ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 7, No. 1, pp. 55-64, January 2017
DOI: http://dx.doi.org/10.17507/tpls.0701.07

A Stylometric Analysis of Iranian Poets

Sohrab Rezaei
Allameh Tabatabai University, Iran

Nasim Kashanian
Allameh Tabatabai University, Iran

***Abstract*—This paper presents an investigation into the extent to which the lexical choices made by different poets are distinctive. When a writer, writes, s/he makes lexical choices that make them different from other writers and the writing to some extent can be considered as their fingerprint or in the other word their signature. Authorship analysis by means of textual measurements has been the interest of so many linguists. Authors have their own styles and the stylometrist is interested in finding units which can distinct authors from each other. Statistical analysis has provided different tools for this attempt, by different scholars. Over the past 3 centuries many types of textual tools has been introduced to discriminate different authors objectively that developing in computer programing has played the important role for using these models. In this study by writing a computer program, the styles of different Iranian poets, Attar and Molavi, and Nezami, are investigated in terms of their word length and word richness. Result shows differences between their styles in terms of these parameters. This way of analyzing writing of different authors has some implications in different field of sociolinguistic and TOEFL.**

***Index Terms*—computational programing, vocabulary, authorship attribution, stylometry, statistical analysis**

I. Introduction

According to Amuchi, Faith, Al-Nemrat, Alazab and Layton (2012) the statistical analysis of a literary text has been the interest of many scholars since 1851 when mathematician Augustus de Morgan suggested applying average word length to objectively characterize authorship style. Thomas Mendenhall (1887), a physicist, found an author has a “characteristic curve of composition”. These curves are determined by how frequently an author uses words of different lengths. By this method he compared the works of Shakespeare and Francis Bacon. In 1888 William Benjamin Smith, a mathematician, investigated the curve of style based on average sentence lengths to discriminate the authorial style.

The stylometrist is interested in a unit of measuring; it could be a number or a curve, which is unique to the style of each writer. On the other words, the stylometrist is looking for attributing a number or a curve to a text which by that number or curve the author can be identified.

Statistical analysis has been use to find the writer of anonymous writing, solve the problem of plagiarism, and end to the controversies about the authorship of texts and these issues have a long history, perhaps extending back to the advent of writing. As an instance, for many years there was a controversy over who is the writer of the book Mormon. By the technique of stylometry, researchers were interested to find the writer of the book (Roper, Fields, & Schaalje 2012).

Investigators of authorship have proposed many textual measurements over the past three centuries. There have been introduced over 40 textual measurements that each has its advantages and disadvantages.

In his article Holmes, 1994, classified and defined different textual measurements which have been introduced and used by different scholars. word length, syllables, sentence length, distribution of part of speech, function words, the type token ratio, simpson’s index (D), Yule’s characteristics (K), Entropy, Vocabulary distributions, word frequency, Hapax legomena, are among textual measurements which were introduced by Holmes in that article.

Recently the complex networks theory appears as the suitable approach for studying authorship analysis. Mehri, Darooneh and Shariati (2011) employed complex networks theory to tackle authorship analysis; they focused on some measurable quantities of word co-occurrence network of each for authorship characterization.

As limitations of stylometry, it must be noted here that although stylometry is sometimes referred to as wordprint analogues to fingerprint, it is not that much unique to each author as figureprint. The description of the stylometry as verbal DNA is an even less applicable overstatement (Roper, M., Fields, P. & Schaallje, B., 2012). Writing style is not singularly unique to a person. A writer may adapt his or her style to a particular topic, audience, and genre; sometimes they imitate other’s style. These styles may change through the passage of time. Stylometry can judge about the similarity and differences between texts but it cannot prove personal identification as fingerprints do. However, recently attempts have been made to increase the approximation along with the improvement in statistical method and computational programming.

This paper investigates the differences between poems of different Persian poets, Molavi, Nezami, and Attar from the word length frequency, entropy, and type token ratio view. In this approach we can see their similarity and differences with the lens of word length frequency, entropy.

II. Review of Related Literature

Traditionally stylometry or statistical analysis has been used as a quantitative method to find the author of unanimous text. When reading a unanimous text, a reader may often guess who the writer of text is by using his or her memory and attribution the familiar phrase or words to a familiar author. But this is a very subjective judgment. The statistical analysis of texts not alone be used for authorship analysis but also it has application in plagiarism and autoscoring, and also it can be used by a sociolinguist to compare different style of authors within a period of time or through the passage of time and see the differences in different genres.

In order to analyze a text, there have been introduced thousands of stylistic markers and all are potentially useful for textual discrimination. Among these we refer to word length and word richness or word diversity. For a comprehensive review we direct the reader to Holmes (1985).

A. Word Length

The first tool was used in this study is a word length distribution. This measurement “consists of the relative frequency of 1-character words, 2 character words, etc. In a text, the relative frequency of each word-length is calculated by dividing the total number of words of that length in the text by the total number of words” (Grieve, 2007, p.252).

