A Study of Ambiguous Authorship in Tamil Articles using Multivariate Statistical Analysis

R. Lakshmi Priya
Department of Statistics
Dr. Ambedkar Govt. Arts College
Vyasarpadi, Chennai, India

G. Manimannan
Department of Statistics
Madras Christian College
Tambaram, Chennai, India

ABSTRACT
One of the areas of applications of *Stylometry* is authorship attribution to articles of ambiguous authorship. The present paper deals with the classification of articles of unknown authorship to the articles written by contemporary Tamil scholars of the same period, namely Mahakavi Bharathiar (MB), Subramniya Iyer (SI), and T. V. Kalyanasundaranar (TVK). These three popular scholars had written number of articles on India's Freedom Movement during the pre-independence period and published in the magazine called, *India.* Initially, all the three writers contributed their articles by attributing their names. The oppressive attitude of the then British regime compelled all the three patriots to write articles on the same theme for anonymous publications without mentioning their names. Later, all the three patriots wrote the articles on the same subject for unidentified publications without their names due to the oppressive attitude of the then British rule.

In this paper, the assignment of articles of ambiguous authorship to contemporary writers, namely, Mahakavi Bharathiar (MB), Subramaniya Iyer (SI), and T. V. Kalyanasundaranam (TVK), using the applications of Principal Component Analysis (PCA) and Multivariate Discriminant Analysis (MDA) is discussed. Different sets of stylistic parameters of the above mentioned three Tamil scholars are considered in the analysis and the writing styles of these authors are quantified, using morphological and function words. Almost the entire set
of articles of ambiguous authorship is attributed to Mahakavi Bharathiar.

***Keywords*:** Authorship Attribution, Stylometry, Classification, Stylistic features, Principal Component Analysis, Multivariate Discriminant Analysis.

1. INTRODUCTION
The present field of authorship attribution began with discipline known as stylometry. Stylometry is the study of the quantifiable of human language, or the statistical analysis of literary style (Holmes, 1995). Stylometry mainly concerns itself with authorship attribution studies, although chronological studies on the dating of work within the corpus of an author have also investigated. 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 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.

The availability of modern computing facility has provided a unique opportunity for many stylometricians to introduce many multivariate methods like factor analysis, cluster analysis and correspondence analysis for analyzing experimental observations in multi dimensional space and also to widen the frontiers of stylometry (Roger Peng, 2001).

In recent years, many scholars have successfully demonstrated that this technique of machine learning field can be applied to authorship attribution. Merriam and Mathews (1993, 1994) have trained a multi layer perception network to distinguish the works of Shakespeare and Marlowe. Tweedieet al. (1996) has provided a useful review of the applications of ANNs (Artificial Neural Network) in the area of computational stylometry and has used this machine-learning package for the reanalysis of the Federalist Papers. Kjell (1994) have taken up authorship study using letter-pair frequency features with neural network classification. Recently, authorship identification problem is also attempted by the authors using the Radial Basis Function Network (Chandrasekaran R. and Manimannan G. 2013). This multivariate technique is also used for measuring the extent to which groups of words have similar patterns of high or low use values of various writers.

2. DATABASE
The present research deals with the literary works of three Tamil Scholars in the British Government ruling period, namely, Mahakavi Barathiar (MB), T.V.Kalyanasundaranar (TVK) and Subramaniya Iyer (SI). In the Pre–Independence period, these three scholars have written number of articles on India's Freedom Movement in the news bulletin called *India*. In the initial stage, all the three scholars have written articles by attributing their names.

