# *An Experimental Study of Stylometry in Bangla Literature*

Prapti Das, Rishmita Tasmim, Sabir Ismail
Computer Science and Engineering,
Shahjalal University of Science & Technology
Sylhet-3114, Bangladesh.
prapti.das.06@gmail.com, rishmita.sustcse015@gmail.com,sabir-cse@sust.edu

*Abstract*— **Every writer has a different style of writing of their own. By analyzing various kinds of features we can identify and specify some characteristics in a writer's writing which is known as stylogenetics. In this paper we gathered Bangla blogs written by four different Bangladeshi writers. Using machine learning methods we tried to identify special Stylometry features in their writing style. We analyzed various features in their writings, for example, percentage of unique words, word length, sentence length, and frequency of some parts of speech, number of suffix, frequency of first word, second word, second last word and last word of a sentence, counting average number of question marks per document, frequency of word by its position in a sentence etc. We gathered statistical data from analyzing those features and tried to find the variance among these writers using the statistical data.**

***Keywords— stylogenetics, stylometry, machine learning, unique word, word length, sentence length, parts of speech***

## I. INTRODUCTION

Now-a-days language technology has reached to a different level of art. This enables the systematic study of the variation of linguistic properties in texts like author detection, find time period of the author, genre of writing, gender of the author etc. In short it helps to detect the characteristic and personality of the author.

Stylogenetics is clustering-based stylistic analysis of literary corpora [1]. It is a way of analyzing written texts to learn about the writer. Some features are often included unconsciously by the writer, for example gender, age, geographic location of the writer, personality characteristics, etc. A writer may use two specific consecutive words most frequently, may start and end a sentence with specific parts of speech, may write using a specific tense most of the time etc. These are some of the features that can be used to identify a writer.

Stylogenetics helps to detect who is the actual author of a specific writing and may assist in identifying fraudulent writing or in cases where multiple authors claim the ownership of a particular writing. Stylogenetics will analyze the writing and find a specific pattern or characteristic of that writing. Through which we can detect the original author.

Stylogenetics is a new area of research in the field of Bangla literature. Researchers in the past have worked with English literature, particularly with worked famous English writers like Shakespeare, Jane Austen, Charles Dickens etc.

However, there has not been any analysis on Bengali writers. There are many talented and skillful writers in our country and we often find interesting style and pattern in their writings. By analyzing their writing we can distinguish among the characteristics and features of individual writers.

Bangla is our mother tongue and Bengali literature is very rich and famous all around the world, which prompted us to initiate the first stylogenetics research in this area.

In this paper we investigated on the writings of four different Bangladeshi writers namely AH, EZ, MZI, SS (Due to Copyright Protection original name of the writers are not used in this paper).The reason behind choosing these four specific writers is the availability and popularity of their writings. We collected the blogs in a random way. A writer may write on a wide range of topics, for example, political issues, educational infrastructure, economical development, critics/novel etc. In this paper, we tried to identify the specific topic used most frequently by a writer.

Through our analysis we found some features that proved to be important to detect an author. But we did not find any absolute way of author detection. So our end-goal is to find a way which can identify an author perfectly.

The methodology we propose is to work with features like unique words used by an author, word length, sentence length which means number of words used in a sentence, number of parts of speech like conjunction and pronoun, number of question mark etc. There several other possible features, for example, use of punctuations, use of stop words, use of simple/complex/compound sentences etc, which we plan to incorporate into our study in the future. Since in Bengali literature lexical and morphological analyzer is not rich, we could not work with some small details yet.

## II. RELATED WORKS

Stylogenetics is the statistical analysis of variations in literary style between one writer or genre and another. Stylogenetics analyzes written texts to learn about the author. In various papers they worked with different features for the purpose of analysis.

Ramyaa et. al. [6] investigated some features like type-token ratio, mean word length, mean sentence length, standard deviation of sentence length, mean paragraph length, number of commas, question marks, exclamation marks per thousand tokens.

Daelemans et. al [1] adopted a methodology from topic detection research, which included more complex features than the simple lexical features suggested by traditional approaches. They used authors or group of authors as a prediction of class, with clustering methods to indicate the differences and similarities between authors.