In 1901, Mendenhall reduced the concordances of Shakespeare and Bacon to distributions of word lengths and plotted these distributions as graphs. His “characteristic curves” serve as an early example of the use of graphics in distinguishing authorship. By comparing the curves he concluded that Bacon probably did not write any of Shakespeare’s work. Brinegar (1963) also applied word length distributions to find if Mark Twain had written the Quintus Curtius Snodgrass (QCS) letters. He used _2 tests and two-sample t-tests on the counts of 2, 3, and 4 letter words to test the agreement of the QCS letters with Twain's known writings. A second series of studies was done by the German physicist Wilhem Fucks (Fucks 1952, Fucks 1954, fucks & Lauter 1965), the syllable based word length distribution were examined and they concluded that word length could be the best indicators. Although such markers seems to work well in some cases but they become failure for their lack of generalization (Smith 1983, 1985).

B. Vocabulary Richness

To find the vocabulary richness a lot of tools have introduced in the litreture.10n formulate were introduced by Grieve which is presented as following:

The formulate for ten of these measurements are presented below, where N is the total number of words in a text (i.e. word tokens), V is total number of vocabulary items in a text (i.e. word types), V_i is the total number of vocabulary items that occur exactly i-times in a text, P_i is the relative frequency of the v-th most frequent vocabulary item in a text, and ɑ is an arbitrary constant.

(1) TTR = Type-Token Ratio= V/N
(2) K= 104 (∑ i2 Vi - N)/ N2
(3) R= V/ √N
(4) C=log V/log N
(5) H= (100 log N)/ (1-V1 / V)
(6) S = V2 /V
(7) K = log V/ log (log N)
(8) LN = (1- V2)/(V2 log N)
(9) Entropy = -100 ∑pv log pv / log N
(10) W = N v-ɑ. (Grive,2007, pp.252-253)

C. Type Token Ratio

As the Type token is sensitive to the length of the text, the type token is limited to the first n-number of words in the shortest writing sample.

Tallentire (1973) asserts lexical markers are the obvious method to initiate stylistic investigations, since more documents lie at the lexical level than at any other in the structure of computed concordances. Richness or diversity of an author’s vocabulary is one of the important notions in stylometry analyses. The basic idea is that each writer has specific range of vocabularies which favors most and chooses more in his or her writing. This feature which is approximately unique to each individual may best discriminated authors from each other. Kjetsaa (1979) studied the behavior of E in formula 1 and restricted his study to N= 500. Kjetsaa in his study found an close regular distribution of types per 500 tokens in texts investigated.

According to Holmes (1994), Baker (1988) introduces the inverted of the TTR as “pace” i.e. the degree at which new vocabularies are produced by an writer. In a research of the research projects of Marlowe and Shakespeare, Baker investigated that the pace of a text is “extremely characteristic of an author’s style” and also it is separate of text length and genre. In addition, he suggests the pace contribute to the sophistication and improvement of an author. Chaski (2001) in his study to evaluate author identification techniques, found out that Tape Token Ratio, and Pace are not good technique for discriminating.

D. Entropy

The formula for the entropy introduced and used in this study was extract from the work of Holmes (1994):

$$\text{Entropy} = - \sum p_v \log p_v$$

where p_v is the likelihood of occurrence of the vth word (fund by dividing the number of occurrence of the vocabulary by the total number of the vocabularies in the document).

This variable is based on a thermodynamic concept of a literary text, namely, that with an increase in internal structure entropy decreases and with an increase in disorder or randomness the measure of entropy increases. Since the value will change according to how much text is analyzed, the formula may be refined in order that works of different length may be compared. Using

$$\text{Entropy} = -100 \sum pv \log pv / \log N$$

Absolute diversity for any length text is measured as 100while absolute uniformity remains zero (Holmes, 1994, p.93).

Johnson (1979) says:

There seems to me no doubt of the measures which have been seriously considered in the literature the most satisfactory indexes of diversity for vocabulary studies are those based on estimates of the repeat rate. The claim has already been made in their favor that, on empirical evidence, they are extremely robust with respect to variation in the sample size. The knowledge that they are also unbiased estimates of an easily interpreted population value, together with the added bonus of some associated sampling theory, in my opinion can only enhance their usefulness to those scholars who are concerned with the measurement of style and vocabulary (Johnson, 1979).

III. Method

In this attempt we compare poems of different poets each belongs to different period of time. The significant of this study is that we analysed the whole poems of these poets. The stylometric measurement that we used here was using the word length relative frequency, type token ratio (TTR), and the entropy.

Subjects

We chose our subjects from the poets belong to different period of time brif biography of these famous poets are extract from the ‘en.wikipedia.org’ and ‘preprints.stat.ucla.edu’:

1. Jalāl ad-Dīn Muhammad Rūmī (جلال‌الدین محمد رومی), also known as Jalāl ad-Dīn Muhammad Balkhī (جلال‌الدین محمد بلخی), Mawlānā (مولانا, "our master"), Mevlânâ, Mevlevî(مولوی*Mawlawī*, "my master"), and more popularly simply as Rūmī (1207 – 17 December 1273), was a 13th-century Persian poet, jurist, Islamic scholar, theologian, and Sufi mystic. Rumi's influence transcends national borders and ethnic divisions: Iranians, Tajiks, Turks, Greeks, Pashtuns, other Central Asian Muslims, and the Muslims of South Asia have greatly appreciated his spiritual legacy for the past seven centuries. Rumi's poetry is often divided into various categories: the quatrains (*rubayāt*) and odes (*ghazal*) of the *Divan*, the six books of the *Masnavi*. The prose works are divided into The Discourses, The Letters, and the *Seven Sermons* (en.wikipedia.org).

