Sankhyā : The Indian Journal of Statistics
1975, Volume 37, Series B, Pt. 2, pp. 139-154.

A STATISTICAL STUDY OF WORD-LENGTH IN BENGALI PROSE-II

By N. BHATTACHARYA
Indian Statistical Institute[1]

SUMMARY. If the lengths of words of a given text be recorded in the normal reading order, one gets what may be called a word-length series. The *approximate* randomness of such series is demonstrated in this paper, for each of a number of works in Bengali prose. Word-length was measured in syllables. The autocorrelation coefficients $r_1, r_2, \ldots$, are estimated from the probability samples of words described in the earlier papor (Bhattacharya, 1974).

The words in fiction were classified into two categories, viz., "words used in conversations" and "other words". The estimated percentage of "conversational" words (p_c) and the average lengths of the two classes of words, denoted x_c and x_o, deepen our understanding of the between works variation in average word-length and the historical changes in average word-length.

The series of word-lengths in a Bengali fiction has alternate 'patches' of shortish conversational words and longish 'other' words—which is a departure from perfect randomness.

1. Introduction

If the lengths of the words of a given text be recorded in the normal reading order, one gets what may be called a word-length series. Word-length may be measured in terms of syllables or phonemes or letters. The randomness of such series was discussed by Fucks (1954), who measured word-length in syllables and suggested many methods for studying (i) the correlations between lengths of consecutive words (termed *Nahordnung*) and (ii) correlations between lengths of words which are not consecutive (termed *Fernordnung*). Some of his methods are of dubious value, e.g., the measurement of skewness of the joint distribution of consecutive word-lengths or the estimation of characteristic functions. He, however, showed that the autocorrelation coefficient r_1 of first order is nearly zero for six works in German and one in English (viz. Shakespeare's *Othello*); actually the values of r_1 range from -0.065 to 0.013 and three out of the seven values are positive.[2] The joint distributions of lengths of two consecutive words showed their approximate independence in the statistical sense.

[1]Adapted from a dissertation (Bhattacharya, 1965) for Ph.D. degree of the Indian Statistical Institute. 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.

[2]Fucks did not mention any sample sizes; presumably his figures were based on complete counts. Some work on English partly contradicting the finding of Fucks (1954) is reported in a separate communication.

This paper demonstrates the approximate randomness of the series of word-lengths in a number of works in Bengali prose. In Section 2 we consider the estimates of r_1 obtained for sixteen works, mostly fiction, using the probability samples of words described in the earlier paper (Bhattacharya, 1974). Word-length has been measured in syllables. Section 3 presents some estimates of autocorrelation coefficients of higher order. Section 4 examines the autocorrelations, etc., within some short passages selected from two works in a purposive manner. Of course, the within passage autocorrelation can be appreciably different from the autocorrelation for an entire work.

From Section 5 onwards we consider a classification of words into two categories, viz., "words used in conversations" and "other words". The percentage of conversational matter p_c is itself an interesting indicator of style. Besides, two word-length distributions or two averages of word-length ($\bar{x}_c$ and $\bar{x}_0$), one for conversational matter and the other for the remaining words, characterize the word-length series for any work more satisfactorily than does one over-all distribution or average ($\bar{x}$). We show that $\bar{x}_0$ is generally higher than $\bar{x}_c$, so that the over-all average depends on p_c. Actually, the classification of words into two categories leads to a better understanding of the between works variation in word-length and of the historical changes in the average of word-length.

As conversational words and other words tend to occur in long runs, the series of word-lengths in a Bengali fiction has alternate "patches" with longish non-conversational words and shortish conversational words, the contrast being pronounced when $\bar{x}_0$ is appreciably larger than $\bar{x}_c$. This *non-random* feature of the series may be partly responsible for the positive values of r_1, r_2, etc.

Section 5 presents the percentages p_c and the word-length averages for the two classes of words. The percentages p_c are discussed in Section 6 along with the averages $\bar{x}_c$ and $\bar{x}_0$.

Section 7 concludes the paper with some observations.

2. Autocorrelation coefficients of the first ofder (r_1)

Sixteen works in Bengali prose were covered for the estimation of r_1 (*vide* Table 2). A probability sample of words was drawn from each work by selecting 100 lines by srswr and taking all words on these lines to form the sample (Bhattacharya, 1974). The sample of 100 lines was split into four independent and interpenetrating subsamples (SS), each containing

25 lines, and this gave the four subsamples of the probability sample of words.[3]

We examined the joint distribution of lengths of consecutive words utilizing the probability sample from each work. In order that all pairs of consecutive words may be represented in the sample of pairs, the first word of the line following each sample line was also used. In Table 1 we present these distributions for the (combined) probability sample for two selected works. Generally speaking, these distributions show that the lengths of neighbouring words are nearly independent in the statistical sense.