On the basis of the stylistic genome of authors, they tried to cluster them into closely related and meaningful groups. They also reported on experiments with a literary corpus of five million words consisting of representative samples of female and male authors.

## III. OUR METHODOLOGY

We collected around 50 blogs each written by four different Bangladeshi writers for this study. The features we used and the detail experimental setup are described below.

### *A. Features Used :*

1) Word frequency:

Word frequency means how many times a word was used in a writers writing. We counted frequency of both one word and two consecutive words. This is used as a feature to find the most frequent words used by a writer.

2) Word length:

The distributions of words of different length have been used as a feature. We calculated the frequency of words of different length.

3) Sentence length:

The number of words present in a sentence is used as a feature.

4) Type-token ratio:

The type-token ratio is *N/V*, where *V* is representing the size of the vocabulary of the sample, and *N* is representing the number of unique words. It is a measure indicating the vocabulary richness of an author.

5) Distribution of parts-of-speech:

Syntax-based features are not under the conscious control of the author and therefore it is considered as a feature. We mainly calculated the number of conjunction and pronoun in the texts.

6) Position specific word frequency:

We counted the frequency of words in first, second, last, second last and other random positions.

7) Writer specific unique words:

We created a list of words for each writer containing words which are mainly used by that particular writer. If another writer uses any of those words more than five times, then those words were not be considered as a unique word for that specific writer.

8) Previous and following words of a unique word:

We collected the previous and following words in a sentence for every unique word used by a writer.

9) Frequency of words related with "আমি", "করা","পারা" :

When we considered the word "আমি" we selected words like, আমি, আমরা, আমার, আমাদের etc. For "করা" we selected করা, করর, কররনা etc. And for "পারা" we selected পারা, পারর, পাররনা, etc.

### *B. Experimental Study:*

1) Calculating Word Frequency:

A writer may frequently use some specific words and by calculating frequency of those words we can rank them as words most frequently used by a writer.

2) Frequency of one word:

We have implemented a program using java programming language to calculate the frequency of every word. Fig. 1 shows the words most frequently used by the four writers.

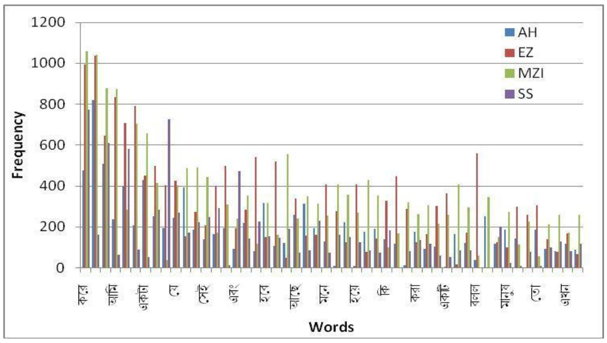

Figure 1. One word frequency of four writers.

As shown in Fig. 1 the words 'বলল' were most frequently used by EZ compared to other writers. So we can say that this can be used as a significant feature in case of EZ.

3) Frequency of two consecutive words:

Writers often use two specific words consecutively. We counted the frequency of two consecutive words using another program implemented in java similar to the program used above. Fig. 2 shows the most frequent two consecutive words used by the four writers.

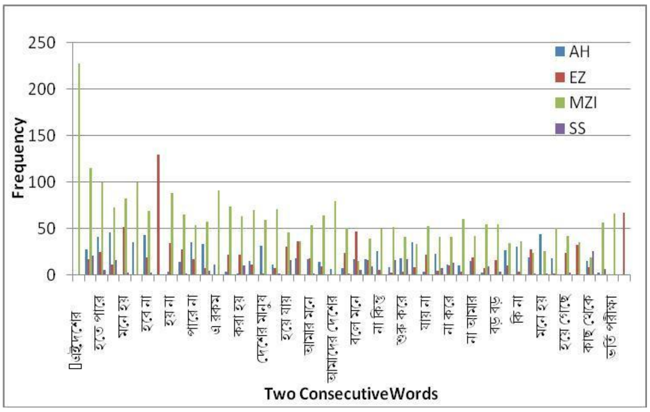

Figure 2. Two consecutive word frequency of four writers.

As shown in Fig. 2 the pair 'এই দেশের' were used in a higher frequency by MZI than the others. The similar differences are shown in case of other writers.

