Metodološki zvezki, Vol. 4, No. 2, 2007, 149-163

Authorship Attribution and Statistical Text Analysis

Rohangiz Modaber Dabagh[1]

Abstract

In the study of ancient literature, a major problem is to deal with uncertain authorship. Ambiguity about authorship is not limited to the works from remote era. Different reasons cause uncertainty in authorship, such as reproduction of books by hand, prestige, having good sells for works with forged reputable names on them, and sometimes social or political pressures. Whatever the reason, authorship case studies offer the statistician an interesting opportunity to deal with various applied problems, where many standard statistical techniques have been introduced.

In statistical analysis of literary texts one tries to apply an objective methodology to works that have received impressionistic treatment for a long time. In subjective analysis of literary style, experts use literary style of the text, which is not quantifiable, as an important criterion in their judgments. Subjective approach can rarely lead to a unique solution acceptable to all the scholars. Statistical quantitative methods provide objective components for judgments.

In the quantitative approach, by carefully analyzing the style of the text one tries to find out how to characterize the style of an author numerically and determine sets of features (variables) in a text that most accurately describe the author's style.

Much work has been done covering different aspects of this field. Different variables are proposed as distinguishing characteristics of writers, a wide range of mathematical methods is employed, and there is still a lot to be done in the future.

The paper presents a brief history and a review of the statistical analysis of literary style, looks at several variables that have been used as stylistic criteria of authors, as well as the methods used. This is followed by some illustrations on Farsi text, implying that there are some general rules that hold for different languages.

[1] Faculty of Mathematics, Alzahra University, Tehran, Iran; rmodaber@Alzahra.ac.ir

1 Introduction

Authorship attribution is one of the applications of stylometry; and stylometry is the science of measuring literary style. It is believed that every author has an inherent style of writing, which is peculiar to himself. A traditional literary scholar captures the peculiarities in style of an author by impression. What statisticians offer to this filed is to help quantify the style, and hence to change a subjective method into an objective technique which is referred to as “Non-Traditional Stylometry”.

Methods have been tried on texts of different languages. Here we review a few of early attempts in this field, and have a quick look at different variables and methods used. Using samples from Persian poetry and prose, we show how well statistical techniques can discriminate between authors in Farsi (Persian) language.

2 Previous works

Thomas Corwin Mendenhall (1841-1924), surely was not the first who thought of how to apply statistical methods to linguistic problems, but for sure he was the first who undertook extensive work to show that some simple statistical methods may prove useful to solve questions of disputed authorship. He suggested they may also be utilized in comparative language studies, in tracing the growth of a language, in studying the growth of the vocabulary from childhood to manhood, and in other directions.

Mendenhall (1887) proposed forming relative frequency curve of number of letters per word (word-length), which he called “word-spectrum” or “characteristic curve” as a method of analysis leading to identification or discrimination of authorship. He constructed word-spectra for works of two contemporary novelists; Dickens and Thackeray, and a few other writers, to show that texts with the same average word-length might possess different spectra.

He assumed that every writer makes use of a vocabulary which is peculiar to itself and the character of which is persistent over time. He examined blocks of writings containing 1000 words each to determine the extent of which an author agreed with himself (Figure 1), and the extent of which he differed from others. He found that when the number of words in a block was increased to five thousand and then to 10000 the accidental irregularities began to vanish, the curve became smoother, approximating more closely the normal curve which was assumed to be the characteristic of the writer (Figure 2).

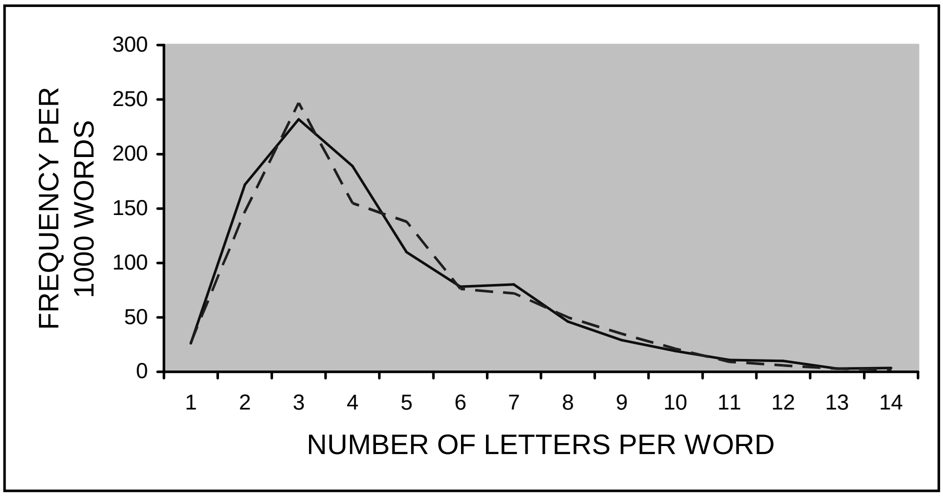

Figure 1: Word frequencies for two samples of 1000 words from "Vanity Fair".

