Chapter 1

HISTORY AND REVIEW OF STATISTICAL LINGUISTICS

1.1 LANGUAGE

Language is a means of communication. It has various dimensions like power of expression, richness, easy rendering of meaning, lucidity, aesthetic, rhetorical etc.

Languages are also bookish, colloquial, poetic, factual, scientific etc.

One can express views, thoughts, emotions etc. through language. One way of knowing a person is studying his language.

Different people use same language in different manners. Description of a particular event or expression of a feeling through language differs from person to person.

A single word "Hello" on a telephone helps us to distinguish persons. It is because, the style of uttering the word "Hello" differs from individual to individual.

The way people use language gives us information about their age, their geographical ethnic, social background, the context in which they are communicating etc. This is because the factors like geographical location, caste, urban-rural culture,
education, age etc. influence language and results in variation of language.

Apart from these factors, which are responsible for variation in language, one more factor “style” is also responsible for such variations.

1.2 STYLE

What is style?

Style may be viewed as a set of language features that make people distinctive.

If a single word like “Hello” constitutes style and is able to distinguish persons, then style of speaking or writing of a language which contains lakhs of words, patterns of sentences, punctuations, synonyms etc. used by crores of people for speaking and by lakhs of people for writing will constitute infinite number of styles.

This study is restricted to writing style. Writing style is a selection of set of linguistic features and their interactions from all the possibilities in a language. The variation in the effect of these features can be understood intuitionally by linguists. But there are often occasions when we have to have an analytical approach than to have intuition only. Rather intuition needs to be supported by more objective account of style. It is this approach which is known as stylistics or statistical linguistics or stylostatistics.
1.3 STATISTICAL LINGUISTICS

A study of linguistic style or linguistic problems, through statistical techniques is termed as stylometry. Stylometry is also called statistical linguistics, stylostatistics, mathematical linguistics, quantitative linguistics.

A stylostatistician is different from language critic. Literary critic sees that writing style of Mahatma Gandhi is different from writing style of Pandit Jawaharlal Nehru. But stylostatistician shows that these differences are objective and measurable.

The stylostatistician applies statistical methods to literary aspect and thereby either confirms the suitability of existing tools or introduces new statistical tools, whereas literary stylist views at applications of statistical tools & methods as a solution to problems of literary style like authorship, comparison of styles etc.

Many fields of humanities have been benefitted from the applications of mathematical and statistical models and techniques. Not the least among these is the study of creative use of language. Stylistics and study of texts have become a crossroads for the interests of literary scholars and critics, linguists, psychologists, psychotherapists, sociologists, mathematicians etc. Everyone among these has his own interest and approach to look at the text.

Sociologist is interested in studying, whether some social forces are at work, psychologist may be interested in the study of writer's psychology through text, literary scholar is interested to know literary value of the text, linguist may be busy in knowing or discovering patterns and variability both, in language
system, lawyers are interested in language from legal point of view. Nowadays, stylistic analysis is carried out for speech samples, where the aim is to establish similarity in accused person’s style or language and that heard on a tape recorder.

Stylostatistics is often used to attribute authorship to anonymous writers or in case of disputed documents. It has legal as well as academic and literary applications, ranging from authorship to forensic linguistics.

1.4 STYLE; A PROBABILISTIC APPROACH

Apparently stylistic features are consciously controlled only to a limited degree. It is said that style is not a coat, that can be put on and put off, it is the very skin of a writer. An author is not expected to exhibit identical values of different parameters of style in all his texts. Hence it is said that style is a probabilistic concept.

The probabilistic approach reveals stylistic stability, that is a stable probability distribution. By adopting specific statistical procedures and statistical tests, one can distinguish between insignificant fluctuations in style, which are chance fluctuations or fluctuations that can be attributed to sampling, and significant fluctuations which indicate differences in styles.

It is noted that style is inherent only to some extent. It is also a function of factors like education, age, geographical location etc.
1.5 HISTORY OF STATISTICAL LINGUISTICS

The history of modern statistical stylistics apparently begins in 1851, when Augustus de Morgan suggested that word-length might prove to be a character, which can distinguish writing styles. T. C. Mendanhall, an American geophysicist (1887) was stimulated by the suggestion due to De-Morgan and made an attempt to test it, in case of: Dickens, Thakeray & Mill. He realised that graphical display of relative frequencies of word lengths would be more appropriate than simple mean word length as suggested by Augustus. In this early stage, word length was used as a parameter to study the plays attributed to Shakespeare in 1901.

Gustav Herdan (1966) studied the mathematical foundation of the topic and he showed that vocabulary richness can be studied without depending upon sample size. Many leading investigations were carried out by Yule (1944).

Recently Douglas Biber (1995) moved stylometry away from individual authorship studies to historical evolution of style and also to style differences in English and other languages, by means of factor analysis applied to grammatical features of text. Some numerical measures were used to measure resemblance between pairs of languages. Ernst Forsteman (1952) and Jan Czekanowsk i (1927) used phonological and grammatical properties to measure resemblance.