TABLE 1. JOINT DISTRIBUTION OF LENGTHS OF CONSECUTIVE WORDS BASED ON PROBABILITY SAMPLES OF WORDS FROM TWO SELECTED WORKS IN BENGALI PROSE

length of preceding word in syllables	length of following word in syllables							total
	1	2	3	4	5	6	7	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
				(a) *Gorā*				
1	20	75	30	11	2	—	—	138
2	72	199	118	26	7	1	1	424
3	29	110	86	14	7	1	—	247
4	9	32	12	3	3	—	—	59
5	3	9	4	2	2	—	—	20
6	—	—	1	—	—	—	—	1
total	133	425	251	56	21	2	1	889

length of preceding word in syllables	length of following word in syllables								total
	1	2	3	4	5	6	7	8	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
					(b) *Palliamāj*				
1	41	91	39	6	3	—	—	—	180
2	91	198	108	19	10	1	1	—	428
3	41	101	59	15	5	1	—	1	223
4	6	16	10	3	—	—	—	—	35
5	2	9	6	1	1	—	—	—	19
6	—	2	—	—	—	—	—	—	2
7	—	1	—	—	—	—	—	—	1
8	—	1	—	—	—	—	—	—	1
total	181	419	222	44	19	2	1	1	889

[3]This is part of the material utilized in the earlier paper on word-length, which gives the definition of word-length and examines the properties of estimators based on the probability samples and the non-probabilistic systematic samples.

We examined the sampling properties of the estimated autocorrelation coefficients r_1, r_2, etc., based on the probability samples of words, following the method adopted in Bhattacharya (1974). It seemed reasonable to assume that the sample covariance in the numerator of the autocorrelation coefficient is normally distributed around its true value as a fair approximation. Then under the null hypothesis of zero autocorrelation, the estimated covariance would be normally distributed around zero mean and the estimated autocorrelation coefficient would have zero median, even though it may be biased and its sampling distribution different from normal.

Table 2 presents the subsample-wise and combined estimates of r_1 for the sixteen works. The simple average of the subsample estimates exceeds the combined estimates for only 2 out of the 16 works. Assuming that the bias

TABLE 2. ESTIMATED VALUES OF r_1, THE CORRELATION COEFFICIENT BETWEEN LENGTHS OF CONSECUTIVE WORDS IN SYLLABLES, SEPARATELY FOR SIXTEEN WORKS IN BENGALI PROSE

work	no. of word-pairs in sample	estimate of r_1 by subsamples					simple average of subsample estimates
		ss1	ss2	ss3	ss4	comb.	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Śakuntalā	696	0.031	0.038	0.004	−0.032	0.030	0.010
Sītār Vanavās	750	0.051	0.013	0.007	0.023	0.023	0.024
Durgeṣnandinī	577	0.152	0.052	−0.076	0.087	0.064	0.054
Viṣavṛkṣa	611	0.184	0.085	0.179	0.200	0.139	0.162
Gorā	889	0.018	0.081	0.060	0.077	0.061	0.059
Śeṣer Kavitā	735	−0.030	0.060	0.037	−0.038	0.016	0.007
Chār-Yārī Kathā	872	−0.007	−0.005	0.062	0.044	0.028	0.025
Birbaler Hālkhātā	1041	0.085	0.067	0.105	−0.110	0.040	0.036
Pallisamāj	889	0.062	0.028	0.124	0.058	0.074	0.068
Pather Dābī	815	0.011	0.070	0.058	0.011	0.043	0.039
Pather Pāñchālī	922	−0.038	0.070	0.093	0.193	0.088	0.081
Devayān	931	0.126	0.138	−0.019	0.083	0.087	0.082
Dṛṣṭipāt	772	−0.022	−0.006	0.274	0.020	0.070	0.067
Janāntik	690	−0.026	0.085	0.018	0.170	0.076	0.062
Chāchā Kāhinī	778	−0.060	0.132	−0.055	0.121	0.048	0.035
Deśe Videśe	791	0.107	0.012	−0.010	0.120	0.063	0.057
average						0.0594	0.0542

of an estimate based on k sample lines is of the order of $1/k$, this comparison reveals that the estimates have a significant downward bias. The straight average of the combined estimates for the 16 works is 0.0594, while the corresponding average of the simple averages of the subsample averages is 0.0542. An "almost unbiased"* estimate of the average of the true r_1's of all 16 works may be worked out as

$$0.0594+\frac{1}{3}[0.0594-0.0542] = 0.0611.$$

However, the bias does not appear to be important, being much smaller than sampling errors.

Even if one forgets the small negative bias, the values of r_1 must be regarded as significantly positive, *on the whole*, though not for all the works taken individually. The combined sample estimate as well as the average of the subsample estimates is positive for each of the 16 works. Also, out of the 64 subsample-wise estimates in the whole table, only 15 have the minus sign.