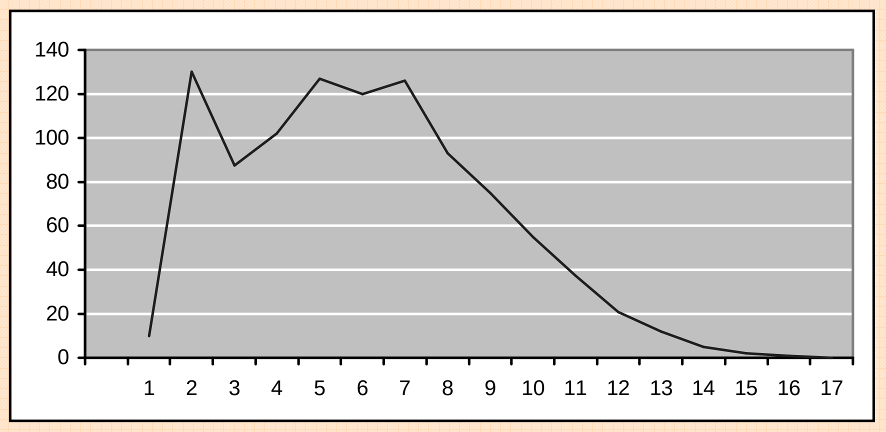

Figure 2: Group of five thousand five hundred words from Caesar's "commentaries".

3 Is Bacon other than Shakespeare?

In an attempt to settle forever Shakespeare Bacon controversy, a controversy which will doubtless remain unsettled forever, Mendenhall (1901), by an intensive word counting for all texts written by Shakespeare and by Bacon, analyzed their spectra and compared characteristic curves of the two authors. He discovered that the most frequent word length (Mode) of Shakespeare was four, in sharp contrast to three being the Mode of Bacon (Figure 3). In this way the conjecture that Shakespeare might be none other than Bacon was rejected.

In the same study, characteristic curve of Christopher Marlowe was found in a close agreement with that of Shakespeare.

Mendenhall's conclusion was later criticized by Williams (1975).

Author:	Shakespeare	Bacon
Mode of word-length:	4	3

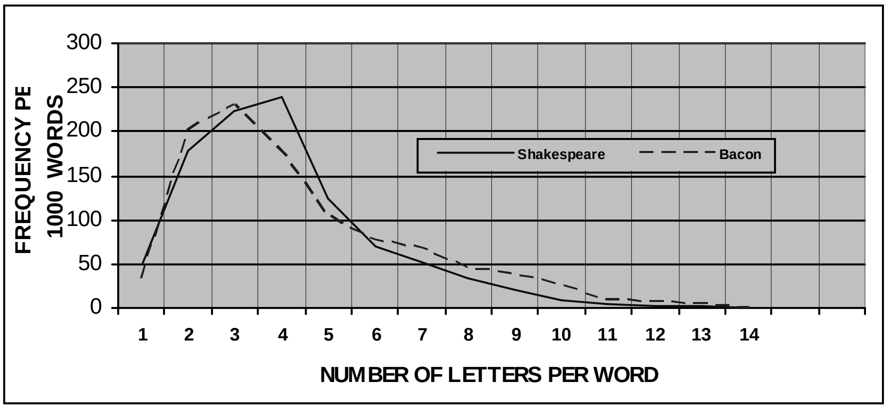

Figure 3: Estimated word frequencies for large samples from works of Shakespeare and Bacon.

4 Was Mark Twain the writer of "Quintus Curtius Snodgrass" (QCS) letters?

Mark Twain's role in the Civil War has been a subject of dispute for years. The evidence regarding Twain's possible military connection in New Orleans was drawn entirely from content of ten letters published in New Orleans' Daily Crescent in early 1861. In these letters, which have largely been credited to Twain and were signed "Quintus Curtius Snodgrass" (QCS), the writer described his military adventures.

On the basis of Mendenhall's method Bringar (1963) applied statistical tests to QCS letters. To determine Mark Twain's characteristic curve 11000 words in total were counted in three groups from writings (before and after 1861) that were indisputably his. These three items formed the control group for the test. Although 11000 words sound short of Mendenhall's work, but the three word groups presented a perfect consistency.

Then ten QCS letters were counted in three groups and their frequency distributions were obtained (Figure 4).

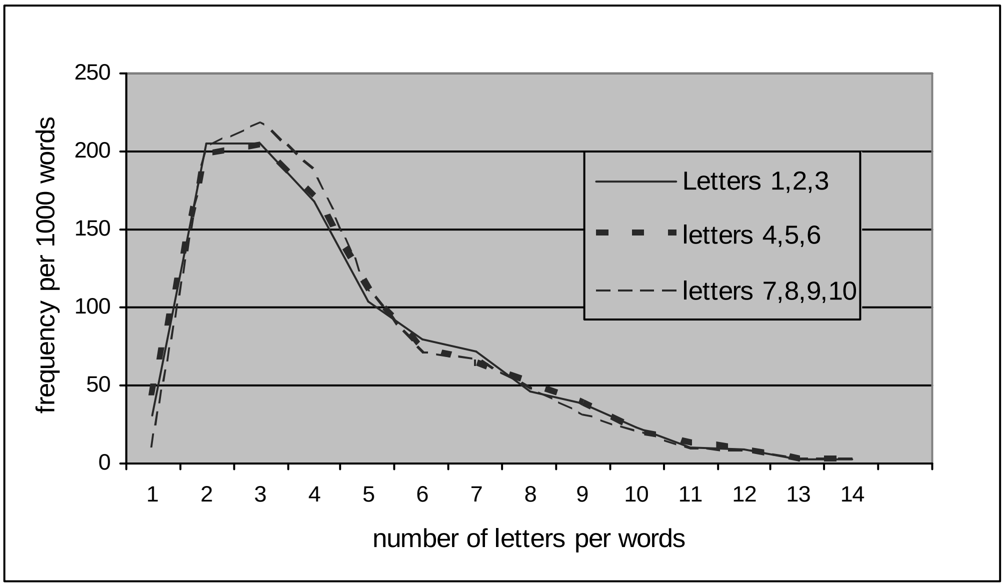

Figure 4: Word frequencies for QCS letters.