Another statistical approach to language was based on generalization about word frequencies stated by Estoup (1916) but later publicised by George Kingsley Zipf, whose name is associated with Zipf’s law about rank-frequency of words. The law
was discussed and commented by Christopher Manning, Hinrich Shutze Conden (1928), Carrol (1938), Benoit Mandelbrot (1954), Parker, Rhodes & Joyce (1956), George Miller (1957), Bosanquet (1957), Good (1957) etc.

Yule (1944) suggested a measure to specify author’s richness of vocabulary. But vocabulary size is only one aspect of the style. Distribution of word frequencies is also means of expressing individual style.

The discovery of laws underlying the distribution of vocabulary (Muller Charles – 1965) and an attempt to illustrate an author’s vocabulary development through time (Hart – 1943) both lead to significant discoveries about an author’s style. Such work may direct attention to the psychological factors that influence lexical habits or attention to literacy and cultural influences on writer’s behaviour. [Johnson 1944, Mann 1944, Chotlos 1944, Fairbanks 1944]. But as Dolezel (Essay on a framework for the statistical analysis of style) pointed out that while studying style of an author, one must be able to distinguish between contextual influence and pattern of structure of language imposed by individual.

Besides word distribution, word length was also considered as stylistic trait (Krishnamurti 1950, Brineger 1963, Muller 1984).

Further Fucks and Lauster (1965) used relative frequency of i-syllabled words, their distribution in the text (length of gaps between i-syllabled words) and the entropy of the syllable structure of the text as parameters of style.

Many studies have been undertaken to study style of a writer by using sentence length. (Yule – 1939, Willam Wake – 1957).
1.6 APPLICATIONS OF STATISTICAL LINGUISTICS

1.6.1 Problem of Disputed Authorship

Statistical stylistics is mostly treated as statistical methods concerning cases of disputed authorship. But in fact solutions to such problems reflect only one application of the methods of this discipline.

Authorship problems are dealt with identification of linguistic characters in text by unknown author and its comparison with known writings of the authorship candidate. When a particular list of candidates has been determined, the chances that each one wrote the disputed text, can be calculated by using one, two, three ......(as is necessary) parameters of style, (e. g. word frequency, word length, verb–adjective ratio,... etc.) and the problem of authorship is solved to some extent.

Yule (1938–39), tried to solve the problem of authorship of “The Imitation of Christ” based on parameter sentence length. He concluded that sentence length is not a sufficient indicator to determine authorship. In 1944 he devised his well known characteristic–k. It was devised on the assumption, that occurrence of a given word is based upon chance and follows Poisson law. Ross (1950) led to reformulation of k. Herdan (1955) defined k as the coefficient of variation of means.

Other measures and techniques have been applied to problems of disputed authorship. Ellegard (1962) to Junius problem, Masteller and Wallace (1963 & 1964) to Federlist essays.
1.6.2 Problem of Undated Texts

The problem which is very much related to the issue of unknown authorship is of placement of undated texts. An author’s work can be placed chronologically or undated text may be placed in a time span by comparison with other texts.

Studies of an author’s literary development indicate that certain traits show a relatively constant change through time within specifiable limits. But the degree of significant variation in the development of these traits, must be specified before any firm answers are given to chronology problems.

In the last decade of nineteenth century, Prof. L. A. Sherman of University of Nebraska raised another sort of chronology problem, which is concerned not with an individual author but with the whole body of English literature. His work is parallel to Jespersen’s theory of “progress in grammar” which appeared about the same time.

1.6.3 Other Important Areas of Application

A certainly better and different type of work was carried out by Miss Josephine Miles on history of style (1967). She has worked on typical diction and favoured sentence pattern of each of the major periods of English and American literature. In another study of favoured words, she has used factor analysis to distinguish significant “clusters” of favoured words (Miles and Solvin 1966). Due to her influence statistical approaches to literary problems have been accepted by American literary critics.

Statistical methods are also used to study poetic devices.
Statistical studies of line and stanza pattern are common in Europe. Skinner (1939) studied alliteration and sound patterning of Shakespear's sonnets. He formulated "coefficient of alliteration". Others also considered the problem of alliteration, in verse from statistical point of view (Sebeok & Zep 1959, Goovaerts 1966).

Thus there are varied areas wherein Statistical Linguistics is applied.

As already stated the wide range of applications of the subject includes determination of authorship to forensic linguistics. Some of these are,

i) To know about author's style, consistency in his style.

ii) Objective comparison of styles of two or more authors.

iii) Comparative study of linguistic features of two or more languages.

iv) Comparative study of a single language with respect to time periods.

v) To decide the status of an author by analysing the effect on reader.

vi) To solve problems of disputed authorship.

vii) To solve problems of Forensic linguistics.

viii) To study language variation from social (historical, geographical, educational, etc.) point of view.

ix) Helpful to psychiatrists / psychotherapists to understand idiosyncratic (a way of behaviour especially when it is unusual) linguistic behaviour of their patients.
x) To deal with patients of specific language impairment and aphsia (language disorder resulting from brain damage).

xi) Problem of undated texts.