Only 6 works, viz., '*Sītār Vanavās*', '*Viṣavṛkṣa*', '*Gorā*', '*Pallīsamāj*', '*Pather Dābi*' and '*Deśe Videśe*', show positive values of r_1 for all the subsamples. Assuming that under the null hypothesis of zero autocorrelation, the subsample r_1's are distributed around zero as median, we may conclude that the estimates of r_1 are significantly positive (at one-sided 6½% level) for these six works. In view of the downward bias of the estimates of r_1, this test is probably erring on the safe side.

3. Autocorrelation coefficients of higher order

Four out of the sixteen works were covered for the estimation of higher order autocorrelation coefficients r_2, r_3, r_4 and r_7. The same probability samples of words were utilized. For each work, we examined the two-way distribution of lengths of all pairs of words having a specified number of words (s) intervening between them ($s = 1, 2, 3$ and 6, in turn). In order to ensure proper representation of all possible word-pairs in the text with s words between them, the $s+1$ words of the line following each sample line were also used for preparing the afore-mentioned two-way distributons, so that each word of each sample line became in turn the preceding word of one pair.

These two-way distributions are not presented for reasons of space. In general, they pointed to the approximate statistical independence of neighbouring word-lengths.

Vide Murthy and Nanjamma (1959) for the underlying formula for *almost unbiased* ratio estimators based on the subsamplewise and combined sample ratio estimators.

Table 3 presents the estimates of r_2, r_3, r_4 and r_7 for the different works, separately by subsamples and for the combined samples. Even r_7 appears to be statistically significant. The combined estimates are positive for all the four works and two out of the sixteen subsample-wise estimates are negative. The average value of r_7 is roughly about 0.06, as against an average value of 0.08 for r_1 for the same works. Actually, the coefficients r_1, r_2, r_3, r_4 and r_7 do not seem to be rapidly falling to zero. This seems plausible in view of the presence of patches in the works discussed in the following sections; since the patches are usually quite long, two words with only six words between them would usually fall in the same patch.

TABLE 3. ESTIMATED VALUES OF AUTOCORRELATION COEFFICIENTS r_2, r_3, r_4 AND r_7 BETWEEN LENGTHS OF NEIGHBOURING WORDS IN SYLLABLES, SEPARATELY FOR FOUR WORKS IN BENGALI PROSE

work	no. of word-pairs in sample	autocorrelation coefficient	estimates by subsamples					simple average of subsample estimates
			ss1	ss2	ss3	ss4	comb.	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Viśavṛkṣa	611	r_2	0.122	0.068	0.022	−0.022	0.047	0.047
		r_3	0.126	−0.092	−0.111	−0.154	−0.069	−0.058
		r_4	−0.083	0.059	0.010	−0.048	−0.022	−0.015
		r_7	0.122	0.170	0.081	0.064	0.112	0.109
Gorā	889	r_2	0.083	0.061	0.046	0.086	0.068	0.069
		r_3	0.062	0.100	0.070	0.046	0.071	0.070
		r_4	−0.042	0.107	0.118	−0.124	0.023	0.015
		r_7	0.121	0.059	0.030	−0.032	0.042	0.044
Pather Dābi	815	r_2	0.102	0.282	−0.078	0.121	0.057	0.107
		r_3	0.091	−0.062	−0.013	0.024	0.012	0.010
		r_4	−0.084	−0.064	0.088	0.164	0.019	0.026
		r_7	0.048	0.132	0.057	0.001	0.031	0.060
Dṛṣṭipāt	772	r_2	−0.005	0.121	0.187	0.073	0.098	0.094
		r_3	0.033	0.001	0.137	0.104	0.072	0.069
		r_4	0.227	−0.080	0.168	0.001	0.076	0.079
		r_7	0.087	0.115	0.078	−0.004	0.065	0.069

Anyway, for '*Dṛṣṭipāt*' and '*Pather Dābi*' the coefficients r_1, r_2, etc., hardly show any trend; there seems to be some declining trend in the coefficients for '*Gorā*'; the coefficients for '*Viśavṛkṣa*' suggest a cycle with r_3 and r_4 below zero while r_7 is positive. If the estimates for '*Viśavṛkṣa*' are taken seriously, r_3 and r_4 cannot be regarded as significantly positive, on the whole, and the possibility of oscillations in the correlogram may have to be recognised.

4. Evidence from short passages

Two passages were selected from '*Viśavṛkṣa*' and five from '*Dṛṣṭipāt*' for examining autocorrelations within homogeneous passages of moderate length and also for examining the presence of "patches" with unusually high or unusually low or medium levels of average word-length. Most of the passages are indeed "patches" and were chosen in a preliminary search for different types of patches.

