Quantification of Stylistic Traits: A Statistical Approach

Mappillairaju Bagavandas, G. Manimannan

Department of Statistics – Madras Christian College – Chennai-600 059 – India
mbdas49@hotmail.com, manimannang@yahoo.co.in

Abstract

It is often recognized that authors have writing styles and it is possible to find a simple statistical mo del, which explains reasonably what makes an author unique. This paper makes an attempt to identify the distinct stylistic features of three Tamil Scholars of the same period and also tries to quantify the writing styles of these authors using eighteen stylistic features. These stylistic features of this study are eleven morphological variables, four habitual words and three function words. ANOVA technique, two sample t-statistic and Factor analysis are used for measuring such stylistics traits and also identifies those traits, which are most densely packed.

Keywords: author style statistics, ANOVA, two-sample t-statistic, factor analysis.

1. Introduction

It has been recognized that an author has a unique writing style which is expressed in the form of subconscious stylistic features. Style of an author can be quantified by counting his\her choice of words for expressing his\ her ideas under the assumption that the writer favouring a stock of words for the expression of ideas is regarded, to some extent, subject to chance (Holmes and Forsyth, 1995). Hence given a certain personality and thus a certain style, as its expression, the characteristic properties of style can be described in terms of statistical law (Herdan, 1964). Bailey (1979) says that the stylistic features of a matured writer will be salient, structural, frequent and easily quantifiable. Thus style reflects personality of a writer and this unconscious process is consistent in the case of matured writers (Holmes, 1985).

Statistical stylistic study not only compliments the traditional scholarship of literary experts but also provides an alternative method for investigating the works of doubtful provenance (Holmes, 1998). These studies provide authentic results if they work within the same genre and also work within as close a time period as possible. Stylistic markers which occur most frequently in a given passage are also identified by these methods (Mealand, 1997). These stylistic studies inhabit two types of problems, the first being the selection of suitable set of stylistic variables and the second being the selection of appropriate techniques. There is no general agreement on the stylistic variable that should be used in stylistic studies. In general, when choosing the stylistic variables, one must use something that has large variation across authors and relatively little variation among an author's own work. Initially, lexical variables have predominated in the stylometry studies, yet this decade has seen the application of syntactic and semantic variables (Holmes, 1998).

Mathematician like Fucks (1952) may be considered pioneers in laying a foundation for more vigorous and objective stylistic analysis through his attempts to quantify stylistic features. Mosteller and Wallace's study (1964) is considered as the first authentic stylometric study

soundly based on modern statistical procedure using computer as its major research tool. John Burrows (1987) through his series of seminal papers introduced stylometry studies as a viable tool for authorship attribution problem. The availability of modern computing facility has provided a unique opportunity for many stylometricians to introduce many multivariate meth ods like factor analysis, cluster analysis and correspondence analysis for conducting experiments with high dimensional data and also to widen the frontiers of stylometry (Peng, 2001).

Factor analysis is considered as an ideal method for determining the relationship between stylistic features and stable personality traits (Sommers, 1966; Herdan, 1964). This analysis helps to find out whether different writers really represent different distinct forms of behaviours or whet her they draw from a limited stock of vocabulary (Miles and Selvin, 1966). This multivariate technique is also used for measuring the extent to which groups of words have similar patterns of high or low use of various writers.

Herdan (1941) was the first to use the factor analysis for analysing the relation between six authors and for identifying the one who uses the most difficult words. This analysis was also applied to establish the common ancestors of a number of proto Indo-European languages (Johnson and Kotz, 1967). Roger Peng and Nicolas Hengartner (2002) have used factor analysis to examine each individual author's function-word counts and also to filter out words which account for very little of the variation between authors in the group. This analysis was used by David Mealand (1997) to establish that samples of different genres from the Gospel of Mark vary in style and also to identify the stylistic markers which are most heavily used in these passages.