The three authors have written many articles in the same topic anonymously in the same magazine during the British Government. All the known and unknown articles written on India's Freedom Movement in that magazine were compiled and brought out as a book entitled *Bharathi Dharisanam* in the year 1975. For this quantitative stylistic study, all known articles of these three scholars written on India's Freedom Movement in the year 1906 are considered. Sample lists of variables of this study with their meanings are given in *Table 1* and *Table 2.*

Table 1. List of Twenty Four Function Words

Function Words	Translation	Function Words	Translation
Um	Also	Ai	Unmarked
Aakiyaal	As	Nodu	With
Entraal	For	Lall	With
Aavatu1	For	Aall	Unmarked
Aaka	As	Ukku	To

Mikavum	Very much	Atu	My
Pola	Like	Ill	In
Entru	For	Utan	With
Pearil	On	Um1	At every
Irrunthu	From	Enum	At least
Kooda	Also	Utaiya	Of
Ul	Inside	Pattri	About

Table 2.Lists of Morphological Variables of this Study with Abbreviations

Abbreviations	Variables Name
P_NOUN	Nouns
P_INT	Introductory
P_INF	Intensifiers
P_PRO	Pronouns
P_NUME	Numerals
P_TWO	Two letter Words
P_THRE	Three letter Words
P_FOUR	Four letter Words
P_VOWE	Vowels
P_VERB	Verbs
P_SYLLA	Syllables
P_POST	Postpositions
P_CLITIC	Clitics
P_CASE	Case Markers
P_ADVERB	Adverbs
P_CONJUN	Conjunctions
TENSES	Tenses
VOICES	Voices

Our study is based on thirty four blocks of Mahakavi Bharathiar (MB), twenty seven blocks of Subramaniya Iyer (SI) and thirty one blocks of T. V. Kalyanasundaram (TVK). Here a block refers to ten sentences. The entire un-attributed articles consist of thirty nine blocks in both stylistic parameters.

3. METHODOLOGY

3.1 Factor Analysis

In the present study, factor analysis is initiated to uncover the patterns underlying morphology and function words of stylometry analysis in Tamil corpus. Factor analysis is often used in data reduction to identify a small number of factors that explain most of the variance that is observed in a much larger number of obvious variables. A method of Principal Component analysis is concerned with explaining the variance and covariance structure of given set of variables through a linear combinations of these variables. In general to reduce the variable space to a smaller number of patterns that retain most of the information contained in the original data matrix. In factor extraction method the number of factors is decided on the proportion of sample variance explained. Orthogonal rotations such as Varimax and Quartimax rotations are used to measure the similarity of a variable with a factor by its factor loading. In this paper we used varimax rotation to identify the pattern of authorship attribution problem.

3. DISCRIMINANT ANALYSIS

Discrimination and classification are multivariate statistical techniques concerned with separating sets of observation and with allocating new observation to previously defined groups (R. A. Johonson and D. W. Wichern, 2009). Many researchers have used apriori group of information for classification and model buildings using Discriminant Analysis (DA) to achieve their objectives. In the present study, discriminant analysis is used to exhibit groups graphically and identify the disputed authorship of the Tamil scholars.

In order to explore the discriminating power of the selected variables in authorship attribution we used Multivariate Discriminant Analysis (MDA). MDA involves deriving a variate, the linear combination of two (or more) independent variables that will discriminate best between apriori defined groups. Discrimination is achieved by setting the variate's weight for each variable to maximize the between-group variance relative to the within-group variance (Hair *et al*., 2009).

4. RESULT AND DISCUSSION

Analysis section consists of two parts. Part one to identify the stylistic pattern of each author. Second part deals with the classification of the writing style of each author and assignment of the unknown writings. All the morphological and function word features are highly loaded in the first seven factors based on the Principle Component Analysis (PCA), which cover nearly 75.78% and 71.50%, respectively, of the total variation present in this data set (Table 3 to 8). In other words, these stylistics parameters are grouped into seven factors on the basis of the inter-relationship among themselves.

Table 3. Total variation of MahakaviBharathiar (Morphology)

Component	Rotation of Sum of Squares		
	Total	Percentage of Variance	Cumulative Percentage
1	2.880	15.998	15.998
2	2.206	12.256	28.255
3	2.134	11.854	40.108
4	1.874	10.408	50.517
5	1.654	9.190	59.707
6	1.530	8.499	68.206
7	1.362	7.566	75.772

Table 4. Total variation of T. V. Kalyanasundaram (Morphology)

Component	Rotation of Sum of Squares		
	Total	Percentage of Variance	Cumulative Percentage
1	4.453	24.739	24.739
2	1.750	9.723	34.462
3	1.630	9.057	43.518
4	1.613	8.962	52.481
5	1.593	8.851	61.332
6	1.565	7.694	69.026
7	1.456	6.091	75.117