The joint distributions of lengths of neighbouring words are omitted for considerations of space. Estimated autocorrelation coefficients are presented in Table 4 and their significance assessed by means of the non-parametric test due to Wald and Wolfowitz (1943).

TABLE 4. AUTOCORRELATION COEFFICIENTS BETWEEN LENGTHS OF NEIGHBOURING WORDS IN SYLLABLES IN SELECTED PASSAGES FROM TWO BENGALI WORKS

work	passage no.	no. of words	average word-length	circular autocorrelation coefficients			
				r_1		r_{10}	
				value	critical ratio	value	critical ratio
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Viśavṛkṣa	1	257	1.988	−0.023*	−0.293	−0.012	−0.118
	2	241	3.083	0.139	2.303	−0.054	−0.868
Dṛṣṭipāt	1	216	1.949	−0.085	−1.185	—	—
	2	218	2.262	0.081	1.279	—	—
	3	212	2.717	0.047	0.074	—	—
	4	211	2.446	0.088	1.297	—	—
	5	221	2.389	0.133	1.982	—	—

*Here the non-circular coefficient is quite different, −0.115.

The coefficient r_1 is negative for one of the two passages from '*Viśavṛkṣa*' and for one of the five passages from '*Dṛṣṭipāt*'. So the coefficient is not significantly positive at the 5% level of significance, by the sign test.

The average value of r_1 is about 0.06 for the passages from '*Viśavṛkṣa*' and nearly 0.04 for those from '*Dṛṣṭipāt*'. Interestingly, these values are smaller than the corresponding estimates for the entire works, viz., 0.139 for

'*Viṣavṛkṣa*' and 0.070 for '*Dṛṣṭipāt*'. It is doubtful, however, whether they are significantly positive. One of the two critical ratios for '*Viṣavṛkṣa*' is significant, but the sum of the two ratios is 2.010, which is not significant, even using a one-sided test. Again, while one of the five critical ratios for '*Dṛṣṭipāt*' is significant, the sum of the five ratios is far from significant. However, the sum of all seven critical ratios is significant at the 5% level.

The values of r_{10} for both the passages from '*Viṣavṛkṣa*' are near zero and non-significant.

It is remarkable that a preliminary search can bring to light unusual patches in either work with such variation in the average word-length. The average length is about 2.46 for '*Viṣavṛkṣa*', as a whole, and the s.d. is 1.07 (Bhattacharya, 1974). At a rough estimate, this work has 35,000 words and can be split into 140 non-overlapping passages of about 250 words each. If the word-length series for '*Viṣavṛkṣa*' were perfectly random, the averages of word-length in these passages would be approximately normally distributed with mean 2.46 and s.d. $1.07/\sqrt{250} = 0.07$. It would be extremely unlikely that among the 140 averages, there would be one as high as 3.08 and another as low as 1.99. Similarly, if we split '*Dṛṣṭipāt*', having nearly 40,500 words, into 190 non-overlapping passages of about 215 words each, the average lengths in these passages would be approximately normally distributed with mean 2.40 and s.d. $1.04/\sqrt{215} = 0.07$ (*vide* Bhattacharya, 1974, for estimates of overall mean and s.d. of word-length in '*Dṛṣṭipāt*'). Again, it would be extremely improbable to find among the 190 averages, one as large as 2.72 (passage no. 3) and another as small as 1.95 (passage no. 1).

If all possible sets of consecutive 250 words or 215 words are considered instead of a particular set of mutually exclusive passages, the probabilities of getting such extreme types of passages would be somewhat increased. However, the conclusion that the series of word-lengths are not perfectly random cannot be altered by such arguments. It may be noted here that a systematic search may reveal even more unusual passages in either work, excepting that passage no. 2 from '*Viṣavṛkṣa*' cannot be beaten.

Thus, due to differences between conversational passages and other passages and due to other reasons like variation of topic, more or less conspicuous patches exist in both the works, having medium or high or low levels of word-length.

5. Words within and outside conversations

We now classify words in a fiction into two categories, viz., "conversational" and "others". A writer of fiction in Bengali during the formative period of Bengali prose had to make two distinct choices regarding the language, one for the language of the conversational matter and the other for the remaining narrative. The chaste style was used throughout in all the works by Vidyasagar, Bankimchandra, and Tagore upto *'Chaturanga'* (*vide* Table 1) excepting *'Gorā'* and also in *'Kābuliwālā'* and *'Kṣudhita Pāsān'*; the colloquial style was used in conversations and the chaste style outside in Tagore's *'Gorā'* and in the works by Saratchandra and Bibhutibhusan, excepting *'Devayān'*; in all the remaining works, including *'Devayān'*, the colloquial style was employed everywhere.[4]

A narrow definition of conversational matter was adopted : it included words actually uttered in conversation with persons present. Letters, soliloquies, words addressed to gods or to absent persons, words spoken in dreams, etc., were included in the "others" category.[5]

Some of the works, viz., *'Chaturanga'*, *'Ghare Bāire'*, *'Chār-Yārī Kathā'*, *'Dṛstipāt'*, *'Chāchā Kāhinī'*, *'Deśe Videśe'*, *'Kābuliwālā'* and *'Kṣudhita Pāsān'*, are written as speeches or reminiscences of the author or of leading characters, who are mentioned in the first person. This lends a conversational character to even the non-conversational matter in these works.

