72.2	71.2	71.6
**Kṣhudhita Pāṣāṇ*	,,	1192	3008	66.7	68.4	70.0	70.8	69.0
**Laboratory*	,,	1228	2617	74.6	68.4	69.4	69.4	70.6

The percentage of type B_2 syllables was of the order of 5 for all the works examined; these works are marked with an asterisk in Table 4.

6. THE FORM OF THE WORD-LENGTH DISTRIBUTION

We considered fitting theoretical distributions to the estimated proportions $p_x(x = 1, 2, ...)$ of words of length x (syllables). Elderton (1949) fitted the geometric distribution to certain distributions like that from Fitzgerald's 'Rubaiyat' of Omar Khayyam. Fucks (1955) stated that $x-1$ is approximately distributed in the Poisson form for eight out of the nine languages examined by him, Arabic being the only exception. The lognormal distribution has been fitted to distributions of English words with word-length measured in terms of letters (Williams, 1956; Herdan, 1958).

Since p_2 is considerably larger than p_1, the geometric law fails completely for Bengali. The Fucks law and the lognormal distribution were tried for the 28 works in Bengali prose (*vide* Table 2)—the short stories and essays were excluded. Only the over-all (combined sample or probability-plus-systematic-samples) distribution was considered and the small deviations from srswr were ignored. The lognormal distribution was fitted in two ways: first (referred to as LN(a)) by supposing that the observed x-values 1, 2, ... represent intervals 0-1, 1-2, ... of the underlying continuous variate (Aitchison and Brown 1957, pp. 92-3), and second (referred to as LN(b)) by supposing that the observed values 1, 2, ... represent intervals 0-1.5, 1.5-2.5, ... etc., of the underlying variate.

We refrain from presenting the estimates of parameters or the fitted distributions of word-length. The goodness of fit was examined by three criteria, χ^2, the Kolmogorov distance K and $D = \sum_x |p_x - \hat{p}_x|$, where $\hat{p}_x$ is the 'fitted' proportion of words of length x. The index D was closely correlated with χ^2/n, where n is sample size, and was employed purely as a descriptive measure. The Kolmogorov test is extremely 'conservative' in the present situation because of the discreteness of the word-length distribution and because parameters have been estimated from sample data. Table 5 shows the results of such examination.

The Poisson fit was generally poor and inferior to the lognormal, except for the older works, e.g., '*Sītār Vanavās*'—The variance of x, S_x^2 is usually less than $\bar{x}-1$ (*vide* Table 2). LN(b) gave a better fit than LN(a) for 20 works out of 28 and the χ^2-test was applied to examine the LN(b) hypothesis.

It must be noted that for the sake of convenience the estimation of the lognormal parameters was not done by a fully efficient method as required for the χ^2-test. We wanted to use the method of quantiles (Aitchison and Brown, 1957, Chap. 5), but various considerations, especially the curvature of the

TABLE 5. GOODNESS OF FIT OF POISSON AND LOGNORMAL DISTRIBUTIONS* TO OBSERVED DISTRIBUTIONS OF WORD-LENGTH IN SYLLABLES, SEPARATELY FOR 28 WORKS IN BENGALI PROSE