Table 5. Total variation of Subramania Iyer (Morphology)

Component	Rotation of Sum of Squares		
	Total	Percentage of Variance	Cumulative Percentage
1	3.346	18.589	18.589
2	2.250	12.502	31.090
3	2.173	12.071	43.162
4	1.968	10.935	54.096
5	1.646	9.144	63.240
6	1.573	7.739	70.980
7	1.432	5.231	76.211

Table 6. Total variation of Mahakavi Bharathiar (Function Words)

Component	Rotation of Sum of Squares		
	Total	Percentage of Variance	Cumulative Percentage
1	2.832	14.907	14.907
2	2.462	12.959	27.866
3	2.091	11.004	38.870
4	1.740	9.160	48.030
5	1.599	9.416	57.446
6	1.488	8.832	65.278
7	1.170	7.158	71.436

Table 7. Total variation of T. V. Kalyanasundaram (Function Words)

Component	Rotation of Sum of Squares		
	Total	Percentage of Variance	Cumulative Percentage
1	3.472	18.276	18.276
2	3.400	17.895	36.170
3	2.441	12.845	49.015
4	1.690	8.897	57.913
5	1.486	7.822	65.735
6	1.247	3.565	69.300
7	1.135	2.976	71.276

Table 8. Total variation of Subramania Iyer (Function Words)

Component	Rotation of Sum of Squares		
	Total	Percentage of Variance	Cumulative Percentage
1	4.630	24.366	24.366
2	2.235	11.763	36.129
3	2.066	10.875	47.003
4	2.009	10.575	57.579
5	1.479	7.784	65.362
6	1.322	5.959	71.321
7			

They are named as, the first factor habitual word factor, second function word factor, third morphological factor, fourth tense factor, fifth postposition factor, sixth syllable factor and last conjunction factor.

The second part attempts to use the Multivariate Discriminat Analysis (MDA) as one of the suitable statistical classification tool. Discriminant analysis is a multivariate method developed for testing the significance of two or more pre-defined groups of objects. In this analysis, the three authors, namely Mahakavi Bharathiar, T. V. Kalyanasundaram and Subrmaniya Iyer, are designated as author 1, author 2, and author 3 respectively.

As there are three authors to be differentiated, we get two canonical discriminant functions. The first canonical discriminant function accounts for 90% of the variance between authors. The second canonical discriminant function accounts for the remaining 10% of total between authors variance. Each canonical discriminant function is a linear combination of morphology and is orthogonal to the other in the case of morphology.

In the case of function words category, we get two canonical discriminant functions. The first one accounts about 86% of the variance and the accounts for the remaining 14% of total between authors variance. Each canonical discriminant function is a linear combination of morphology and is orthogonal to the other in the case of function words.

On comparing both stylistics parameters there exist significant canonical correlation between authors. Here (r=0.954), (r=0.963) and (r=0.724), (r=0.832) are first and second canonical discriminant functions which, clearly distinguishes the authors in case of morphology and function words.

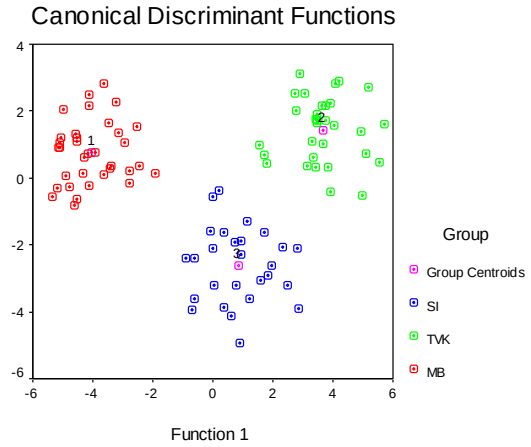

Figure 1. Classification Map with known Articles (Morphology)

Table 9. Classification Results with known Articles

		Predicted Group Membership			Total
	Group	1	2	3	
Original Count	1	34	0	0	34
	2	0	31	0	31
	3	0	0	27	27
Percentage	1	100.0	0	0	100.0
	2	0	100.0	0	100.0
	3	0	0	100.0	100.0