Table 5 presents the estimates p_c for 19 works in Bengali prose including three short stories shown at the end. For works from which both probability and systematic samples of words were drawn, the estimates are given only for the pooled sample. Table 6 shows the corresponding averages $\bar{x}_c$ and $\bar{x}_0$. Standard errors were not computed for any of these estimates; rough ideas may be formed from the subsamplewise estimates. The length-distributions for the two classes of words are not presented for reasons of space.

As regards the sampling properties of these estimates, the estimates p_c may be written as

$$p_c = \sum_i n_i^{(c)} / \Sigma n_i$$

[4] Colloquial forms are occasionally found in conversations in *'Devī Chaudhurāṇī'* by Bankimchandra and in *'Kābuliwālā'* by Tagore.

[5] Such words were quite frequent in *'Viṣavṛkṣa'* by Bankimchandra.

TABLE 5. PERCENTAGE OF WORDS USED IN CONVERSATIONS ESTIMATED FOR DIFFERENT WORKS IN BENGALI PROSE

author	work	type of sample	no. of sample words	percentage of "conversational" words by subsamples				
				ss-1	ss-2	ss-3	ss-4	comb.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Vidyasagar	*Śakuntalā*	prob.	696	57.6	53.6	35.4	59.1	51.7
	Sītār Vanavās	prob.	750	35.8	62.3	47.6	48.0	48.4
Bankimchandra	*Durgeśnandinī*	pooled	2359	20.8	32.6	32.2	27.4	28.0
	Kapālkuṇḍalā	syst.	493	23.1	37.0	25.4	33.9	30.2
	Viṣavṛkṣa	pooled	2463	19.7	14.7	17.0	15.6	16.8
	Kṛṣṇakānter Will	pooled	2526	28.4	29.3	33.8	31.5	30.8
	Ānandamaṭh	pooled	1910	33.5	36.2	39.9	39.1	37.1
	Devī Chaudhurāṇī	pooled	2007	35.6	43.1	43.2	43.3	41.2
	Rājsimha	pooled	1930	37.9	40.5	33.1	36.5	37.0
Rabindranath	*Bauṭhākurāṇīr Hāṭ*	pooled	2419	34.4	36.8	53.4	41.0	41.4
	Rājarṣi	pooled	2321	33.9	20.2	25.1	20.6	24.9
	Chokher Bāli	syst.	1318	28.2	31.5	24.5	31.0	28.8
	Gorā	pooled	2713	41.6	38.9	34.2	37.8	38.2
	Chaturaṅga	pooled	2312	18.5	24.2	25.1	22.3	22.6
	Ghare Bāire	prob.	1901	30.0	21.5	25.2	34.0	27.9
	Śeṣer Kavitā	pooled	2019	45.7	48.1	53.2	58.6	51.4
	Yogāyog	syst.	1187	29.7	26.7	34.4	32.8	30.8
Pramatha Choudhury	*Chār-Yārī Kathā*	prob.	872	29.0	34.7	37.2	35.9	34.2
Saratchandra	*Pallisamāj*	prob.	890	45.9	32.9	39.1	49.8	41.9
	Pather Dābī	prob.	815	34.1	50.8	46.0	43.4	43.3
Bibhutibhusan	*Pather Pāncālī*	pooled	2552	28.7	28.0	27.3	20.4	26.2
	Aparājita	syst.	1894	24.1	73.4	29.2	14.1	26.4
	Devayān	pooled	3176	56.0	51.4	59.6	53.6	55.2
Jajabar	*Dṛṣṭipāt*	pooled	2363	7.29	9.05	20.5	9.9	11.9
	Janāntik	prob.	690	27.0	37.2	12.9	24.3	25.5
Muztaba Ali	*Chāchā-Kāhinī*	prob.	778	20.1	19.0	8.2	24.3	17.9
	Deśe Videśe	prob.	791	46.2	37.9	37.3	21.2	35.5
Rabindranath	*Kābuliwālā*	syst.	779	2.0	19.5	6.8	13.8	10.5
	Kṣudhita Pāṣāṇ	syst.	1192	3.3	2.8	4.0	7.0	4.3
	Laboratory	syst.	1228	67.3	64.7	70.1	72.9	68.8

where n_i is the number of words on the i-th sample line (i.e., cluster of words), $n_i^{(c)}$ the number of conversational words among these, and $\sum_i$ denotes summation over all the k lines (clusters) in the sample or sub-sample. Large sample properties of ratio estimates do not seem to be possessed by most of the p_c's, as the c.v.'s of the sample averages of $n_i^{(c)}$ over lines are usually above 10% (Cochran, 1963, Chap. 6).