Referring to distributions found, Bringar concluded that the curves of the QCS letters were quite unlike those of known Mark Twain's writings (Figure 5); hence, Mark Twain was not the author of the disputed letters. He used a X-squared goodness of fit test and a two-sample t-test to support his conclusion.

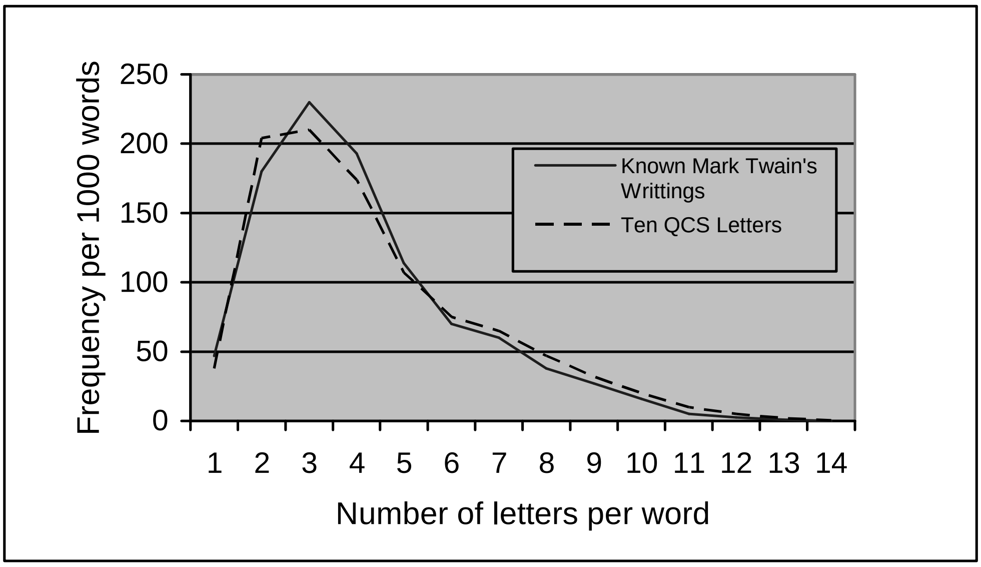

Figure 5: Word frequency for known Mark Twain's writings and QCS Letters.

5 Federalist papers: a favourite testing ground for researchers

The Federalist papers were published anonymously (in 1787 - 1788) by Alexander Hamilton, John Jay, and James Madison to persuade New Yorkers to adopt a new constitution of the United States. Of the 77 essays, 900-3500 words in length, that appeared in newspapers, it is generally agreed that Jay wrote 5, Hamilton 43 and Madison 14 papers. Three are joint papers, and 12 papers (Nos. 46-58, 62 and 63) are of disputable origin between Hamilton and Madison.

Mosteller and Wallace (1964) compared word-usage and word-length distributions in writings by Hamilton and by Madison with that of disputed papers and finally assigned all 12 disputed papers to Madison; a conclusion that could benefit from historians' support.

It is said that this was the first convincing demonstration that stylometry has the power to distinguish the authorship of a text.

6 Variables: Text discriminators

It is thought that every author's style has certain features that are independent of the author's will. The main problem of how to characterize the style of an author is to determine which sets of features in a text most accurately summarize his style. Bailey (1979) lists the general properties for such variables: “They should be salient, structural, frequent, and relatively immune from conscious control”. Much of work has been done and several variables have been suggested to be used to quantify the style of a writer. Some examples follow:

Mosteller and Wallace (1964), and Peng and Hengartner (2002) used distribution of word-length to identify the style of an author.

Holmes (1992) looked at the idea of richness of vocabulary in his studies.

Williams (1940) and Morton (1965) experimented with sentence lengths to quantify the style.

Some have used syntactic and semantic features of the text to represent the style.

Among the most efficient characteristic measures are function word counts. Mosteller and Wallace (1964) based their analysis on word-usage of authors and used function word counts in their seminal work on Federalist Papers.

7 Function words

Function words are words with very little contextual meanings. These words include pronouns, auxiliary verbs, prepositions, conjunctions, determiners, and degree adverbs. Why many authors use the frequency of certain function words to reveal peculiarities in patterns of a writer's style? These parts of speech have more grammatical than lexical functions. It is thought their usage in a text is not much under conscious control of the writer. The frequencies with which they occur in a text tend to be rather stable within texts of the same author. That is they have large variation across authors and relatively little variation among an authors own works. (See, e.g., Mosteller and Wallace (1964), Holmes (1992), Binongo (2003), Peng and Hengartner (2002), Girón, Ginebra and Ri0ba (2005), Riba and Ginebra (2005) among many others.)

Groups of function word counts construct large scale multi-dimensional observations where computers play their efficient role to help the researchers to analyze the data.

8 Multinomial statistical techniques

On the merit of growing power of computers, both in statistical analysis, in text-reading and word-counting, each text can be considered as a collection of multivariate observations, where standard multivariate methods may be employed for stylistic identification purposes. Most of these methods operate on stylometric characteristics such as distributions of *word lengths* and the frequencies of certain *function words* which are extracted from the text.