2. Nizami Ganjavi (Persian: نظامی گنجوی) (1141 to 1209) (6th Hejri century), whose formal name was *Jamal ad-Dīn Abū Muḥammad Ilyās ibn-Yūsuf ibn-Zakkī*, was a 12th-century Persian poet. Nezāmi is considered the greatest romantic epic poet in Persian literature, who brought a colloquial and realistic style to the Persian epic. His heritage is widely appreciated and shared by Afghanistan, Azerbaijan, Iran, Kurdistan region and Tajikistan.His work consist of: The Story of the Seven Princesses, The Fire of Love: The Love Story of Layla and Majnum, Haft Paikar (preprints.stat.ucla.edu).

3. Abū Ḥamīd bin Abū Bakr Ibrāhīm (c. 1110 – c. 1221; ابو حامد بن ابوبکر ابراهیم), better known by his pen-names Farīd ud-Dīn (فرید الدین) and ʿAṭṭār (عطار, "the perfumer"), was a Persian[2][3][4] Muslim poet, theoretician of Sufism, and hagiographer from Nishapur who had an immense and lasting influence on Persian poetry and Sufism. The question whether all the works that have been ascribed to him are really from his pen has not been resolved. This is due to two facts that have been observed in his works:[3] There are considerable differences of style among these works. Some of them suggest the author's allegiance is to Sunni Islam; others, to Shia Islam. Classification of the various works by these two criteria yields virtually identical results. The German orientalist Hellmut Ritter at first thought that the problem could be explained by a spiritual evolution of the poet (en.wikipedia.org).

4. Materials

The books were gathered from the http://ganjoor.net/ in the HTML format, and then by writing a program changed the format of the HTML format to the text format to be understandable by our next computer program. We wrote a code using Phyton for parsing analyze.

IV. Result and Discussion

1. similarity and differences between poems of different Persian poets with the lens of word length frequency

1.1 Molavi

Figure.1 exhibits the curves of five groups' words, each of one thousand words in raw, from Masnavi-e-Manavi, and it is presented to see the variation among groups of the same author within the same book based on a relative small number of words, the numerical analysis of which is shown in table1.

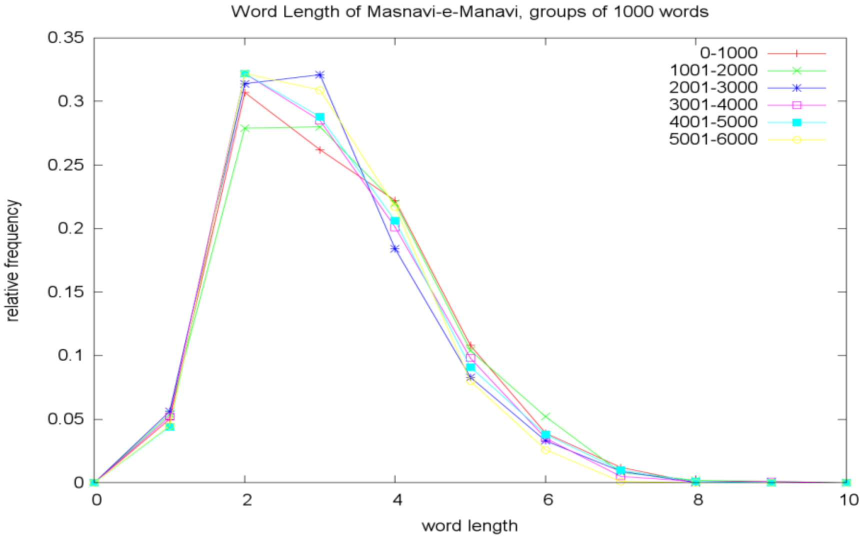

Figure 1 frequency word distribution of five groups of one thousand words each from Masnavi-e-Manavi

In order to decrease the accidental irregularities we increased the number of words in each groups from one thousand to five thousand, total number of words thirty thousand words. For each group we drew the curves of word length distribution as it is shown in figure2.