In view of our findings regarding the over-all averages $\bar{x}$ (Bhattacharya, 1974), it may be presumed that the estimates $\bar{x}_c$ and $\bar{x}_0$ have the large sample properties of ratio estimates, at least to a rough approximation. This may not, however, be true for the estimates $\bar{x}_c$ in cases where p_c is very small, so that the underlying sample size is inadequate.

6. Observations on p_c, $\bar{x}_c$ and $\bar{x}_0$

We may now consider the estimates p_c set out in Table 5. It is evident from the divergence among subsample estimates that the combined estimates may be in error by more than 5% or even 10%, for some of the works. Actually, since conversational words and other words tend to occur in long runs, precise estimates p_c can only be obtained from much larger samples.

The percentages p_c vary almost continuously from only 4.3 for '*Ksudhita Pāsān*' to 68.8 for '*Laboratory*', both the works mentioned being short stories. Even among typical novels, the range is fairly wide, from 16.81% for '*Visavrksa*' to 51.4% for '*Śeser Kavitā*'.

Some within author differences appear to be statistically significant, according to sign tests based on the subsample-wise estimates. One may compare, for instance, '*Pather Pānchāli*' and '*Aparājita*', on the one hand, with '*Devayān*', on the other, or '*Visavrksa*' with '*Krsnakānter Will*'. More striking differences can be pointed if one compares the works of Bankimchandra or Tagore or Muztaba Ali or Jajabar, without regard to differences in subject-matter, etc.

Coming to Table 6, we notice that $\bar{x}_0$ is larger than $\bar{x}_c$, on the whole, and the difference is statistically significant for many works. (In addition to the sign test, we may here apply Student's t-test on the subsample-wise differences $\bar{x}_0 - \bar{x}_c$ for testing whether the true difference is zero or not.) '*Ghare Bāire*' is a clear exception. One may also mention '*Chaturanga*', '*Śeser Kavitā*', '*Devayān*', '*Chāchā Kāhinī*' and '*Deśe Videśe*', besides '*Ksudhita Pāsān*' and

TABLE 6. AVERAGE LENGTH OF "CONVERSATIONAL WORDS" AND "OTHER WORDS" ESTIMATED FOR DIFFERENT WORKS IN BENGALI PROSE

author	works	type of sample	no. of sample words		average length in syllables by subsamples										
					conversational words					other words					all words
			conversational	others	ss-1	ss-2	ss-3	ss-4	comb.	ss-1	ss-2	ss-3	ss-4	comb.	comb.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Vidyasagar	*Śakuntalā*	prob.	360	336	2.735	2.330	2.638	2.542	2.553	2.986	2.643	2.991	2.824	2.866	2.704
	Sītār Vanavās	prob.	363	387	2.493	2.613	2.551	2.663	2.587	2.710	2.750	2.776	2.968	2.796	2.695
Bankimchandra	*Durgeśnandinī*	pooled	660	1699	2.485	2.315	2.545	2.302	2.407	2.648	2.677	2.691	2.627	2.658	2.588
	Kapālkuṇḍalā	syst.	149	344	2.333	2.140	2.424	2.071	2.208	2.925	2.882	2.742	2.805	2.834	2.645
	Viṣavṛkṣa	pooled	414	2049	2.186	2.247	2.129	2.363	2.227	2.508	2.527	2.585	2.411	2.505	2.459
	Kṛṣṇakānter Will	pooled	777	1749	2.122	2.055	2.137	2.094	2.103	2.430	2.271	2.503	2.483	2.459	2.350
	Ānandamaṭh	pooled	709	1201	2.167	2.249	2.253	2.136	2.202	2.580	2.596	2.606	2.491	2.582	2.441
	Devī Chaudhurāṇī	pooled	827	1180	2.188	2.014	1.979	2.066	2.053	2.421	2.397	2.275	2.537	2.405	2.260
	Rājsimha	pooled	714	1216	2.410	2.286	2.260	2.288	2.311	2.667	2.661	2.500	2.613	2.630	2.512
Rabindranath	*Bauṭhākurāṇīr Hāṭ*	pooled	1002	1417	2.291	2.136	2.168	2.218	2.199	2.466	2.508	2.577	2.597	2.530	2.393
	Rājarṣi	pooled	577	1744	2.196	2.213	2.193	2.121	2.184	2.523	2.520	2.483	2.557	2.521	2.437
	Chokher Bāli	syst.	380	938	2.147	2.186	2.177	2.128	2.158	2.401	2.583	2.428	2.409	2.451	2.366
	Gorā	pooled	1035	1678	2.127	2.160	2.165	1.988	2.110	2.492	2.499	2.488	2.454	2.483	2.341
	Chaturaṅga	pooled	522	1790	2.407	2.180	2.099	2.210	2.209	2.412	2.291	2.325	2.356	2.346	2.315
	Ghare Bāire	prob.	530	1371	1.948	2.147	2.010	2.106	2.049	2.090	2.072	2.186	2.100	2.109	2.093
	Śeṣer Kavitā	pooled	1038	981	2.159	2.197	2.152	2.037	2.131	2.256	2.227	2.263	2.303	2.268	2.198
	Yogāyog	syst.	366	821	2.000	1.878	2.010	2.094	2.000	2.184	2.231	2.378	2.188	2.244	2.168