Holmes (1992), in an example of the use of statistical multivariate techniques, used hierarchical cluster analysis to detect changes in authorship of The Book of Mormon.

Peng and Hengartner (2002) used canonical discrimination analysis and principal component analysis to identify structure in the data and distinguish authorship.

Binongo (2003) used principal component analysis in his work on The Royal Book of Oz attributing authorship of the 15th book of Oz (1921) to Thompson rather than Baum.

Riba and Ginebra (2005), and Girón, Ginebra, and Riba (2005) employed correspondence analysis and cluster analysis of multinomial observations in an attempt to settle the debate around the authorship of Tirant lo Blanc (1460-1464), the main work in Catalan literature, which is considered as the first modern European novel. In their conclusion it was remarked that "even though the statistical analysis supports the existence of two authors, it is not up to us to exclude the possibility that the stylistic boundary could be explained otherwise".

9 Do multivariate techniques discriminate between Persian authors?

To test capability of multivariate techniques in discriminating between Persian authors, we selected some arbitrary books from two great Persian poets: Nezami Ganjavi (1141-1209) and Shahriyar (1906-1988), and two contemporary prose writers: Dr. Abdolhossein Zarinkoob (1923-1999) and Simin Daneshvar (1921-). These authors obviously have different styles. We wanted to test how well multivariate methods could distinguish between authors of Farsi writings.

writings of Nezami (N) and Shahriyar (SH):

We used 14 sample blocks from the book of Nezami (Khamseh) and three sample blocks from the book of Shahriyar (Divaan). To obtain samples first we selected random pages of each book to determine starting points to take random pieces of text containing at least 1000 words each as units of our observation (see Table 5 for description of the data). Within each block the frequencies of more than 150 function words were tabulated. Some of these function words had Zero frequency(ies) for at least one sample block. These were omitted from the list of variables. There were left 14 function words altogether with Non-Zero frequencies for all sample blocks. These words and their equivalents in English are listed in Table 1.

Because of different total number of words in each block, we used frequency percentages instead of absolute counts. Table 2 shows frequency percentages of fourteen function words on works of two Poets.

Initially each author's works were examined, using Box-plots by themselves to identify possible outliers or unusual blocks (with respect to the function word counts). Figure 6 shows box-plot of the data from Nezami's book.

Table 1: Fourteen the most frequently used function words and their equivalents in English in writings of Nezami and Shahriyar.

1	Aan	*that*	6	Beh	*to*	11	Dar	*in*
2	Az	*from*	7	Baa	*with / by*	12	Va	*and*
3	Een	*this*	8	Choan	*because / how*	13	Har	*each*
4	Taa	*till*	9	Raa	object-marker	14	Ze	*from*
5	Bar	*on*	10	Keh	relative clause-marker			

Table 2. Frequency percentages of 14 function words, in the works of Nezami and Shahriyar.

sample	**Aan**	**Az**	**Een**	**Baa**	**Bar**	**Beh**	**Taa**	**Dar**	**Raa**	**Ze**	**Keh**	**Va**	**Har**	**Choan**
N1	0.73	2.00	1.36	0.73	1.82	2.18	1.45	2.82	1.18	1.72	3.09	1.72	0.36	1.82
N2	1.61	0.25	0.93	0.76	1.86	3.04	0.59	1.61	1.44	2.11	1.27	3.21	0.08	1.01
N3	1.31	3.05	0.87	0.78	2.26	3.31	0.61	2.79	1.22	1.22	1.74	3.14	0.26	1.39
N4	1.63	2.31	0.86	1.03	1.03	4.80	0.94	1.28	2.23	1.03	2.14	2.23	0.43	0.77
N5	1.55	1.63	0.34	0.34	1.55	2.24	0.86	2.06	2.06	1.63	2.24	3.78	0.77	0.77
N6	0.53	2.11	0.44	0.79	1.76	2.73	0.53	1.93	1.76	1.14	2.02	2.46	0.53	0.53
N7	0.76	1.78	1.18	0.76	1.61	3.13	0.34	1.27	1.86	1.44	2.79	2.79	0.51	0.68
N8	1.58	2.19	0.44	0.96	1.67	3.25	0.18	1.23	2.72	1.58	3.69	1.32	0.26	0.44
N9	2.14	3.48	0.27	0.45	1.61	4.38	0.62	1.16	2.05	2.14	2.23	2.50	0.62	0.62
N10	2.36	3.15	0.44	0.18	1.58	1.84	0.52	1.58	1.05	2.10	2.01	2.89	1.05	0.79
N11	0.99	2.07	0.99	0.66	0.91	2.57	0.08	1.16	1.99	1.66	2.65	2.32	0.50	0.83
N12	0.78	2.69	1.21	0.95	0.95	4.16	0.09	1.13	1.56	1.21	1.99	2.17	0.43	0.52
N13	0.74	2.54	1.31	0.66	1.39	2.38	0.33	3.70	3.53	1.39	3.44	2.46	0.57	0.98
N14	1.04	2.36	0.85	0.28	2.46	2.64	0.38	2.08	0.94	2.27	2.83	1.51	0.38	1.60
SH1	0.90	2.59	0.90	0.70	1.19	0.40	0.20	1.49	1.49	0.20	0.40	4.28	0.70	0.60
SH2	0.20	2.70	1.10	0.70	1.30	0.30	1.00	1.40	0.50	1.10	1.20	4.60	0.40	0.90
SH3	0.40	3.29	0.60	1.20	0.80	0.10	0.30	1.99	1.69	0.70	1.29	4.68	0.20	1.10