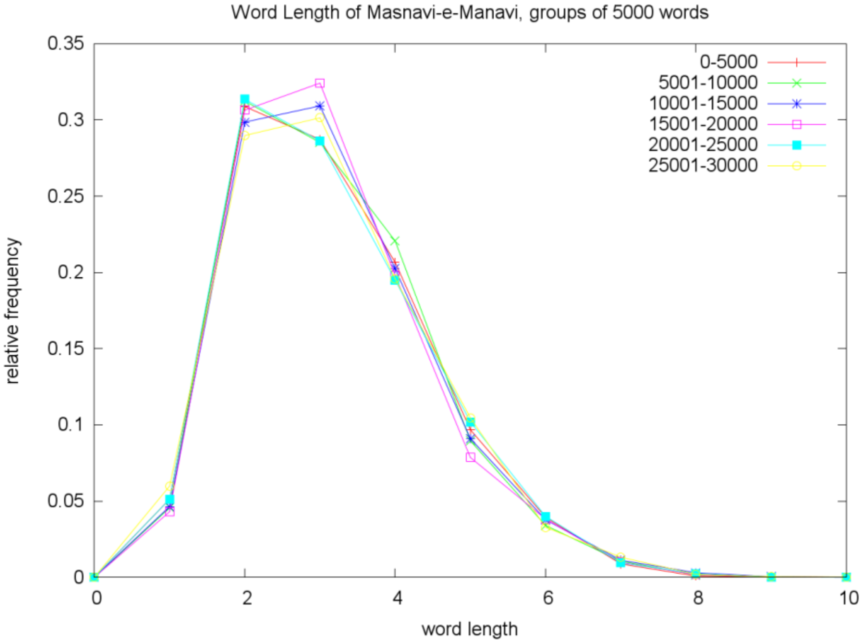

Figure 2 frequency word distribution of five groups of five thousand words each from Masnavi-e-Manavi

We could see from the figure2 that this time while the curves differ considerably, in a general way, they get closer to co each other more than one thousand groups which we had before. Although all the groups of words belong to the same book of one single writer but the curves are different and the most suprising part is that these curves follow only two distinctive shapes not more. We could conclude from this figure that we cannot attribute a single specific curve to a specific author. At this stage the odd things we are dealing is that why these curves follow two specific shapes? We would expect all the curves have the same pattern because they belong to the same book or genre and the same author.

In order to attribute a single curve to Molavi we averaged the word length of those previous five groups of five thousand words from the book Masnavi-e-Manavi and attribute one single curve to the book. And also we did this work for other books of Molavi Ghazalyaat, and Robaeyaat and the result for these three books is illustrated in figure 4.

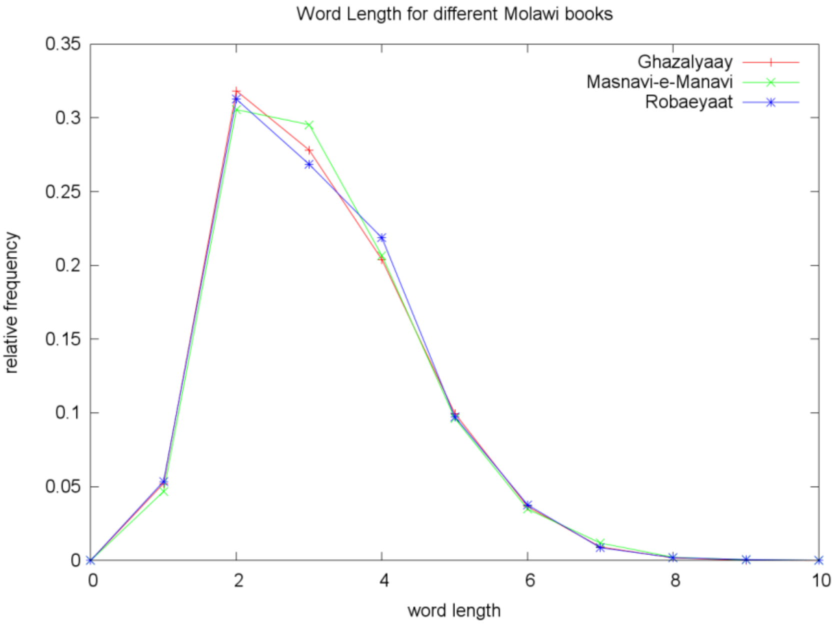

Figure 3 word length distribution of Molave's book, Ghazalyaat, Masnavi-e-Manavi, and Robaeyaat

Interestingly, it is obviously seen that there is a big differences between the distribution curve of these three book, Ghazalyaat, and Robaeyaat are approximately the same but the Masnavi-e-Manavi has different distribution, the differences could be the result of different source, it could be the result of passage of time, or could be result of different genres of each book etc.

1.2 Nezami

In the manner very similar to Molavi, we analyzed the six book of Nezami, Haft Peykar, Khosro-va-Shirin,, Makhzan-ol-Asraar, Leili-o-Majnoon, sharaph nameh, and Kherad Nameh. Result is represented in figure4, which each curve belongs to different books of this poet.