2. Data and methods

The present study deals with the literary works of three contemporary Tamil scholars,namely, MahakaviBarathi (MB), V.Kalyanasundaram (VK) and Subramaniya Iyer (SI). In the Pre–Independence period, these three scholars have written number of articles on India's Freedom Movement in the magazine called *India*. Initially, all the three scholars have written articles by attributing their names. The oppressive attitude of the then British Regime made all the three writers to write articles on the same topic anonymously in the same magazine. All the attributed and unattributed articles written on India's Freedom Movement in that magazine were complied and brought out as a book entitled *Bharathi Dharisanam* in the year 1975. For this quantitative stylistic study, all attributed articles of these three scholars written on India's Freedom Movement in the year 1906 are considered. Our study is based on nineteen articles of Bharati, six of Kalyanasundram and seven of Subramaniya Iyer.

In stylometry, there are important decisions to be made about the features to be selected and the methods to be used (Mealand, 1997). Eighteen stylistic features are consider ed for this study. They include eleven morphological variables, four habitual words and three function words. The exact lists of variables of this study with their abbreviations are given in Table 1.

For a comparative analysis the frequency counts of the stylistic features must be normalized to the text length in an article. In this study since each sentence is considered as a sample, to normalize the stylistic features, the raw frequency counts of each stylistic feature is divided by the number of words in each sentence and then multiplied by hundred to express it in percentage. Eighteen stylistic features are identified from each sentence. These features include parts of speech, habitual words and function words. Both voices and tenses are expressed in frequencies but not in percentages. If we have n sentences and if we identify p stylistic features

from each sentence, then we have a data matrix of size n x p. Thus, each article was converted as a data matrix and these data matrices form the basis for this quantitative study.

A chi-square analysis of the nineteen articles of Bhararthi establishes that these articles do not differ from one another in terms of the frequency distribution of occurrence of these stylistic features. Similar results were obtained in the case of other two scholars (Manimannan and Bagavandas, 2001). Hence all the nineteen articles of Bharathi are considered as one article for this study. So also, the six articles of Kalyanasundram and seven articles of Subramaniya Iyer. In this study, each sentence is considered as a sample. Hence the nineteen articles of Bharathi consist of three hundred and fifty three sentences, six articles of Kalyanasundram consist three hundred and eighty two sentences and seven articles of Subramaniya Iyer consist of three hundred and fifteen sentences. As there are three authors, there are three data matrices and their sizes are (353x18), (382x18) and (315x18) respectively. Hence the aim is to compare the data matrices of the linguistic features of the three scholars. Average values, two-sample t-statistic values and Euclidean distance values are given in Table1.

3. Analysis

This analysis section consists of two parts. Part one identifies the special stylistics features of each author and Part two quantifies the writing style of each author.

3.1. *Identification of Special Stylistic Features*

This univariate analysis compares the average values of the stylistic features of the three scholars. This comparative study is made in two stages. In the first stage, the hypothesis of equality of the means of a particular feature of three authors is tested using ANOVA (one-way) technique. The acceptance of this hypothesis indicates that particular stylistic feature has no discriminatory power. However, if this hypothesis is rejected, then mean difference of a feature between any two authors is tested using the conventional two-sample t-statistic.

This two-stage comparative analysis indicates that the stylistic features like two-letter word, three-letter word and pronoun do not discriminate these three scholars from one another. That is, these three scholars had the habit of using the same number of these features in writing a sentence. This result indicates that all these three scholars had used, on an average, one pronoun, one two-letter word and two three-letter words in a sentence of ten words. Also the smaller percentages of occurrence of features like intensifier, infinity and adverb indicates that these three authors had used these three features very rarely.

The percentages of occurrences of stylistic features like noun, post-position, clitic, case makers and conjunctions differentiate these three authors statistically from one another. This result indicates that Bharathi is identified as the least user of these features whereas Kalyanasundram is identified as the maximum user of the same stylistic features. Subramaniya Iyer is not identified with any distinct stylistic features because the percentages of the occurrences of stylistic features of this author indicate that the writing style of this author shares equally the special features of the other two authors. The Euclidean distance values confirm this result in Table 1.