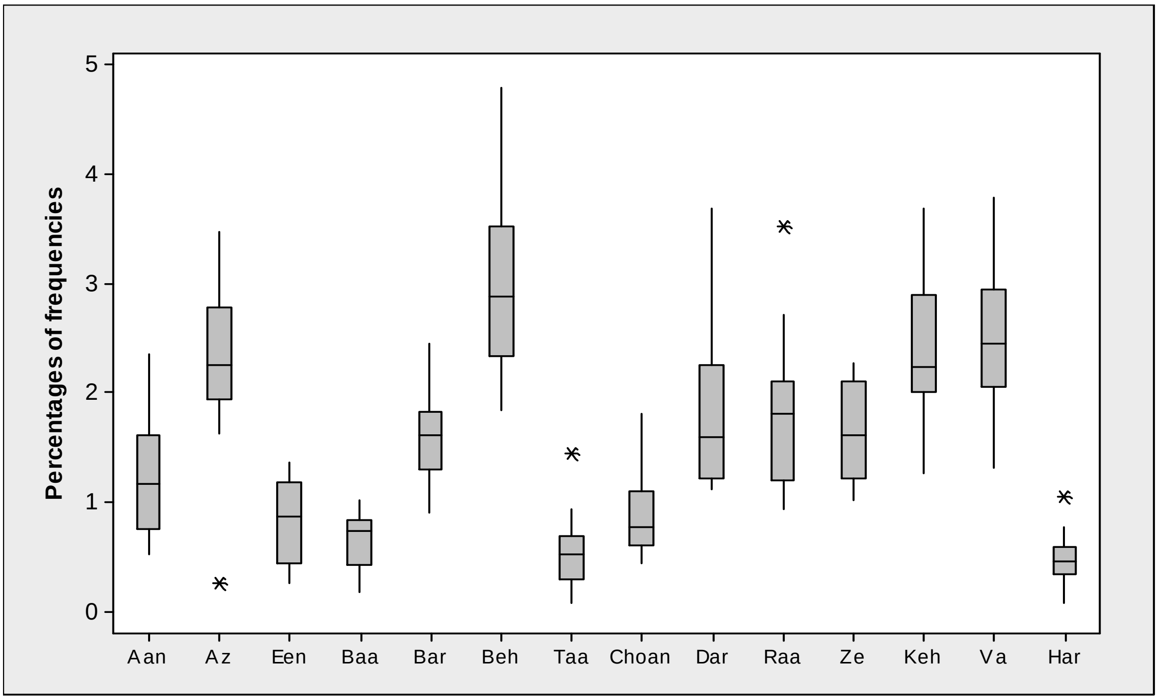

Figure 6: Box-plot of function words: Aan, Az, Een, … , for Nezami's data.

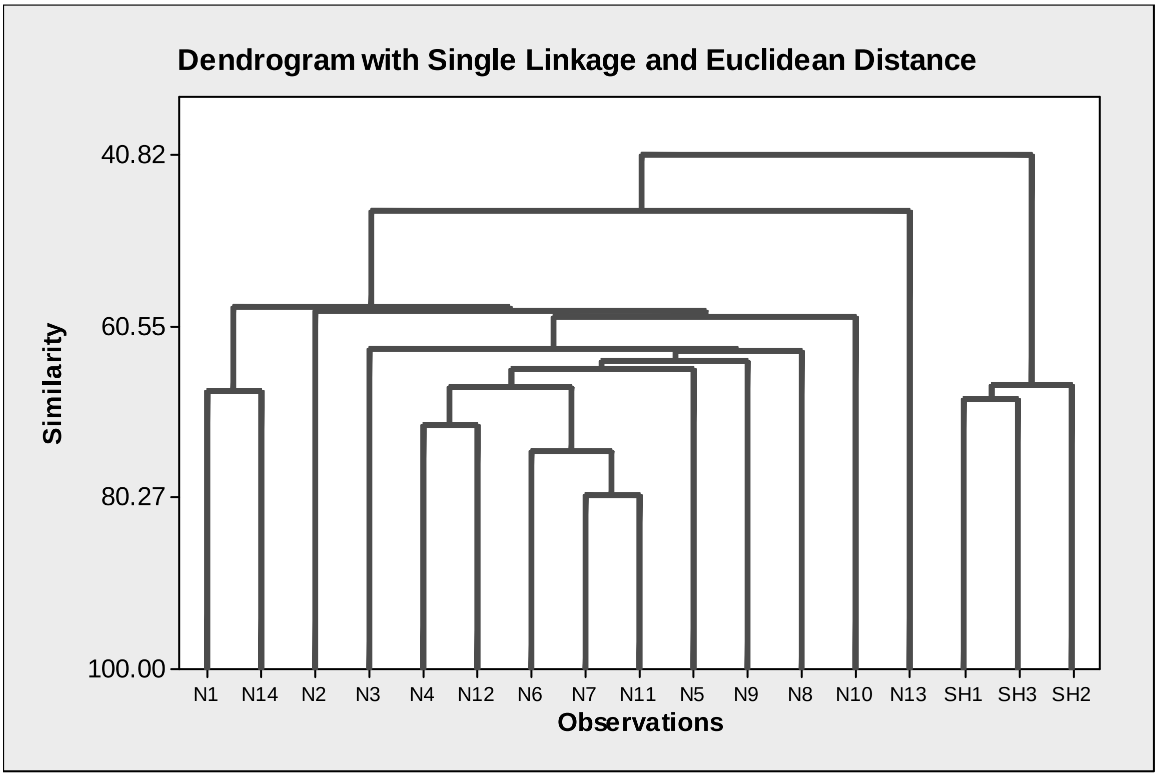

Figure 7: Discrimination between writings of Nezami (N) and Shahriyar (SH).

