Sankhyā : The Indian Journal of Statistics
1975, Volume 37, Series B, Pt. 2, pp. 165–167.

A NOTE ON THE WORD-LENGTH SERIES IN ENGLISH WORKS

By N. BHATTACHARYA
Indian Statistical Institute, Calcutta

SUMMARY. Fucks (1954) demonstrated the approximate statistical independence of consecutive word-lengths for six German works and one English work (*viz.*, "Othello" by Shakespeare). This note presents some results for English works partly contradicting this finding.

1. Introduction

If the lengths of the words of a given text be recorded in the natural reading order, one gets what may be called a word-length series. Fucks (1954) examined the randomness of such series for six German works and one English work (*viz.*, Shakespeare's "Othello"), measuring word-length in terms of the number of syllables. Consecutive word-lengths appeared to be approximately independent in the statistical sense in all these works and the auto-correlation coefficients r_1 of first order were close to zero. Fucks did not mention any sample size; presumably his results were based on complete counts.

Below are reported some results for English works partly contradicting the above finding of Fucks.[1]

2. Results

A probability sample of words was drawn from Chapters I to 32 of "Pride and Prejudice" by Jane Austen by selecting 200 lines by srswr. This was split into four independent and interpenetrating sub-samples, each obtained from 50 randomly selected lines. A (non-probabilistic) systematic sample was drawn from "The Tale of Two Cities" by Charles Dickens, by selecting the 5th line from top of every 5th page, beginning with page 5. This sample was split into two sub-samples by assigning the lines on pages 5, 15, 25, ... to sub-sample 1 and the remaining lines to sub-sample 2.[2]

[1] More extensive investigations on Bengali prose are reported in Bhattacharya (1975); these point to the approximate randomness of the series of word-lengths in many works in Bengali prose. The methodology adopted here is discussed fully in the afore-mentioned communication.

[2] A similar systematic sample was drawn from "Pride and Prejudice" and showed the approximate equivalence of such samples and probability samples. This systematic sample is not utilized here. *Vide* Bhattacharya (1974) for a detailed account of such methods of (*i*) probability sampling and (*ii*) non-probabilistic systematic sampling, and of the approximate equivalence between the two types of samples.

Word-length was measured in letters. We refrain from presenting the joint distributions of lengths of consecutive words. The estimates of the first order autocorrelation coefficients are shown in Table 1. Needless to say, the first word of the line following each sample line was used for estimating the coefficients.

For the sake of interest, the computations for "Pride and Prejudice" were done in three ways; once by considering only those word-pairs where both the words were included in conversational matter, then for word-pairs where both the words were outside conversational matter, and finally for all the word-pairs in the sample.

TABLE 1: ESTIMATES OF AUTOCORRELATION COEFFICIENT r_1 BETWEEN LENGTHS IN LETTERS OF CONSECUTIVE WORDS.

work	type of sample	sub-sample	no. of word-pairs in sample			estimate of r_1		
			both words conv.	both words non-conv.	all pairs	both words conv.	both words non-conv.	all pairs
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Pride and Prejudice (Chs. 1-32)	prob.	1	198	256	460	−0.133	−0.150	−0.125
		2	198	210	412	−0.152	−0.185	−0.160
		3	154	282	440	−0.105	−0.219	−0.198
		4	267	186	460	−0.156	−0.144	−0.132
		comb.	817	934	1772	−0.150	−0.174	−0.151
A Tale of Two Cities	syst.	1	—	—	337	—	—	−0.131
		2	—	—	352	—	—	−0.092
		comb.	—	—	689	—	—	−0.110

Fucks (1954) measured word-length in syllables, but the distinction between syllables and letters may not be crucial. Fucks examined only one English work, *viz.*, "Othello" by Shakespeare, which is a drama, partly in verse. He found r_1 was −0.0209 for this work, and concluded that r_1 was nearly zero for English. Our estimates partly contradict this conclusion and show that r_1 can be significantly negative and around −0.1 or −0.15 for other works in English.

The difference in average length between words used in conversations and other words could conceivably give rise to positive values of r_1. This expectation was, however, not realized. Indeed, the estimates of r_1 are about equal in the last three columns of Table 1.

The real explanation of the negative r_1 seems to be the tendency of the shorter *grammar* words and the longer *content* words to occur alternately in

English works (Miller *et al*, 1958; Herdan, 1956, pp. 111-5). Compared to this, the presence of alternate patches of shortish "conversational" words and longish "other" words seems to have much smaller effect. The situation seems to be different from that obtaining in Bengali prose (*vide* Bhattacharya, 1975).

We also examined two short passages chosen in a subjective manner from Shakespeare's "Othello"[3] and one passage from "Pride and Prejudice". The main results are shown in Table 2.

TABLE 2: CIRCULAR AUTOCORRELATION COEFFICIENTS r_1 BETWEEN LENGTHS IN LETTERS OF CONSECUTIVE WORDS.

work	passage no.	no. of words	circular r_1		
			estimate	standard error	critical ratio
(1)	(2)	(3)	(4)	(5)	(6)
Othello	1	224	−0.008	0.107	−0.033
	2	276	0.064	0.060	1.128
Pride and Prejudice	1	211	−0.159	0.073	−2.105

Wald-Wolfowitz's non-parametric test (Wald and Wolfowitz, 1943) was applied for judging the significance of the estimates of r_1. The estimate is significantly negative for the passage from "Pride and Prejudice", and nearly equal to the estimate for the same work presented in Table 1. But, for "Othello", the estimates are non-significant and not far from the small value found by Fucks, *viz.*, −0.0299.

References

Bhattacharya, N. (1974): A statistical study of word-length in Bengali prose. *Sankhyā*, Series B, 36, pt. 4, 323-47.

——— (1975): A statistical study of word-length in Bengali prose-II. *Sankhyā*, Series B, this issue.

Fucks, Wilhelm (1954): On nahordnung and fernordnung in samples of literary texts. *Biometrika*, 41, 116-32.

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

Miller, G. A., Newman, E. B. and Friedman, E. A. (1958): Length-frequency statistics for written English. *Information and Control*, 1, 370-89.

Wald, A. and Wolfowitz, J. (1943): An exact test for randomness in the non-parametric case based on serial correlation. *Ann. Math. Stat.*, 14, 378-88.

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

[3] The first passage comprised the first 224 speech words of the drama, entirely in verse, and the second was from scene 5.2, line 23 onwards, where Othello and Desdemona talk before the killing.