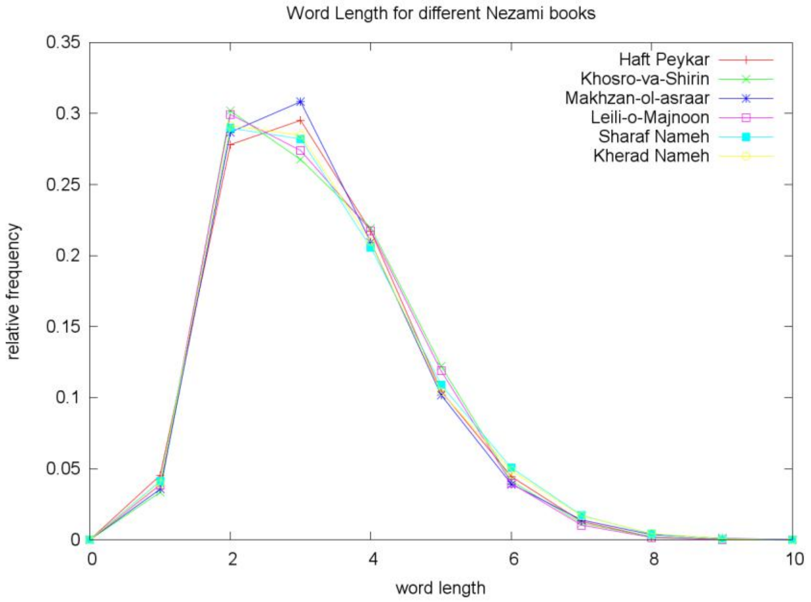

Figure 4 the average word length distribution of Nezami's book, Haft Peykar, Khosro-va-Shirin,, Makhzan-ol-Asraar, Leili-o-Majnoon, sharaph nameh, and Kherad Nameh.

Although figure 4 and figure3 shows different information numerically but it is seen the same pattern like we had in Molavi's curves which is that curves follow only two distinctive curves. One with the high frequency in two-letter words and the other high frequency with three-letter words.In other words Leili-o-Majnoon, Sharaf Nameh, and Kherad Nameh are high in two word frequency and the others high in three- letter words.

1.3 Attar

The next poet that we are interested in is Attar. The curves of his books are illustrated in figure5. Here again each curve belongs to each book of him, Ghazalyaat, Ghasaayed, Mantegh-o-teir, and Pand Nameh.

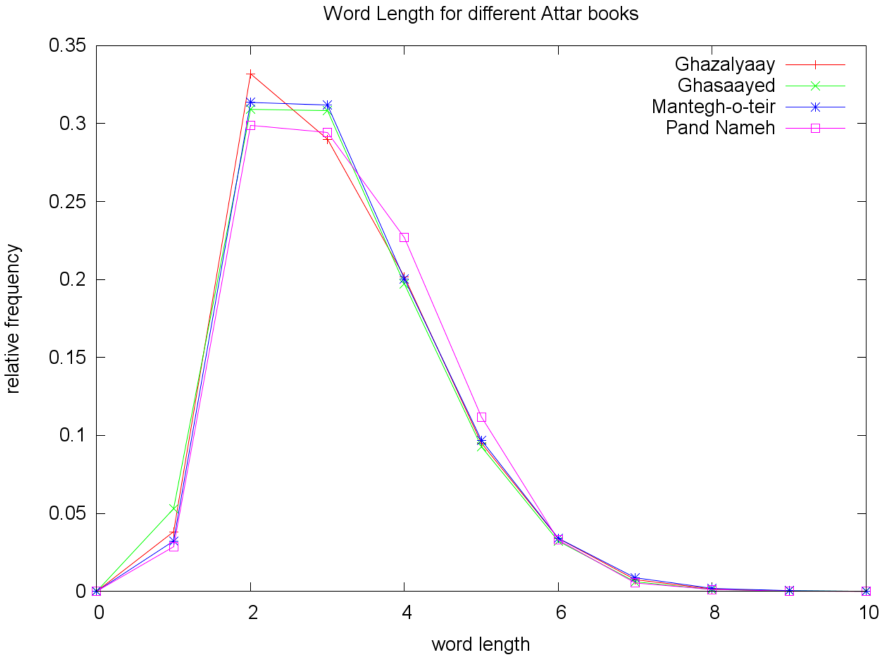

Figure 5 the average word length distribution of Attar's book, Ghazalyaat, Ghasaayed, Mantegh-o-teir, and Pand Nameh.

The relative distribution of words for Ghasaayed, Mantegh-o-teir, and Pand Nameh are the same, 0ne letter words and three-letter words have the same relative frequency. But the book Ghazalyaat has different characteristic which is high in one-letter frequency.

In our final investigation of word length distribution we made average of curves of each poet', books separately and at the end we put these three curves in the same coordinator so that to have the vivid picture of their similarities and differences. Figure 6