work	no. of sample words	$D = \Sigma\|p_x - \hat{p}_x\|$			K (per cent)		χ^2 for $LN(b)$	
		Poisson	$LN(a)$	$LN(b)$	$LN(a)$	$LN(b)$	d.f.	χ^2
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Śhakuntalā	696	0.219	0.200	0.130	6.20**	4.06	4	15.30**
Sitār Vanavās	750	0.163	0.247	0.196	7.46***	5.40*	4	30.07***
Durgeśhnandinī	2359	0.228	0.194	0.124	6.00***	3.86**	4	40.79***
Kāpalkuṇḍalā	493	0.144	0.139	0.088	3.15	1.26	4	8.60
Viṣavṛkṣha	2463	0.230	0.114	0.044	3.19	1.10	4	6.11
Kṛṣhṇakānter Will	2526	0.211	0.160	0.107	4.62***	2.59	4	33.92***
Ānandamath	1910	0.229	0.181	0.111	5.35***	3.20*	4	24.48***
Devī Chaudhurāṇī	2007	0.288	0.221	0.153	6.28***	4.12**	3	48.67***
Rājsiṃha	1930	0.199	0.143	0.070	4.10	2.02	4	10.15*
Bauṭhākurāṇīr Hāṭ	2419	0.289	0.242	0.182	6.63***	4.40***	4	86.89***
Rājarṣi	2321	0.309	0.178	0.108	5.61***	3.30*	4	29.02***
Chokher Bāli	1318	0.329	0.208	0.133	6.42***	4.06*	3	23.04***
Gorā	2713	0.314	0.107	0.040	2.99*	0.79	4	10.55*
Chaturaṅga	2312	0.326	0.181	0.118	5.30***	3.13*	3	32.16***
Ghare Bāire	1901	0.412	0.079	0.143	1.79	3.63*	3	41.31***
Śheṣer Kavitā	2019	0.360	0.034	0.101	0.73	2.70	3	20.47***
Yogāyog	1187	0.394	0.121	0.179	2.67	4.50*	3	40.48***
Chār-Yārī Kathā	872	0.405	0.111	0.174	2.44	4.26	2	20.43***
Birbaler Hālkhātā	1041	0.190	0.093	0.037	2.12	0.79	3	3.29
Pallisamāj	800	0.308	0.139	0.076	2.83	1.63	3	12.72**
Pather Dābī	815	0.297	0.249	0.173	7.09***	4.91*	2	25.54***
Pather Pānchālī	2552	0.312	0.098	0.027	2.61	0.51	3	0.48*
Aparājita	1894	0.303	0.124	0.062	3.07	0.97	3	11.12*
Devayān	3176	0.390	0.067	0.130	1.14	3.05**	3	57.42***
Dṛṣṭipāt	2363	0.223	0.063	0.036	1.50	0.54	4	6.14
Janāntik	690	0.345	0.054	0.112	1.05	3.10	3	8.85*
Chāchākāhinī	778	0.402	0.022	0.004	0.46	2.54	2	6.58*
Deśhe Videśhe	791	0.387	0.007	0.130	1.20	3.16	2	13.88**

*For explanation of the Poisson and the two lognormal models, are text.

N.B.: Single, double and triple asterisk denote, respectively, significance at 5 per cent level, significance at 1 per cent level and significance at 0.1 per cent level.

ogives on log-probit scale, suggested the following modification for the LN(a) fit. Denoting by P_i the cumulative proportions of observed x-values upto $x = i$, and the normal deviate corresponding to P_i by t_{P_i}, we estimated the parameters θ and λ—mean and s.d. of the underlying logarithmic variate —by solving

$$\log_e 1 + \log_e 2 = 2\theta + \lambda(t_{P_1} + t_{P_2})$$

and
$$\log_e 3 + \log_e 4 = 2\theta + \lambda(t_{P_3} + t_{P_4}).$$

For the LN(b) fit, the quantities on the left-hand side were replaced by $\log_e 1.5 + \log_e 2.5$ and $\log_e 3.5 + \log_e 4.5$, respectively.

The LN(a) fit was generally better for works *wholly* in colloquial style, e.g., '*Ghare Bāire*', while LN(b) tended to be superior where the chaste style is used at least outside conversations. Our choice of the 28 works gave higher weightage to works in the chaste style, and this explains the *over-all* superiority of LN(b) over LN(a) in Table 5.

While the values of D and K show declining time-trends for LN(a), the values for LN(b) seem to fluctuate around a constant level.

