CHAPTER I

GENERAL INTRODUCTION
CHAPTER I

GENERAL INTRODUCTION

1.1 QUANTITATIVE STUDY OF LANGUAGE:

Language is a conventional symbolic activity whose main function is to help us to think and to influence the behaviour of others without too much effort. It originated by a masterpiece of opportunism and gave the mouth a new function, apart from eating, fighting and loving. In other words language is a complex system of phonological, morphological and syntactical structures. In a language, phonemes are its atomic unit and syllables its molecular units while alphabets are its symbolic units. Combinatorial structures of syllables form words of the language and those of words form the sentences of the same. The study of the distribution of phonemes, of the structure of words and of the structure of sentences are known as phonology, morphology and syntax of the language. Each language has its phonology and morphology while each writer has its own syntax.
The quantitative study of language has a variety of applications and is an interesting branch of linguistics in its own right. The concept of quantitative linguistics may be regarded as the quantitative version of structural linguistics on one hand and a distinct branch of applied statistics on the other. This type of the study is generally initiated and persuaded by mathematicians and statisticians rather than by philologists.

Many linguists have shown an increasing interest recently in the application of mathematics and in particular, statistical methods to linguistic problems. To quote a few, statistical methods have been developed to study the problem of

(i) the basic structure of language,
(ii) the historical trend in the characteristics of a language, due to change in phonology and syntax etc,
(iii) the style of different authors and their differences and resemblances,
(iv) the dating of valuable manuscripts, documents and disputed authorship.
1.2 REVIEW OF PREVIOUS WORK

Quantitative aspects of literary style have been a subject of study by mathematicians and statisticians for quite some time. One of the earliest references can be traced to 1851, when Augustus de Morgan hypothesised that word-length measured by counting the number of letters in each word, might be a reliable indicator of authorship. Then, an American physicist named Mandenhall, T.C. (1887-1901) confirmed it by his works, in the writing of Shakespeare, Bacon and Marlow. He showed that the frequency distribution of word-lengths in letters has one peak and a long tail towards right. He observed that Shakespeare used four letter words most frequently as also Marlow.

Yule, G.U. (1939), made similar investigation regarding the problem of disputed authorship from an angle of variation in sentence-length, which latter on proved to be a much more fertile approach. He showed that the frequency distribution of sentence-length (i.e. number of words between successive full stops) is of skew type and by comparing the mean, the median, quartiles and deciles of two different manuscripts, he was able to produce convincing mathematical evidence on the identity of otherwise of their authorship.
When (see Williams, 1927) converted some of the Yule's Tables into diagrams, its resemblance to certain skew distributions was noticed. Further, if the logarithm of the number was taken as a basis for subdivision into groups instead of the number itself, distribution of word-lengths tended towards normality.

A pioneering work in this field was carried out by a linguist Zipf (1949). He, by giving the lowest rank to word with the highest frequency and to words with decreasing frequencies, increasing ranks tried to establish inverse rank-frequency relationship. He found that words used most frequently are shorter in length and more versatile. Zipf arrived at negative correlations between lengths and frequencies of occurrences of different words. Mandelbrot (1954) tried to replace the straight line relationship (both on log scale) of Zipf between rank and frequency by a more general curve in the name of "Mandelbrot Cannonical law", but without any justification to rank allotment procedure, which has been questioned all along. Simon (1955) attempted to fit the empirical distribution of word-frequencies by considering model of certain skew distribution functions. Herdan (1966) studies the above problem and concluded that none of the above mentioned relations put forward by Zipf, Mandelbrot and
Simon explains linguistic phenomena successfully. He made stochastic approach to interpret the results of word counts and showed that

$$\text{prob} \left[ \frac{\log v}{\log N} \rightarrow c \right] \rightarrow 1, \text{ for large } n,$$

where $N$ and $v$ stand respectively for the total number of words and that of distinct words in a text.

Thomson and Thompson (1915) attempted to extrapolate lexicon from the type-token ratio. Yule (1944), Muller (1969), Herdan (1966) are a few to quote who used type-token ratio as an index of style of a particular author. Brainered (1972) tried to fit stochastic models of type-token counts by taking samples from literary works of Shakespeare. He concluded that the simple exponential model yeilds a poor fit, while the mixed model is fairly serviceable one.

The randomness of the word-length series, which forms the basis of a number of investigations about authors characteristics, was discussed by Fucks (1954). He measured word-length in syllables and carried out many important calculations such as (i) correlation between lengths of consecutive words and (ii) correlation between lengths of words which are not consecutive. He
found the autocorrelation coefficient of the first order to be insignificant for German and English literature by considering randomly selected pieces of work. Herdan generalized the above result to the extent that randomness will not be disturbed even if sampling is done at chapters or of pages or of some other suitable units. Bhattacharya (1965) studied the randomness of series of word-length in a number of cases of Bengali prose and estimated autocorrelation coefficients of various orders. He also examined the randomness of sentence-length in words employing certain statistical tests. Mazumder (1985) also tested the randomness of word-length and sentence-length in words for Assamese prose literature and found that autocorrelation coefficients of order up to 7 are all near to zero and are quite small for sentence-length series. Fitting of empirical distribution to different characteristics such as word-length and sentence-length is an important field of investigation of quantitative linguistics. Fucks (1955) measured the word-length \(X\) in syllables for a number of languages and concluded that the quantity \((X-1)\) is approximately distributed as a Poisson variate for eight out of nine languages studied by him. A number of scholars such as Wake (1957), Bhattacharya (1965), Gore, Gokhale and Joshi (1979),
Mazumder (1985) have studied the distributional pattern of word-length and sentence-length for Greek, Bengali, Marathi and Assamese languages respectively. Sichel (1971) proposed a family of discrete distributions particularly suited to represent long tailed frequency data. A Compound Poisson distribution was the one suggested for word-frequency distribution. Sichel (1974) also used this distribution for studying the distributional pattern of sentence-length of Greek, Latin and English texts. He found Compound Poisson fit to be satisfactory for word-length frequencies also. Another important application of statistical methods relates to the investigation into uncertain authorships which has been discussed for many years by a number of authors in different languages. An attempt has been made to solve this problem by applying discriminant analysis, etymological analysis and metrical analysis. Yardi (1946) studied the problem of chronology of Shakespeare's plays by applying discriminant analysis. Brinegar (1963) used statistical test for differences of authorship and showed that Mark Twain did not write Quartus cutius snodgrass 10 letters. Mosteller and Wallace (1964) made an effort to solve the problem of disputed authorship of 'Federalist paper' and used weight-rate analysis for discrimination. He concluded that Madison wrote all the 12 of
the disputed papers. Morton (1965) did a study on the authorship of Greek prose. Gore Gokhale-Joshi (1979) studied the authorship of some writings of Kesari and used word-length and sentence-length as discriminators.