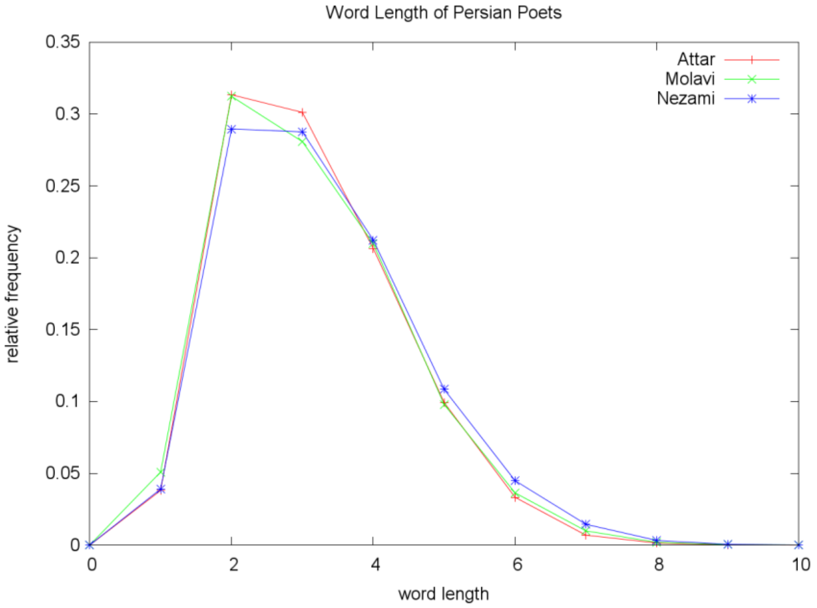

Figure 6 the average word length distribution of Attar, Nezami and Molavi.

It will be seen that the average word length distribution graph from his books for Molavi is high at two-letter words as the same as Attar but with this difference that attar in three-letter word relative frequency has higher rank than Molavi, Nezam has different distribution, Nezami approximately has the same relative frequency in two and three-letter word.

2. similarity and differences between poems of different Persian poets with the lens of Entropy

This part is devoted to the analysis of all the books of all poets from the word richness or word diversity view. If the veriety of words which is used by an author increases in the text the entropy of that text, which is derived from the formula of entropy, become larger. It is hypothesized that if a writer have great lexicon, the entropy of his/her text become higher than the one with the less entropy.

2.1 Nezami

The entropy of words of Nezami's books named Haft Peykar, Khosro-va-Shirin,, Makhzan-ol-Asraar, Leili-o-Majnoon, sharaph nameh, and Kherad Nameh. is represented in figure7.

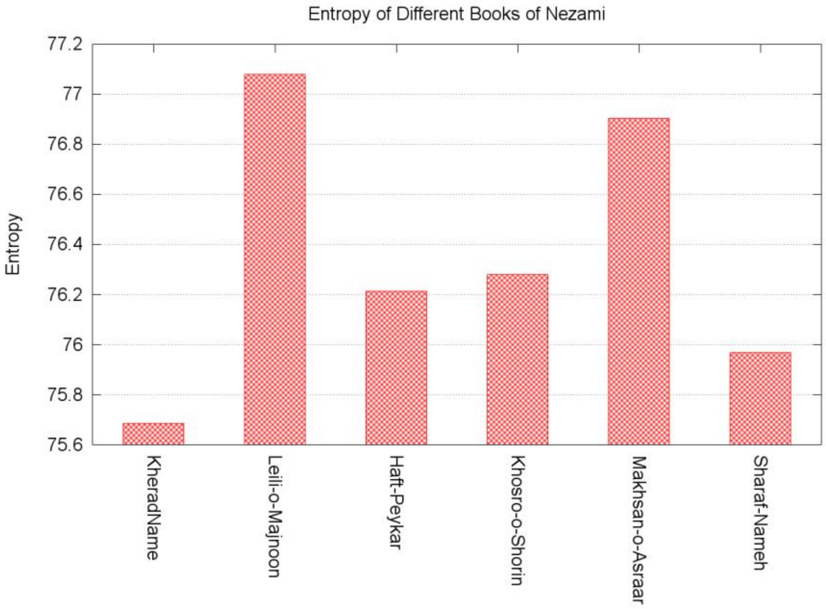

Figure 7 Entropy of the books, Haft Peykar, Khosro-va-Shirin,, Makhzan-ol-Asraar, Leili-o-Majnoon, sharaph nameh, and Kherad Nameh.

Entropies of these different books have the range between71.7 to 74.4; the differences between them are not significantly different. the maximum difference approximately is 2.7. Result of this analyze is illustrated in figure7.

2.2 Attar

Figure8 present the entropy attribute to the Attar's different book, Ghazalyaat, Ghasaayed, Mantegh-o-teir, and Pand Nameh.

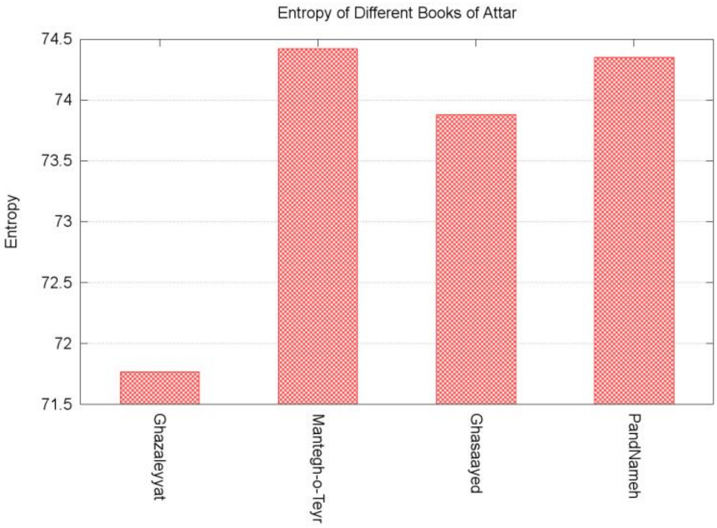

Figure 8 Entropy of the books, Ghazalyaat, Ghasaayed, Mantegh-o-teir, and Pand Nameh.