The χ^2-test and even the K-test gives significant results in many cases and, evidently, on the whole. The sum of the 28 χ^2's is 680.68, which is a remarkably high value for a χ^2 with 92 d.f. In an absolute sense, the fits are often fairly good, as shown by the small values of K, but the small deviations are statistically significant as the sample sizes are large.

We spent some time in re-examining the distributions for nine languages presented by Fucks (1955). The Poisson fit was better than for Bengali works, with $D = 0.03$ for Esperanto, 0.08 for German and 0.10 to 0.15 for the other languages. For Arabic, however, D is 0.31. The difference $s_x^2-(\bar{x}-1)$ is well below zero for Arabic, Latin and Turkish, near zero for Esperanto and German, and fairly above zero for the four remaining languages. The sample sizes being presumably large, the fit cannot be said to be really satisfactory.

Fucks' approach of studying one 'average' distribution for each language is, in fact, open to serious criticism : the concept of an 'average' distribution is ill-defined. We therefore tried the Poisson model to word-length distributions for individual works in English, German and Russian found in Elderton (1949), Fucks (1952) and Herdan (1950). Obviously, the model cannot apply both for individual works and for the 'average' distribution.

For most works, s_2^2 exceeds $(\bar{x}-1)$ by an appreciable margin. Among English works, Gray's *Poems* ($D = 0.02$) and *Genesis* ($D = 0.03$) showed excellent agreement, but works by Macaulay and several others showed D around 0.25. The findings were similar for German works. For the four Russian works reported in Herdan (1956), the value of D ranged from 0.10 to 0.17. Tests of goodness of fit would give significant results in most cases.

7. Some observations on Bengali poetry

We did some hurried examination of Bengali poetry, covering a very small sample. Actually, we examined (i) the first 200 lines of '*Meghanādabadha Kāvyā*', an epic in blank verse, by Michael Madhusudan Dutt and (ii) 22 poems of Tagore selected in a purposive manner, spread over his poetical life-span, including many famous poems and representing different types of poetry with varying themes, moods and meters. All the poems of Tagore were subjected to complete counts, excepting one long poem, viz., '*Puraskār*', where every 8th stanza starting from the second was chosen.

We refrain from presenting the word-length distributions. The extract from "*Meghanādabadha Kāvyā*' has $\bar{x}$ between 2.55 and 2.6, which is not at all high.

There is little evidence of any time-trend in the averages for Tagore's poems. This is in sharp contrast to the picture for Tagore's novels, essays and short stories. The highest $\bar{x}$ is 3.35 for '*Varshāmangal*'; next comes '*Meghadūt*' (2.85) and '*Ūrvaśhī*' (2.86). At the other end of the scale, we get '*Krshnakali*' (2.10) and two poems for children, '*Vīrpurus*' (1.95) *and* '*Khelābholā*' (2.00). In between the two extremes, one finds almost continuous variation : '*Pranām*' (2.76), '*Swapna*' (2.69), '*Tapobhanga*' (2.65), '*Africa*' (2.56), '*Puraskār*' and '*Satyendranāth Dutta*' (2.48), '*Balākā*' *and* '*Orā Kāj Kare*' (2.47), '*Sandhyā*' (2.46), '*Śhājāhān*' (2.39), '*Niruddeśa Yātrā* (2.35)', '*Bānśī*' (2.32), '*Āmi*' (2.18), '*Badhū*' (2.13), '*Sonār Tarī*' (2.12) and '*Nirjharer Swapnabhanga*' (2.11).

Apparently, a Bengali poem can easily have $\bar{x}$ anywhere from 2 to 2.9, broadly speaking. This is a wide range. A high $\bar{x}$ does not seem to be as unnatural in Bengali poetry as it does in Bengali prose today, because poetry need not employ everyday language. Word-length does not have the same significance in Bengali poetry as it does in Bengali prose. Different poems in the same work of Tagore often show conspicuous variation in $\bar{x}$.