TABLE 6. (*Contd.*) AVERAGE LENGTH OF "CONVERSATIONAL WORDS" AND "OTHER WORDS" ESTIMATED FOR DIFFERENT WORKS IN BENGALI PROSE

author	works	type of sample	no. of sample words		average length in syllables by subsamples										all words
					conversational words					other words					
			conversational	others	ss-1	ss-2	ss-3	ss-4	comb.	ss-1	ss-2	ss-3	ss-4	comb.	comb.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Pramatha Choudhury	*Chār-Yārī Kathā*	prob.	298	574	1.862	1.865	1.963	2.064	1.943	2.126	2.094	2.102	2.158	2.120	2.060
Saratchandra	*Pallisamāj*	prob.	373	517	1.930	1.840	1.774	2.000	1.898	2.475	2.333	2.580	2.383	2.439	2.212
	Pather Dābī	prob.	353	462	2.000	2.140	1.978	1.906	2.011	2.441	2.464	2.358	2.306	2.394	2.228
Bibhutibhusan	*Pather Pānchālī*	pooled	669	1883	1.935	1.883	1.936	1.886	1.912	2.432	2.387	2.391	2.375	2.396	2.269
	Aparājita	syst.	499	1395	1.959	2.006	2.046	1.968	2.000	2.309	2.393	2.400	2.393	2.371	2.274
	Devayān	pooled	1753	1423	2.108	2.095	2.037	2.138	2.093	2.148	2.233	2.221	2.170	2.193	2.138
Jajabar	*Dṛṣṭipāt*	pooled	281	2082	2.140	2.283	1.984	2.000	2.075	2.419	2.437	2.405	2.491	2.439	2.395
	Janāntik	prob.	176	514	2.204	2.000	2.136	2.095	2.091	2.429	2.156	2.389	2.450	2.362	2.293
Muztaba Ali	*Chāchā Kāhinī*	prob.	139	639	2.105	2.000	2.125	2.109	2.079	2.252	2.289	2.089	2.238	2.213	2.189
	Deśe Videśe	prob.	281	510	2.052	2.217	2.194	1.796	2.085	2.214	2.122	2.231	2.213	2.220	2.172
Rabindranath	*Kābuliwalā*	syst.	82	697	1.750	2.053	2.308	2.444	2.207	2.454	2.612	2.556	2.536	2.535	2.501
	Kṣudhita Pāṣāṇ	syst.	51	1141	2.200	1.625	2.500	1.667	1.961	2.465	2.629	2.505	2.602	2.549	2.524
	Laboratory	syst.	845	323	2.089	1.978	2.067	2.022	2.041	2.404	2.340	2.229	2.337	2.329	2.131

'*Kābuliwālā*'.[6] In these cases, the difference between $\bar{x}_0$ and $\bar{x}_c$ is not clearly significant. It is of interest to note that, for reasons stated earlier, the non-conversational matter in many of these works is akin to conversational matter.

If one plots a scatter diagram showing $\bar{x}_0$ against $\bar{x}_c$ for all the works covered, one obtains a picture of the historical decline in average word-length in Bengali prose. The earliest works by Vidyasagar, Bankimchandra and Tagore (upto '*Chokher Bāli*'), written entirely in the chaste style, reveal a declining trend and the points fall around the line $\bar{x}_0 - \bar{x}_c = 0.35$, approximately. (The largest difference (0.63) between $\bar{x}_0$ and $\bar{x}_c$ is found for '*Kapālkundalā*', but the sample sizes are not large for this work.) The next group of five works employing the colloquial style in conversations but the chaste style elsewhere, which includes *Gorā*, seems to have continued this declining trend, or perhaps, the decline in $\bar{x}_0$ became slower than that in $\bar{x}_c$. Finally, when the colloquial style began to be used throughout, the $\bar{x}_c$-values naturally did not show any further decrease, but $\bar{x}_0$ fell greatly, and the $(\bar{x}_c, \bar{x}_0)$ points are spread around a new line, viz., $\bar{x}_0 - \bar{x}_c = 0.1$ or 0.15.[7] However, '*Drstipāt*' and '*Janāntik*,' written completely in the colloquial style, present higher values of the difference $\bar{x}_0 - \bar{x}_c$.