A single-linkage cluster analysis was applied to 17-sample and 14-variable data set using Minitab. The metric employed was Euclidean distance. The dendrogram obtained (Figure 7) shows good discrimination between the two authors.

Writings of Zarinkoob (Z) and Daneshvar (D):

For a prose example, we used six sample blocks selected out of the three books written by Dr. Zarinkoob (Z), and four sample blocks selected from one book of Simin Daneshvar (D), as units of observation. (See Table 6 for description of the data.)

Samples were obtained in the similar way as above. Frequencies of all function words were determined for all of ten sample blocks. The function words which had Zero frequency(ies) for at least one sample block, were omitted. There we had 9 variables with Non-Zero frequencies for all of the samples to be employed in the analysis. These words and their equivalents in English are listed in Table 3.

A single-linkage cluster analysis using Euclidean distance measure was applied to 10-sample and 9-variable data set (Table 4). Using Minitab we obtained the dendrogram shown in Figure 8. It can be seen that works of the two authors are clearly categorized.

Table 3: Nine the most frequently used function words and their equivalents in English in writings of Zarinkoob and Daneshvar.

1	Va	*and*	4	Beh	*to*	7	Aan	*that*
2	Dar	*in*	5	Raa	object-marker	8	Een	*this*
3	Az	*from*	6	Keh	relative clause-marker	9	Baa	*with / by*

Table 4: Frequency percentages of 9 function words in the works of Zarinkoob and Daneshvar.

	Aan	**Az**	**Een**	**Ba**	**Beh**	**Dar**	**Raa**	**Keh**	**Va**
Z1	0.49652	3.77358	1.09235	0.79444	1.19166	4.17080	1.98610	2.78054	6.65343
Z2	0.68966	2.85714	1.87192	0.49261	2.95567	4.43350	2.36453	3.84236	5.51724
Z3	1.95258	4.88145	0.97629	0.90656	3.20781	4.67225	2.85914	4.53278	7.60112
Z4	1.63416	2.74117	0.73801	1.05430	2.95203	4.21719	3.84818	2.95203	6.37849
Z5	1.78759	3.31230	1.41956	1.78759	1.94532	3.83807	4.15352	3.83807	7.30810
Z6	1.60686	3.32084	1.66042	1.23192	2.89234	2.35672	3.42796	3.69577	8.14140
D1	0.39643	1.38751	0.19822	0.29732	1.68484	0.39643	2.67592	2.97324	8.91972
D2	0.89641	1.69323	0.39841	0.99602	2.29084	1.59363	3.78486	2.78884	5.67729
D3	0.19940	1.89432	0.59821	0.39880	2.99103	1.09671	3.68893	1.89432	7.67697
D4	0.19900	1.59204	0.29851	0.99502	1.59204	0.39801	2.88557	2.78607	4.37811

How does reduction of number of variables involved in the analysis affect the strength of discrimination?

In Figure 9 the dendrogram of analysis shows a good discrimination between authors when 7 variables were used (*Baa* and *Een* omitted). But the method does not work well when five variables (Va, Dar, Az, Beh, Beh, Raa) are employed as it can be seen in Figure 10.

It can be seen that cluster analysis using function words have worked well to discriminate between Persian authors. It is notable that the number of variables and choice of words markedly affect the strength of discrimination. In above experiments more investigation is needed to determine the best sub-set of variables which give the best discrimination.