The above studies based their inferences, on the morphology and syntax of language. A new approach advocated by Shannon, C.E. (1948) consists in the application of 'Information theory for studying the properties of language, treating this as an information system. He carried out some information theoretic study for English language. His further study showed that, in large samples the sampling distribution of entropy is normal. This information theoretic approach was subsequently adopted by a number of scholars, eg. by Fucks (1952), Siromoney (1963), Siromoney and Rajgopalan (1964), Mazumder (1985) etc for examining language properties.

In carrying out statistical study of any language, precautionary measures must be taken against different kinds of non-sampling errors. For instance, subjective choice of literary pieces, variation in the definitions of sampling designs, error in counting linguistic characters of different investigators in the
sampled units, printing and punctuation mistakes, etc are a few sources of non-sampling errors and their magnitude must be assessed before drawing any final conclusion.

We close this section simply by stating that the books of Yule (1944), Miller (1951), Roberts (1965), Herdan (1956, 1962), Hosteller and Wallace (1964), Partes (1978), Lupson (1970), Williams (1970), Less (1953) may provide a good insight for research in this unconventional field to statisticians.

1.3 Problems under Investigation:

It is evident from the previous section that statistical study of a language has attracted the attention of many research workers, but major contribution is for Foreign languages. In recent past no doubt some studies have been carried out for some Indian languages like Kannada, Marathi, Bengali and Assamese etc., but it appears that no attempt has been made to investigate into quantitative measures of Sanskrit language, the language at the root of many Indian languages. In this thesis, we have made an honest attempt to fill up this gap. In the course of investigation, we have also tackled some allied problems associated with the linguo-statistical survey e.g. problem
of randomness and that of non-sampling errors.

Chapter II gives the general investigation into certain quantitative aspects of Sanskrit literature by considering different literary eras. The whole study is planned on the sample survey basis and centres around the estimation of word-length, sentence-length and proportion of parts of speech. The analysis given in this chapter is of general nature and lays emphasis on era-wise changes, if any.

Chapter III discusses the analysis of the word count data for certain sampled pieces of Sanskrit prose. It has been tried to compare the results with those available for few foreign languages. This chapter also includes the mathematical analysis of type-token relationship and fitting of the stochastic model to it.

Chapter IV investigates the randomness of the series of word-length (in syllables) and that of sentence-length (in words) for the prose work of Sanskrit literature. The sample data giving the word-length and sentence-length series are based on simple random sampling. The autocorrelations of various orders between successive values of word-length as well as sentence-length have been obtained for studying dependence. Since in our investigation neither word-length nor sentence-length distribution of any text under consideration
follows normal law, a test of randomness, proposed
by Von Neumann [see Tintner, 1962] has been applied.

Chapter V attempts to fit Compound Poisson and Negative binomial distributions to word-length (in syllables) and sentence-length (in words) of written prose of Sanskrit language. For examining the goodness of fit of Compound Poisson and Negative binomial fits to the observed distribution of word-length and sentence-length, a number of measures (same as those of Bhattacharya, 1965) based on the differences of word-length (sentence-length) have been used. Fit with the least value of measure was taken as a good fit, in the absence of knowledge of exact sampling distributions of test statistic and the power of tests based on them.

Chapter VI attempts to study the standing problem of Sanskrit literature i.e. the disputed authorship of the Kavyas "Kumārasambhava" (9th to 17th cantos) and "Rātusamhār". The problem relates to a literary controversy around the conclusive authorship and its solution is the confirmation, whether Kālidāsa is the author of these writings. The analysis of this chapter as a whole can be divided into four groups, namely

(i) Type-token ratio study,
(ii) Yule's characteristic study,
(iii) Weight rate analysis, and 
(iv) General statistical analysis.

The conclusion is based on the fact that whether the characteristic mentioned above for the disputed Kavyas lies within the range of such characteristic computed for various undisputed Kavyas of Kalidasa.

Chapter VII estimates the entropy of Sanskrit language and includes discussion regarding non-sampling errors. The value of entropy for Sanskrit language has been compared with the entropies of other languages—English, German, Russian and Assamese. Our comparison shows that structurally Sanskrit language with syllable as unit, resembles quite a bit with Russian language. This chapter also compares the entropy for Kalidasa's Kavyas with the disputed pieces of literature. It is found that the entropy for both the disputed Kavyas i.e. Kumārasambhava (9th to 17th cantos) and Kusumāntara is different from the entropy of Kalidasa's well known Kavyas, reconfirming the Mallinath's comment that disputed pieces may not be attributed to Kalidasa due to inferior style and language and many other quantitative properties.