This analysis shows that in the sentence of ten words, Bharathi had used, on an average, one postposition, one clitic but three nouns and four case markers and two conjunctions. But on the other hand, in the sentence of the same length, on an average, Kalyanasundram had used five nouns, four conjunctions, three postpositions, three clitics but six case markers. Also it can be seen that the third author, Subramania Iyar had used, on an average, four nouns, three post positions, three clitics, six case markers and four conjunctions.

3.2. *Stylo – Statistical Analysis*

Factor analysis is a variable-oriented multivariate technique. This analysis describes the inter-relationship among many variables in terms of a few underlying, but observable, random qualities called factors (Lawley and Maxwell, 1971). Factor analysis can be considered as an extension of principal component analysis and is used for data reduction and interpretation. This analysis is also used for grouping of variables in such a way that the variables are highly correlated with in groups but have relatively insignificant correlation with variables of different groups. Correlation matrix of the eighteen stylistic features is calculated for each data matrix. The initial statistics are given in Table 2 and groups of stylistic features are given in Table 3.

3.2.1. *The case of Bharathi*

All the eighteen features are highly loaded in the first seven factors, which covers nearly 54 % of the total variation present in this data set. In other words these eighteen features are grouped into seven clusters on the basis of the inter-relationship among themselves. The features like words starting with vowel, verb, two-letter, three-letter and four-letter words are highly loaded in the first factor and hence they form as a cluster. This result shows that the writer Bharathi had preferred verbs and words starting with vowels either as two-letter or three-letter or four-letter words. Since four out of these five features are habitual words, this factor is named as habitual-word factor.

Factor two is highly correlated with features like clitics and case makers. These correlated relationships establish that this writer had the habit of using clitics and case makers in the ratio of 1: 4 in a sentence of ten words. As these two features are function words, this factor is known as function-word factor. Third factor is a contrast between features like noun and pronoun and also they occur in the ratio 4:1 and whenever the occurrence of noun increases the occurrence of pronoun decreases in a sentence. These two features are morphological variables and hence this factor is known as morphological factor. Statistical features like tenses and numeral are accommodated in the fourth factor-the tense factor. This factor indicates that this writer had used to write sentences mostly in past tense with a very few numerals.

Since features like voice and postposition are accommodated in the fifth factor, this factor may be named as voice factor. This factor is a contrast between voice and postposition. This indicates that the writer had favoured to write sentences in the past tense with less number of postpositions. The sixth factor is a syllable factor and it's established that the length of ten words sentence on an average fifteen syllables.

The Seventh factor is contrast between two groups of features. Infinity and adverb are grouped together and intensifiers and conjunction are grouped together. The occurrence of these stylistic features like intensifier, infinity and adverb are rare phenomena. But Bharathi had used at least two conjunctions in a sentence of ten words and hence this factor may be called conjunction factor.

Summarizing, the writer Bharathi had used passive voice sentences in past tense to narrate India's Freedom Movement. In sentence of ten words, he had used, on the average, one clitic, one pronoun, two verbs, three words starting with vowels, four case makers and four nouns. The verbs and words starting with vowels are either two-letter or three-letter or four-letter words. The incre ase in the occurrence of nouns reduces the occurrence of pronouns.

3.2.2. *The case of Kalyanasundram*

The first four factors, which cover nearly 36 % of the total variation present in the data set, had grouped all eighteen features into four clusters. In the first factor features like case maker and clitic are highly accommodated in the ratio 3:1 and hence this factor is known as function-word factor. The features like verb and noun are grouped in the second factor in the ratio 1: 2. This factor is a contrast between verb and noun, which indicates that whenever the occurrence of nouns increases, the occurrence of verbs decreases. This is a morphological factor.