Information from the figure8 shows the range of Entropy change from 71.6 to 74.3. This time again the differences show no significant difference among the books of Attar.

2.3 Molavi

Figure9 represent the entropy related to the books of Molavi, Ghazalyaat, Masnavi-e-Manavi, and Robaeyaat. This entropy is derived from thousands number of words. The differences change from 74.1 to 76.5. the difference, here again is not that much significant. the maximum difference approximately is 2.4.

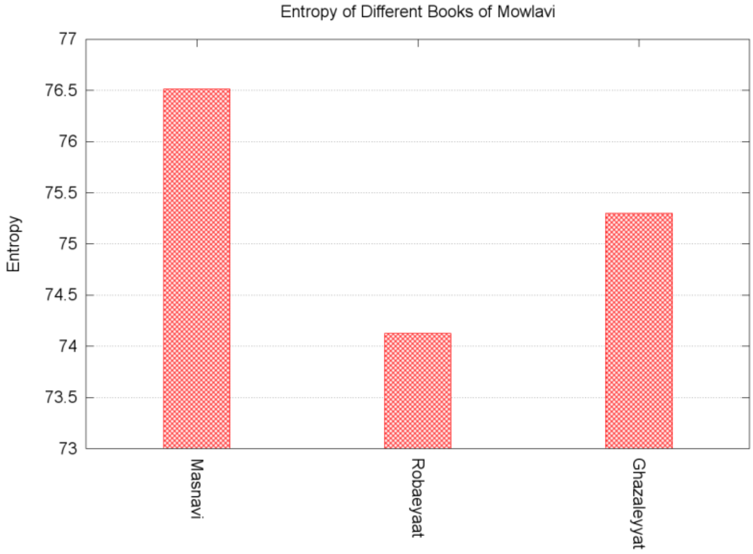

Figure 9 Entropy of the books, Ghazalyaat, Masnavi-e-Manavi, and Robaeyaat

Now let's compare the differences of entropies between Nezami, Attar, and Molavi, Figure10 shows the entropy diagram of these poets. The differences between these poets range from 74.1 to 76.5. In the other word the maximum difference approximately is 2.4. As this difference ranks the same as the maximum different entropy within each poet the entropy could not be considered as a tool to discriminate poets from each other and for doing so we must find the other way around.

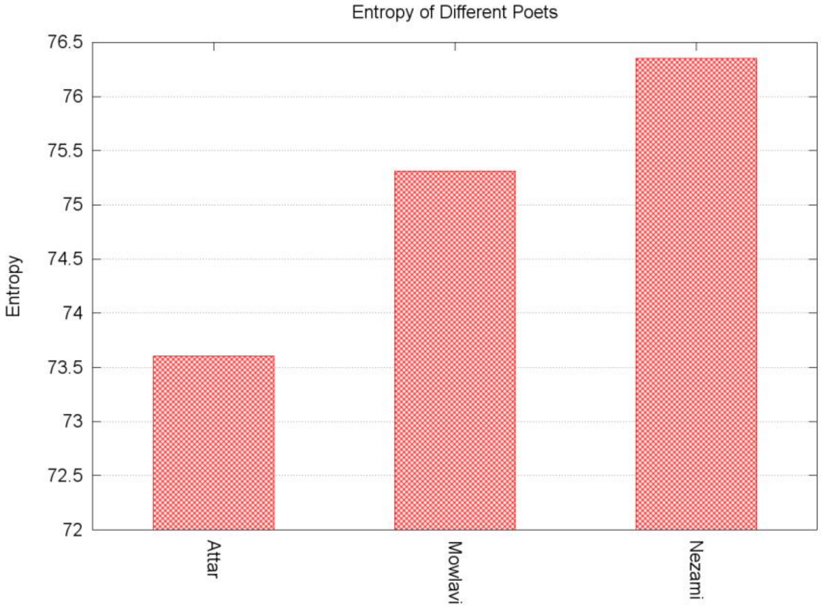

Figure 10 the average entropy of Nezami, Molavi, and Attar

V. Conclusion

The statistical analysis of a literary text has been the interest of many scholars during the last three centuries. Investigators of authorship have proposed many textual measurements over the past three centuries. There have been introduced over thousands textual measurements that each has its advantages and disadvantages.

In this study we aimed to find the characteristics of the books of Nezami, Molavi, and Attar, and find their differences. This paper investigates the differences between poems of different Persian poets, Molavi, Nezami, and Attar, from the word length frequency, entropy view. The computer codes was writtern and was used to analyze huge number of words of their book, Results gave us quantitative characteristics of their relative frequency of word length and entropy. Not only authors had differences in terms of their relative frequency of word length and entropy these measurements also were different within each authors. As these differences had approximately the same magnitude, it was induced that the entropy and word length frequency could not be as a good indicator of discriminating these poets from each other. The new methods have been introduced in literature to compensate this lack of measurement. Even though an author's style may be different, it is not different enough to be considered unique to that author to the exclusion of all other authors in the world. Stylometric characteristics can provide us a comparative description of an author's style, but the writing style exhibited in a text is an indirect and uncertain measure of an author's identity.