A poem with a high $\bar{x}$ is usually on a serious theme, but the converse is not true (e.g., '*Āmi*.). The forms of verbs and pronouns are not very important. '*Meghadūt*' with colloquial verbs has an elevated style ($\bar{x} = 2.85$), while '*Badhū*' and '*Nirjharer Swapnabhanga*' with $\bar{x}$ near 2.1 use chaste forms.

The relative frequencies of type A syllables vary considerably among the poems examined, which vitiates between poems comparisons in respect of $\bar{x}$. The percentage of type A syllables ranges from 65 to 90. The variation seems to be related to the meter employed. '*Bānśhī*', '*Āmi*' and '*Africa*', written in free verse, report percentages between 65 and 72 and resemble the prose works in this respect. The highest percentage, about 90, is found for '*Varshāmangal*' in the '*māttravrtta*' meter.

References

Aitchison, J. and Brown, J. A. C. (1957): *The Lognormal Distribution.* Cambridge University Press.

Bhattacharya, N. (1960): Syllable counts on modrn Bengali prose (Abstract), Part III (p. 37) *Proceedings, Indian Science Congress*, 47th Session, Bombay.

——— (1965): *Statistical Studies on Languages.* Ph.D. thesis, Indian Statistical Institute, Calcutta.

Brinegar, Claude S. (1963): Mark Twain and the Quintus Curtius Snodgrass letters: a statistical test of authorship. *Jour. Amer. Stat. Assn.*, 58, 85-96.

Chatterjee, Suniti Kumar (1921): A brief sketch of Bengali phonetics. *Bulletin of the School of Oriental Studies, London*, Vol. II, Part I.

——— (1945): *Bhāṣā Prakash Bānglā Byākaran.* 3rd Edition. University of Calcutta.

Cochran, W. G. (1963): *Sampling Techniques.* 2nd Edition. John Wiley and Sons, Inc., New York.

Elderton, W. P. (1949): A few statistics on the length of English words. *Jour. Roy. Stat. Soc.* Series A, Vol. CXII, 436-445.

Flesh, Rudolf (1946): *The Art of Plain Talk.* Harper and Brothers Publishers, New York.

Fucks, Wilhelm (1952): On mathematical analysis of style. *Biometrika*, 39, 122-129.

——— (1955): Mathematical theory of word-formation. pp. 154-170 of *Information Theory, Third London Symposium*, edited by E. Colin Cherry. Butterworths Scientific Publications, London.

Herdan, Gustav (1956): *Language as Choice and Chance.* P. Noordhoff Ltd., Groningen, Holland.

——— (1958): The relation between the dictionary distribution and the occurrence distribution of word-length and its importance for the study of quantitative linguistics. *Biometrika*, 45, 222-228.

Mahalanobis, P. C. (1960): A method of fractile graphical analysis. *Econometrica*, 28, 325-51; reprinted in *Sankhyā*, Series A, (1961), 23, 41-64.

Murthy, M. N. and Nanjamma, N. S. (1959): Almost unbiased ratio estimates based on interpenetrating subsample estimates. *Sankhyā*, **21**, 381-392.

Oettinger, A. G. (1954): The distribution of word-length in technical Russian. *Mechanical Translation*, **1**, 38-40.

Ross, A. S. C. (1950): Philological probability problems. (With discussions). *Jour. Roy. Stat. Soc.*, Series B, XII, 19-59.

Williams, C. B. (1940): A note on the statistical analysis of sentence-length as a criterion of literary style. *Biometrika*, **31**, 356-61.

——— (1956): Studies in the history of probability and statistics, IV: A note on an early statistical study of literary style. *Biometrika*, **43**, 248-56.

Yule, G. Udny (1938): On sentence-length as a statistical characteristic of style in prose: With applications to two cases of disputed authorship. *Biometrika*, **30**, 363-90.

——— (1944): *The Statistical Study of Literary Vocabulary*. Cambridge University Press.

Paper received : September, 1972.
Revised : January, 1974.