The third factor is a habitual-word factor as it accommodates all the three habitual words with pronoun. This factor is a contrast between four-letter word and the set of two-letter and three-letter words and pronouns. More the occurrence of pronouns, less it will be four-letter words. Fourth factor is a contrast factor. Adverb, syllable, conjunction and infinity are grouped in one set and voice, tense, word starting with vowel, intensifier, and clitics are grouped in another set. This result shows that this author had written active voice sentence in past tense. This author has provided four conjunctions, three words starting with vowels and three clitics. The occurrence of more conjunctions in a sentence reduces the occurrence of clitics and words starting with vowels.

Finally, the author Kalyanasundram used active voice sentences in past tense to describe India's Freedom Movement. In these sentences of ten words, on the average, three postpositions, three clitics, three words starting with vowels, four conjunctions, four nouns and six case markers are accommodated. The occurrence of more verbs red uces the occurrence of nouns; also the occurrence of more conjunctions reduces the occurrences of clitics and the words starting with vowels.

3.2.3. *The case of Subramaniya Iyar*

All the eighteen features are accommodated in the first seven factors, which covers nearly 56 % of variation present in the given data set. In the first factor, verb, word starting with vowel, syllable and four-letter word are accommodated. This indicates that verb and word starting with vowel will be four-letter words with two or three syllables. This is a morphological factor. Case marker and clitic are highly loaded in the second factor and they occur in the ratio 2:1 in a sentence. This is a function- word factor.

Third factor, a contrast factor, provides high loading for nouns and pronouns in the ratio 6:1. The occurrence of more nouns reduces the occurrence of pronouns. This is a noun-family factor. Fourth factor accommodates voice and conjunction. This author used to write passive voice sentences with at least three conjunctions. In the fourth factor adverb and two-letter word are accommodated and this indicates that the adverbs of this author are identified as two-letter words.

Fifth factor is a tense factor. This writer used to write sentence in present tense. The last factor contrasts between two sets of features. In one set postposition, intensifier and three-letter word are accommodated and in the other set numeral and infinity are accommodated. This result indicates that there will be three post-positions and two three-letter words in a sentence of ten words.

Summarizing, the scholar Subramaniya Iyar made use of passive voice sentences in present tense. There will be six case markers, four nouns, three conjunctions and three postpositions and one pronoun in a sentence. The verb and word starting with vowel will be four-letter words with two or three syllables. The occurrences of more nouns reduce the occurrence of pronouns.

4. Conclusions

This study provides opportunities to introduce statistical techniques for identifying the special stylistic features and also for quantifying the writing styles of three Tamil scholars, namely, Mahakavi Bharathi, V. Kalyanasundram and Subramania Iyer using eighteen stylistic features. Articles written on India's Freedom Movement by these scholars are considered for this study. Bharathi had written sentences in past tense with the least function words. V. Kalyanasundram is identified as a writer who has used maximum number of function words in active voice sentences with past tense. The third writer, Subramaniya Iyar has written sentences in passive voice but in present tense and is not identified with any distinct stylistic features.

References

Bailey R.W. (1979). The Future of Computational Stylistic. *Association for Literary and Linguistic Computing Bulletin*.

Burrow J. (1987). Word-Patterns and Story-shapes: The Statistical Analysis of Narrative Style. *Literary and Linguistic Computing*, vol. (2/2): 61-70.

Fucks W. (1952). On the Mathematical Analysis of Style. *Biometrica*, vol. (39): 122-129.

Herdan G. (1941). *The Advanced Theory of Language as Choice and Chance*. The Hegue.

Herdan G. (1964). On Communication between Linguistics. *Linguistics*, vol. (9): 71-76.

Holmes D.I. (1985). The Analysis of Literary Style: A Review. *Journal of the Royal Statistical Society*, Series A (155): 91-120.