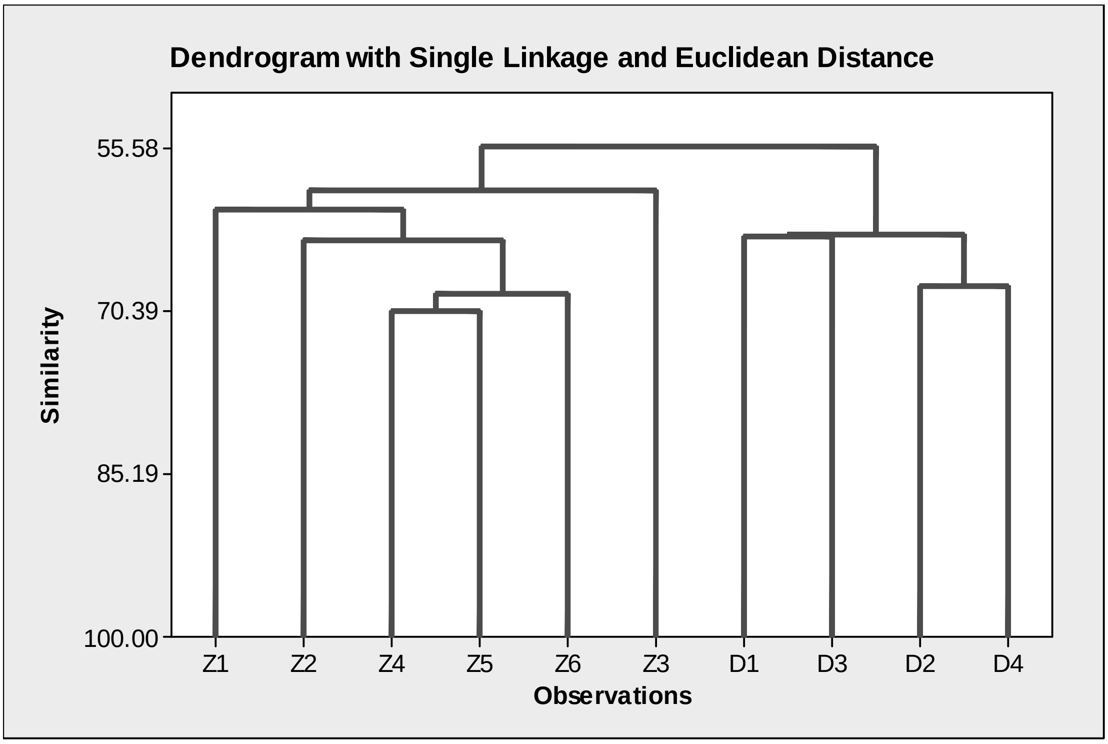

Figure 8: Discrimination between works of Zarinkoob (Z) and Daneshvar (D) using nine variables.

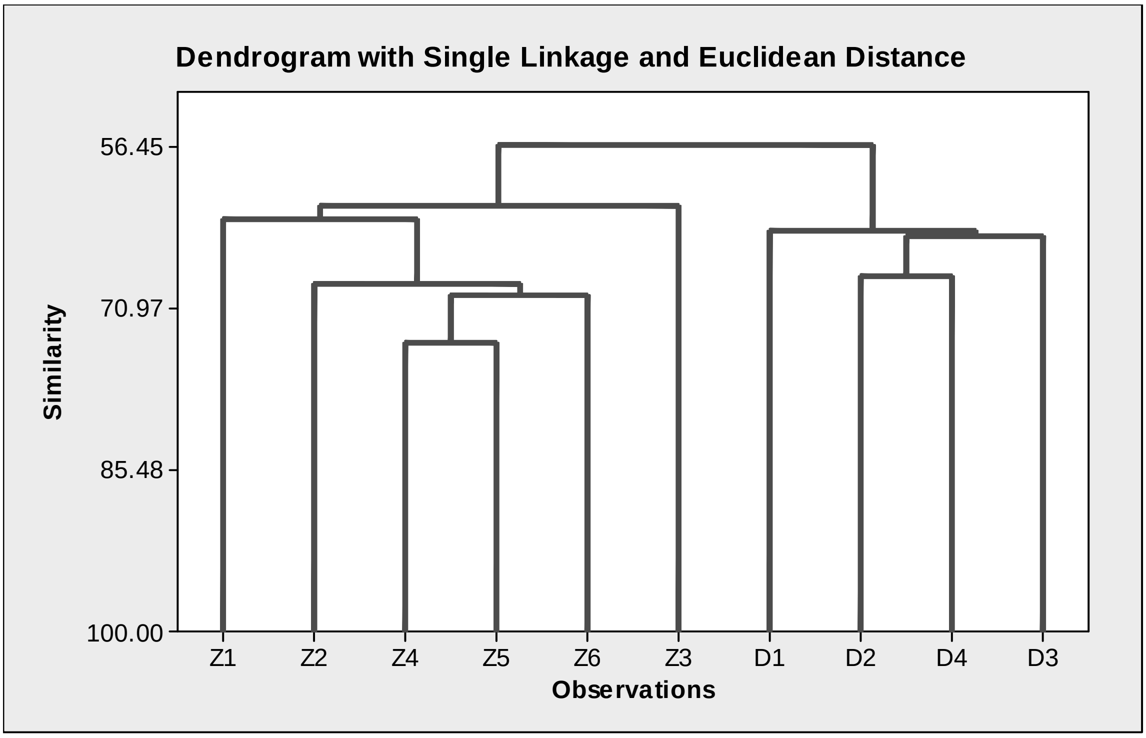

Figure 9: Discrimination between works of Zarinkoob (Z) and Daneshvar (D) using seven variables.