REFERENCES

[1] Amuchi, Faith, Al.Nemrat. A,Alazaab, M. & Layton, R. (2012). Identifying Cyber Predatore through Forensic Authorship Analysis of Chat Logs, 2012. Third Cybercrime and Truthworthy Computing Workshop, 2012

[2] Brinegar, 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] Fucks, W. (1952). On mathematical analysis of style. *Biometrika, 39*: 122–9.

[4] Fucks, W. (1954). On Nahordnung and Fernordnung in samples of literary texts. *Biometrika, 41*, 116–32.

[5] Fucks, W. and Lauter, J. (1965). Mathematische Analyze des Literarischen Stils. In Kreuzer, H. & Gunzenhausers, R. (eds), *Mathematik und Dichtung*. Munich: Nymphenburger Verlagsbuckhandlung.

[6] Grieve, J. (2007). Quantitative authorship attribution. An Evaluation of Techniques. *Literary and Linguistic Computing, 22*(3), 251-270.

[7] Holmes, D. I. (1985). The analysis of literary style-a review. *J. R. Statist. Soc. A, 148***,** 328-341.

[8] Kjetsaa, G. (1979). And quiet flows the Don through the computer. *Association for Literary and linguistic computing Bulletin,* 248-256.

[9] Mendenhall, T. C. (1887). The characteristic curves of composition. *Science*, 11, 237–49.

[10] Mehri, A., Darooneh, A.H., & Shariati, A. (2011). The complex netwrorks approach for authorship attribution of books. *Physica A, 391*, 2429-2437.

[11] Roper, M., Fields, P. & Schaalje, B. (2012). Stylometric Analyses of the Book of Mormon: A Short History. *Journal of the Book of Mormon and Other Restoration Scripture 21/1*, 28–45.

[12] Smith, M. W. A. (1983). Recent experience and new developments of methods for the determination of authorship. *Assosiation for Literary and Linguistic Computing Bulletin, 11*, 73-82.

[13] Smith, M. W. A. (1985a). An Investigation of Morton's Method for the determination of authorship. *Style, 19,* 341-368.

[14] Smith, M. W. A. (1985b). An Investigation of Morton's Method to distinguish Elizabethan playwrights. *Computer and Humanities, 19,*3-21.

[15] Tallentire, D. R. (1973). Towards and Archive of Lexical Norms- A Proposal. In The Computer and Literary Studies (eds Aitken, A. J. & Bailey R. W. & Hamilton-Smith, N.). Edinburgh University Press.

[16] en.wikipedia.org.

[17] preprints.stat.ucla.edu.

Sohrab Rezaei was born in Khalkhal Iran 1968. He received his PH.D. degree in teaching English as a Foreign Language from Istanbul University, Turkey in 2005. He has got his M.A. in Teaching English as a Foreign Language, Islamic Azad University, Tabriz, Iran, 1996 and his B.A.in English Translation, Allame Tabataba'i University, Tehran, Iran, 1992.

He is currently an assistant professor in the Faculty of Persian literature and Foreign Languages, ,Allameh Tabatabai University, Tehran, Iran. His research interests include sociolinguistics and education. At the same time he is working with different universities nation-wide. He has got several positions in Iranian Ministry of Science, Research and Technology as well as Ministry of Education.

Nasim Kashanian, was born on August 18, 1976, in Tehran, Iran. She has degrees in M. S. in Astrophysics, 1999-2002, Institute for Advanced Studies in Basic Science (IASBS), Zanjan, Iran, and B. S. in physics, 1994-1998, Alzahra university, Tehran, Iran. Currently she is M. A student in Allame Tabataba'i University, soon to be graduated.

She has taught several institutes as English and physics teacher.

- 2011 - 2013, Invited Teacher, *Allameh Institute*, zanjan, Iran.
- 2009 - 2015, Invited Teacher, *Soufi University*, Zanjan, Iran. teaching English for ESP students
- 2003 - 2007, Invited Teacher, *Azad Islamic University, Zanjan Branch*, Zanjan, Iran.
- 2003 - 2006, Invited Teacher, *Payam-e-Noor University, Zanjan Branch*, Zanjan, Iran.
- 2006 - 2007, Invited Teacher, *Zanjan Universities, Zanjan Branch,* Zanjan, Iran.
- 2009-2012, invite teacher, Parseh Institute Zanjan University, Zanjan.

Mrs Kashanian currently has interdisciplinary approach toward her researches IN THE FIELD OF APPLIED LINGUISTICS. Her research activities are listed below:

- Paper presentation in the 12th international tellsi conference, 25-27 February, 2015. University of Sistan and Baluchestan, Zahedan, Iran.
- Paper presentation in the 2th conference in interdisciplinary Approaches to Language teaching, Literature and Translation studies, Ferdowsi University of Mashhad, Iran.

- Workshop participation on applications of the appraisal Framework, 28-29, 2016. The University of Shahid Chamran University of AHVAZ, Iran.