Holmes D.I. and Forsyth R.S. (1995). The Federalist Revisited: New Directions in Authorship Attribution. *Literary and Linguistic Computing*, vol. (10): 11-27.

Holmes D.I. (1998). The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing*, vol. (13): 111-117.

Johnson N.L. and Kotz. S. (1967). *Discrete Distributions*. Houghton Miffin Company Boston.

Lawley P.A. and Maxwell. A.E. (1971). *Factor Analysis as a Statistical Method* (2nd Ed.). American Elsevier Publishing and Co. New York.

Manimannan G. and Bagavandas M. (2001). The Authorship Attribution: the case of Bharathiyar. In *Paper Presented at National Conference on Mathematical and Applied Statistics*, Nagpur University.

Mealand D. (1997). Measuring Genre Differences in Mark with Correspondence Analysis. *Literary and Linguistic Computing*, vol. (12/4).

Miles J. and Selvin H.C. (1966). A Factor Analysis of the Vocabulary of Property in the Seventeenth Century. In Leed J. (Ed.), *The Computer and Literary Style*. State University Press.

Mosteller F. and Wallace D.L. (1964). *Applied Bayesian and Classical Inference, The Case of the Federalist Papers*. Addition-Wesley, Reeding.

Peng R. and Hengartner N. (2001). *Quantitative Analysis of Literary Style*. University of California, CA90095.

Peng R. and Hengartner N. (2001). *Statistical Aspects of Literary Style*. Yale University, CC-99.

Stylistic features	Abbre viations	Mean values			Two samples t-statistic values			Euclidean values		
		MB	VK	SI	MB-VK	MB-SI	VK-SI	MB-VK	MB-SI	VK-SI
Noun	P_Noun	34.26	45.47	41.8	09.38	06.21	02.91	125.49	56.73	13.47
Intensifier	P_Int	00.05	06.07	06.1	11.63	10.31	00.09*	31.07	31.79	0.00
Infinitive	P_Inf	00.58	01.33	03.6	02.69	06.14	04.26	0.57	9.18	5.19
Pronoun	P_Pro	07.72	07.73	06.6	00.01*	01.55*	01.64*	0.00	1.19	1.20
Tense	Tense	01.71	01.77	01.4	01.32*	05.84	08.42	0.00	0.08	0.11
Numeral	P_Nume	03.99	05.27	05.3	02.26	02.20	00.13*	1.62	1.83	0.01
Two-Letter Word	P_Two	10.61	09.79	09.5	01.04*	01.31*	00.31*	0.68	1.13	0.06
Three-Letter Word	P_Thre	19.24	20.18	18.3	00.86*	00.69*	01.52*	0.88	0.78	3.32
Four-Letter word	P_Four	20.46	25.66	25.9	04.20	03.86	00.19*	2.08	30.0	0.07
Word starting with vowels	P_Vowe	27.74	33.36	28.8	04.09	00.72*	03.35	31.61	1.27	20.22
Verb	P_Verb	23.43	21.40	23.6	02.52	00.22*	02.19	4.13	0.06	5.18
Voice	Voices	02.25	01.55	02.1	10.88	01.69*	08.69	0.49	0.01	0.34
Syllable	P_Sylla	151.10	119.70	163.0	02.82	00.88*	03.40	989.31	161.21	1949.20
Post position	P_Post	13.43	35.26	31.3	15.00	11.68	02.06	476.38	322.53	14.95
Clitics	P_Clitic	14.14	34.25	33.4	16.31	14.93	00.52*	404.28	371.69	0.68
Case marker	P_Case	38.65	69.95	63.0	15.15	12.45	02.75	980.00	596.68	47.30
Adverb	P_Adverb	04.39	02.26	02.8	04.22	02.78	01.29*	4.54	2.37	0.35
Conjunction	P_Conjun	22.65	42.15	35.5	11.11	07.42	03.48	380.24	165.74	43.90
Total								3458.37	1754.2	2105.60
SQRT								58.81	41.88	45.89