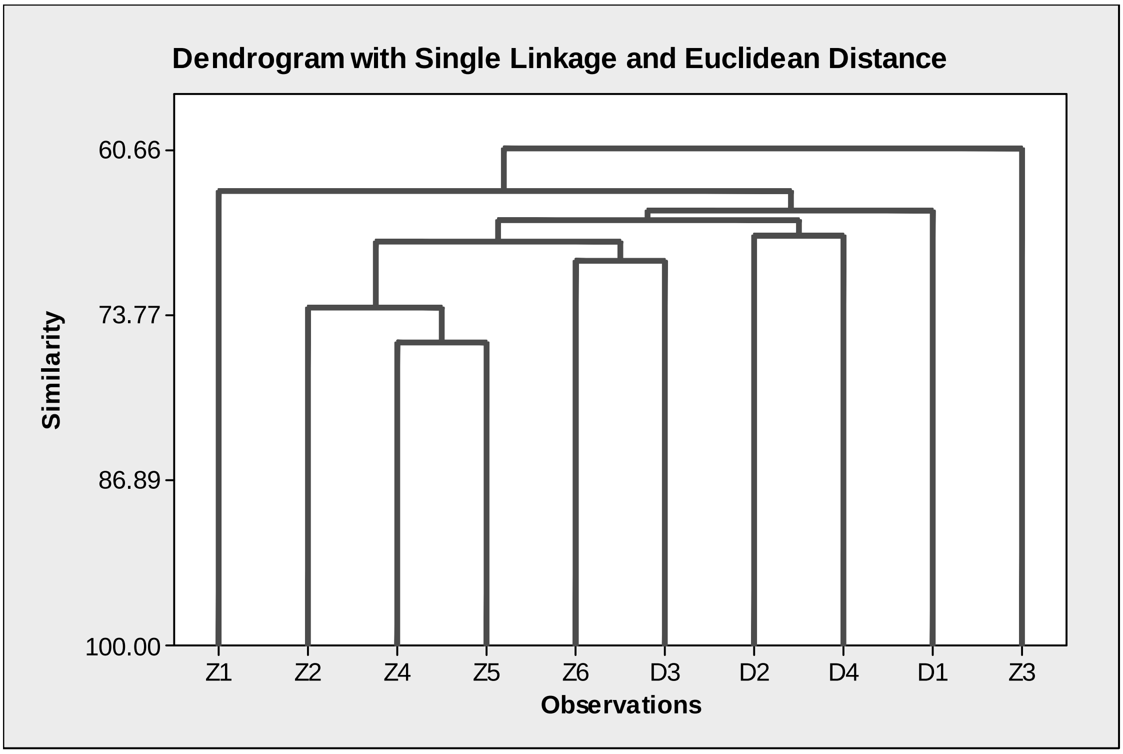

Figure 10: Discrimination between works of Zarinkoob (Z) and Daneshvar (D) using five variables.

Table 5: Seventeen observations on works of two poets Nezami and Shahriyar.

Author	Symbol	Page	Number of words in block	Book
Nezami	N1	639-642	1101	Khamseh
	N2	674-677	1183	
	N3	724-727	1148	
	N4	770-773	1167	
	N5	805-808	1163	
	N6	621-624	1137	
	N7	1322-1325	1181	
	N8	1359-1362	1139	
	N9	1393-1397	1120	
	N10	1405-1408	1142	
	N11	1440-1443	1207	
	N12	1450-1454	1153	
	N13	248-251	1219	
	N14	569-573	1054	
Shahriyar	SH1	415-420	1005	Divaan
	SH2	455-459	1002	
	SH3	491-495	1004	

Table 6: Ten observations on works of two prose writers Dr. Zarinkoob and S. Daneshvar.

Author	Symbol	Page	Number of words in block	Book
Dr. Zarinkoob	Z1	114-117	1007	Az Chizhaye Digar
	Z2	295-298	1015	
	Z3	256-260	1434	Yaddashtha Va Andisheha
	Z4	161-165	1897	Naghsh Bar Aab
	Z5	413-417	1902	
	Z6	555-559	1867	
Simin Daneshvar	D1	26-29	1009	Be Ki Salaam Konam?
	D2	238-241	1004	
	D3	140-143	1003	
	D4	191-194	1005	

10 Conclusions

For more than a century statisticians have found the untapped field of stylometry a great opportunity to introduce and try out different statistical methods. And when their analyses have led to a conclusion in close consistency with that of literary scholars they have felt more confident and motivated to continue their experiments.

The methods proposed so far have provided insight into many literary mysteries, but what has been considered a dream is to introduce a technique that could be used to settle any attributional problem, regardless of genre, language, or time period. Maybe this dream is not that far as other techniques such as Automated Pattern Recognition and Artificial Intelligence as well as other Computer Based Techniques has also come into play.

Acknowledgement

The author would like to thank the referees for their valuable comments. The author would also like to thank two colleagues Malekpor and Ganji-Safar who gave a helpful hand preparing the initial data.

References

[1] Binongo, J.N.G. (2003): Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. *Chance*, **16**, 9-17.

[2] Bringar, C.S. (1963): Mark Twain and the Quintus Curtius Snodgrass Letters: A statistical test of authorship. *Journal of the American Statistical Association*, **58**, 85-96.

[3] Girón, J., Ginebra, J., and Riba, A. (2005): Bayesian analysis of a multinomial sequence and homogeneity of literary style. *The American Statistician*, **59**, 19-30.

[4] Holmes, D.I. (1992): A stylometric analysis of Mormon scripture and related texts. *Journal of The Royal Statistical Society*, Series A, **155**, 91-120.

[5] Mendenhall, T.C. (1887): The characteristic curves of composition. *Science*, **9**, 237-249.

[6] Mendenhall, T.C. (1901): A mechanical solution of a literary problem. *The Popular Science Monthly*, **60**, 97-105.

[7] Morton, A.Q. (1965): The authorship of greek prose. *Journal of The Royal Statistical Society*, Series A, **128**,169-233.

[8] Mosteller, F. and Wallace, D.L. (1964): *Applied Bayesian and Classical Inference: The Case of The Federalist Papers*, NY: Springer.

[9] Peng, R.D. and Hengartner, N.W. (2002): Quantitative analysis of literary styles. *The American Statistician*, **56**, 175-185.

[10] Riba, A.F. and Ginebra, J. (2005): Change-point estimation in a multinomial sequence and homogeneity of literary style. *Journal of Applied Statistics*, **32**, 61-74.

[11] Williams, C.B. (1940): A note on the statistical analysis of sentence-length as a criterion of literary style. *Biometrika*, **31**, 356-361.

[12] Williams, C.B. (1975): Mendenhall's studies of word-length distribution in the works of Shakespeare and Bacon. *Biometrika*, **62**, 207-212.