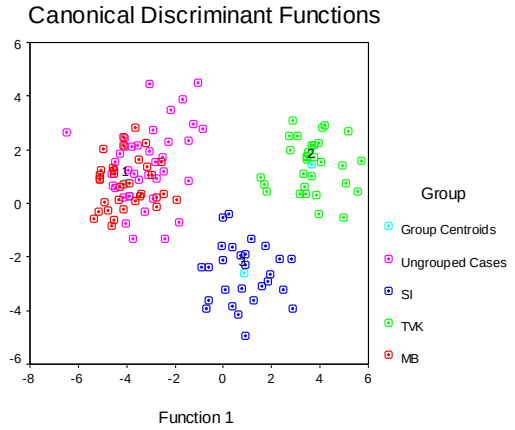

Figure 2. Classification Map with known Articles (Morphology)

Table 10. Classification Results with unknown Articles

	Predicted Group Membership				Total
	Group	1	2	3	
Original Count	1	34	0	0	34
	2	0	31	0	31
	3	0	0	27	27
	Unknown	39	0	0	39
Percentage	1	100	.0	.0	100.0
	2	.0	100	0	100.0
	3	.0	.0	100	100.0
	Unknown	100	.0	.0	100.0

100% of original grouped cases correctly classified

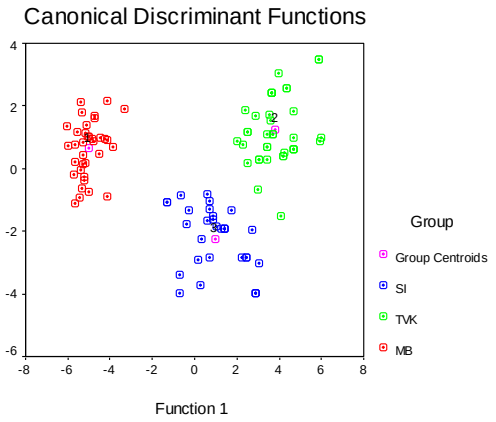

Figure 3. Classification Map with known Articles (Function Words)

Table 12. Classification Results with known Articles

		Predicted Group Membership			Total
	Group	1	2	3	
Original Count	1	34	0	0	34
	2	0	31	0	31
	3	0	0	27	27
Percentage	1	100.0	0	0	100.0
	2	0	100.0	0	100.0
	3	0	0	100.0	100.0

100% of original grouped cases correctly classified

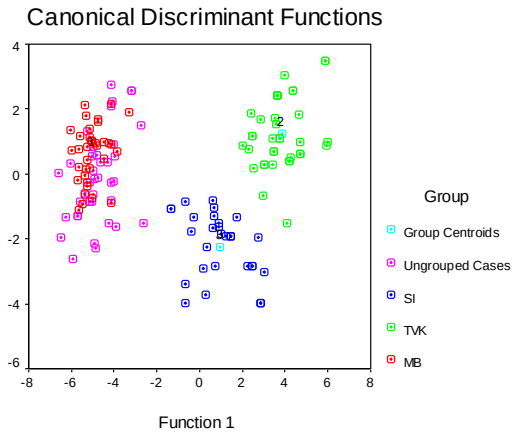

Figure 4. Classification Map with known Articles (Function Words)

Table 11. Classification Results with unknown Articles

		Predicted Group Membership			Total
	Group	1	2	3	
Original Count	1	34	0	0	34
	2	0	31	0	31
	3	0	0	27	27
	Ungrouped case	39	0	0	39
Percentage	1	100.0	.0	.0	100.0
	2	.0	100.0	0	100.0
	3	.0	.0	100.0	100.0
	Ungrouped case	100.0	.0	.0	100.0

100% of original grouped cases correctly classified

The figure 1 to 4 shows the classification map of morphology and Function words parameters. Table 7 to 10 provides the summary of the classification results of this study. The percentages *of cases classified correctly are often considered as* an index of the effectiveness of the derived discriminant functions. The diagonal elements of this matrix are the number of cases classified correctly into groups and the non-diagonal elements are the misclassified cases. The articles of all three authors are classified into three groups correctly.