* not significance at 5% level

Table 1. Mean value, Two-samples t-statistic values and Euclidean distance values

Factors	MB			VK			SI		
	Eigen values	Percentage of variance	Cumulative percentage	Eigen values	Percentage of variance	Cumulative percentage	Eigen values	Percentage of variance	Cumulative percentage
1	2.223	12.3	12.3	2.018	11.2	11.2	2.492	13.8	13.8
2	1.560	8.7	21.0	1.645	9.1	20.3	1.590	8.8	22.7
3	1.420	7.9	28.9	1.463	8.1	28.5	1.483	8.2	30.9
4	1.270	7.1	36.0	1.329	7.4	35.9	1.217	6.8	37.7
5	1.167	6.5	42.4	1.150	6.4	42.3	1.135	6.3	44.0
6	1.103	6.1	48.6	1.068	5.9	48.2	1.094	6.1	50.1
7	1.035	5.7	54.3	1.048	5.8	54.0	1.008	5.6	55.7
8	1.013	5.6	59.9	1.004	5.6	59.6	0.966	5.4	61.0
9	0.961	5.3	65.3	0.943	5.2	64.8	0.903	5.0	66.0
10	0.945	5.3	70.5	0.916	5.1	69.9	0.876	4.9	70.9
11	0.856	4.8	75.3	0.843	4.7	74.6	0.829	4.6	75.5
12	0.837	4.6	79.9	0.784	4.4	78.9	0.772	4.3	79.8
13	0.758	4.2	84.1	0.745	4.1	83.1	0.720	4.0	83.8
14	0.702	3.9	88.0	0.738	4.1	87.2	0.673	3.7	87.6
15	0.646	3.6	91.6	0.685	3.8	91.0	0.658	3.7	91.2
16	0.546	3.0	94.7	0.580	3.2	94.2	0.597	3.3	94.5
17	0.450	2.8	97.4	0.540	3.0	97.2	0.538	3.0	97.5
18	0.460	2.6	100.0	0.502	2.8	100.0	0.447	2.5	100.0

Table 2. Factor analysis-Initial statistics

Factors	MB	VK	SI
FACTOR 1	(.75105) P-VOWE (.67532) P-VERB (.55204) P-TWO (.58042) P-THRE (.66074) P-FOUR	(.65782) P_CASE (.56787) P_POST	(.71539) P_VERB (.54890) P_FOUR (.64105) P_SYLLA (.34474) P_VOWE
FACTOR 2	(.83945) P_CLITIC (.85563) P_CASE	(-.53092) P_NOUN (.73761) P_NUME (-.68164) P_VERB	(.77953) P_CASE (.76640) P_CLITIC
FACTOR 3	(.70597) P_NOUN (.82167) P_PRO	(.75332) P_FOUR (-.67378) P_THRE (.59352) P_TWO (.75897) P_PRO	(-.66287) P_PRO (.62711) P_NOUN
FACTOR 4	(-.66491) TENSE (.69586) P_NUME	(.79795) P_ADVERB (-.39408) TENSES (.83215) P_SYLLA (.66924) P_CONJON (.71794) VOICE (.44038) P_VOWE (.73773) P_INF (.46984) P_INT (.56534) P_CLITIC	(-.71236) VOICE (.59945) P_CONJON
FACTOR 5	(.73510) VOICE (.62676) P_POST		(.79019) P_TWO (.56156) P_ADVERB
FACTOR 6	(.66129) P_SYLLA		(-.76712) TENSE
FACTOR 7	(.18315) P_INT (.66689) P_INF (.57019) P_ADVERB (-.52044) P_CONJON		(.10776) P_INF (.61875) P_NUME (.39798) P_POST (.76640) P_CLITIC (.72258) P_THRE

Factor scores are given in the brackets

Table 3. Grouping of stylistic features according to factor scores