In many instances, the variation in $\bar{x}$ between similar works by the same author (Bhattacharya, 1974) could be partly explained by the variation in p_c; or, in other words, the variation was less in respect of $\bar{x}_c$ or $\bar{x}_0$ than in respect of $\bar{x}$. Thus, if '*Visavrksa*' ($\bar{x} = 2.459$, $p_c = 16.81\%$) had the same p_c as '*Krsnakānter Will*' ($\bar{x} = 2.350$, $p_c = 30.76\%$), while the averages $\bar{x}_c$ and $\bar{x}_0$ were as estimated, the average $\bar{x}$ for '*Visavrksa*' would have been 2.419; similarly, if '*Krsnakānter Will*' had the same p_c as '*Visavrksa*' while its averages $\bar{x}_c$ and $\bar{x}_0$ were as observed, its over-all average $\bar{x}$ would have risen to 2.399. So, the difference in p_c increased the difference in $\bar{x}$ between these works. '*Rājarsi*' and '*Bauthākurānīr Hāt*' present a more striking example : the $\bar{x}_0$'s and $\bar{x}_c$'s are nearly equal, and the difference in $\bar{x}$ is largely due to the unequal weightage of conversational matter. If '*Rājarsi*' had the same p_c as *Bauthākurānīr Hāt*, while its $\bar{x}_c$ and $\bar{x}_0$ were as they actually are, its $\bar{x}$ would have fallen to 2.381, which is close to the $\bar{x}$-value

[6]Actually, the sample for conversational words is very small for '*Kṣudhita Pāṣāṇ*' and '*Kābuliwālā*'.

[7]'*Chaturaṅga*', written in the chaste style, seems to belong to the third group; as stated earlier, the work uses a *de facto* colloquial style, only the verbs and pronouns having the chaste form.

(2.393) for '*Bauthākurānir Hāt*'. Similar statements may be made about '*Pather Pānchāli*' and '*Aparajita*', on the one hand, and '*Devayān*', on the other, and about '*Drstipāt*' and '*Janāntik*'.

On the whole, $\bar{x}_c$ and $\bar{x}_0$ seem to be positively correlated, but there are exceptions. For instance, x_c seems to be lower for '*Pather Pānchāli*' than for '*Devayān*', while the opposite seems to be true for x_0.

7. Concluding observations

That non-probabilistic systematic samples of words behave like probability samples even in respect of sampling errors (Bhattacharya, 1974) also points to the approximate randomness of the word-length series. Or, rather, the series seem to be like stationary series with the autocorrelations vanishing beyond a few lines, there being no trend or periodicity to affect the systematic samples with intervals of the order of one page.

Some evidence of deviations from perfect randomness is available from Table 2 of our earlier paper. If we compare the estimated standard error ($s_{\bar{x}}$) of the sample average of word-length x based on a probability sample with the corresponding value of $s_x/\sqrt{n}$, where s_x is the s.d. of word-length and n the number of sample words, we find that, on the average, $s_{\bar{x}}$ exceeds the value of $s_x/\sqrt{n}$ by about 10%. The percentage difference varies from -10 to $+34$ among the rows (i.e. works) in the aforementioned table, but only 4 values are negative, so that there is significant evidence that $\sigma_{\bar{x}}$ exceeds the standard error of the sample average based on a srswr of the same size. Since our samples consist of line-clusters, this points to a small positive intra-cluster (i.e. intra-line) correlation between lengths of words.

The difference between conversational words and other words may be partly responsible for this positive correlation. Texts of fictions show alternate patches of shortish conversational words and longish "other words", with the average of word-length in the two sets of patches differing by 0.1 to 0.5 syllables. This non-random feature may give rise to (or exaggerate) the positive autocorrelations r_1, r_2, etc., even if neighbouring word-lengths within conversational passages and within non-conversational passages are independent (or relatively independent) in the statistical sense.

Randomness or otherwise of a given word-length series can be examined in many other ways. One may, for instance, compare the word-length distributions for different parts of the work or for speeches made by different characters.

References

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

——— (1974): A statistical study of word-length in Bengali prose. *Sankhyā*, **36**, Series B, 323-347.

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

Fucks, Wilhelm (1954): On nahordnung and fernordnung in samples of literary texts. *Biometrica*, **41**, 116-132.

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

Wald, A. and Wolfowitz, J. (1943): An exact test for randomness in the non-parametric case based on serial correlation. *Ann. Math. Stat.*, **14**, 378-388.

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