The overall percentages of cases classified correctly are 100%. This result indicates that all the thirty four blocks of Mahakavi Bharathiar (MB), twenty seven blocks of Subramaniya Iyer (SI) and thirty one blocks of T. V. Kalyanasundaram (TVK) are correctly classified into three different groups. Here a block refers to ten sentences. The entire un-attributed articles consist of thirty nine blocks in both stylistic features are overlapped in MB articles, this result supported the claims made by many scholars that these 39 blocks could have been written by Mahakavi Bharathiar (MB).

5. CONCLUSION

The problem of classification of articles of ambiguous authorship to the articles written by contemporary Tamil scholars, namely Mahakavi Bharathiar (MB), Subramaniya Iyer (SI), and T. V. Kalyanasundaram (TVK), all of them belonging to of the same period, is taken up in the present research. To begin with, all the three writers contributed their articles by attributing their names. The oppressive attitude of the then British regime compelled all the three patriots to write articles on the same theme for anonymous publications without mentioning their names.

The factor analysis results showed that three Tamil scholars writing styles based on morphology and function words. The first factor *habitual word factor*, second *function word factor*, third *morphological factor*, fourth named as *tense factor*, fifth *postposition factor*, sixth *syllable factor* and last *conjunction factor*. This study provides opportunities to introduce statistical techniques for identifying the special stylistic parameters and also for quantifying the writing styles of three Tamil scholars, namely, MB, TVK and SI using eighteen stylistic parameters.

The MDA yielded the output which assigned all the 39 blocks of unknown authorship to Mahakavi Bharathiar (MB) for morphological as well as functional variables. This result supported the claims made by many scholars that these 39

blocks could have been written by Mahakavi Bharathiar (MB).

6. REFERENCES

[1] Bailey, R. W. (1979), **The Future of Computational Stylistics**, Association for Literary and Linguistic Computing Bulletin, Vol. 7, 4-11. England.

[2] Chandrasekaran, R. and Manimannan, *G.* (2013), Use ofGeneralizedNeuralNetworkinAuthorshipAttribution,InternationalJournal of ComputerApplication,Volume6, Number4, ISSN 0975-8887,New York, NY10001, USA

[3] Hair et. al. (2009), Multivariate Data Analysis, Pearson Education, Sixth Edition, India.

[4] Holmes D.I. and Forsyth, R.S. (1995),The Federalist Revisited: New Directions in Authorship Attribution, Literary and Linguistic Computing, 10, 111-127, England.

[5] Herdan G. (1941). The Advanced Theory of Language as Choice and Chance. *Oxford, The Hegue.*

[6] Kjell, B. (1994), Authorship Determination Using Letter-pair Frequency Features with Neural Network Classifiers, Literary and Linguistic Computing, Vol.9, 119-124, England.

[7] Manimannan, G. and Bagavandas, M. (2001), Authorship Attribution : The case of Bharathiar, National Conference on Mathematical and Applied Statistics, Department of Statistics, Nagpur University, Nagpur.

[8] [8] Merriam, T. and Mathews, R. (1993), Neural Computation in Stylometry I: An Application to the Works of Shakespeare and Fletcher, Literary and Linguistic Computing, Vol.8, 203-209, England.

[9] Merriam, T. and Mathews, R. (1994), Neural Computation in Stylometry II: An Application to the Works of Shakespeare and Marlowe, Literary and Linguistic Computing, Vol.9, 1-6, England.

[10] Roger Peng and Nicolas Hengartner. (2001). Quantitative Analysis of Literary Style. *University of California*, Los Angeles, CA90095.

[11] Richard A. Johnson, Dean W. Wichern (2009), Applied Multivariate Statistical Analysis, PHI Learning Private Limited, New Delhi, India.

[12] Tweedie, F. J, Singh, S., and Holmes, D. I. (1996*)*,Neural Network Applications in Stylometry. The Federalist Papers, Computers and the Humanities, 39(1), 1-10, 1996.