4) Word Length:

Longer words are traditionally associated with more formal style, whereas shorter words are a typical feature of informal spoken language. Shorter words are also a feature of the modern day writes.

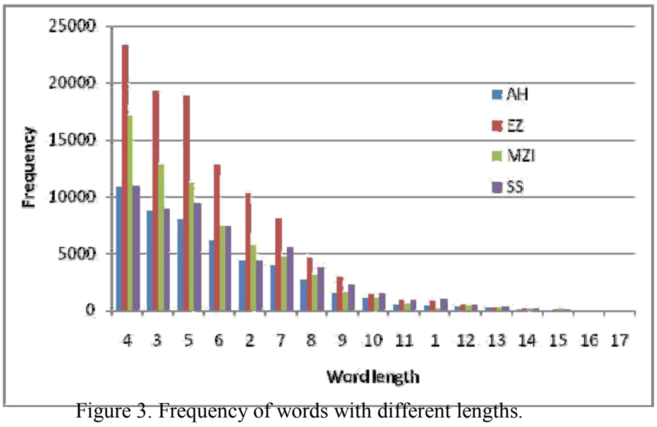

Figure 3. Frequency of words with different lengths.

As shown in Fig. 3 all the writers prefer to use shorter words more frequently than longer words most noticeably the writer EZ.

5) Sentence Length:

While shorter sentences are more indicative of spoken language, longer sentences indicate formal writing. Fig. 4 shows the usage of shorter sentences in a high frequency in case of all four writers.

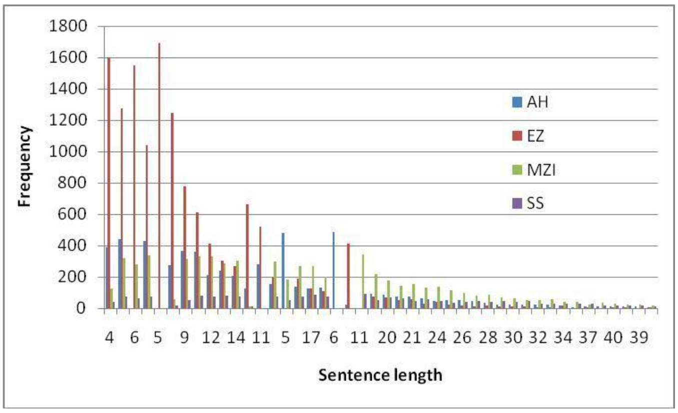

Figure 4. Frequency of sentences with different length.

As shown in Fig. 4 shorter sentences were used most frequently by EZ, while longer sentences were used in a higher frequency in case of MZI.

6) Type-token ratio**:**

The ratio of number of unique words and total number of words present in a document.

$$\text{Type-token ratio} = N/V, \quad (1)$$

Here, N= number of unique words, V= Total words

Type-token ratio helps to find the richness of the vocabulary of a text. The writer that has the larger ratio has a richer vocabulary. Fig. 5 shows a pie-chart of the type-token ratio of all four writers.

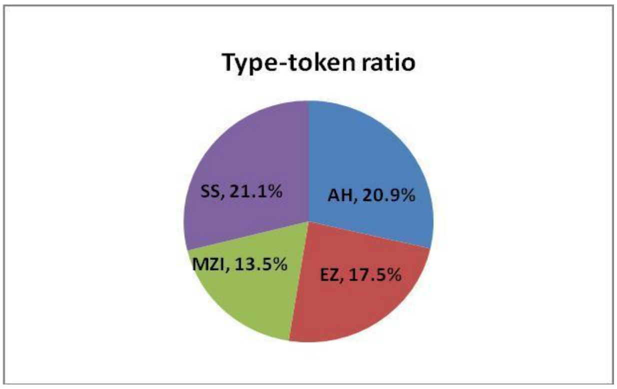

Figure 5. Type-token ratio of four writers.

The pie-chart above shows that SS and AH has higher type-token ratio compared to MZI and EZ. SS has highest type-token ratio, while MZI has the lowest ratio.

7) Distribution of parts-of-speech**:**

As part of identifying the distribution of parts-of-speech, we calculated the number of "conjunction" and "pronoun" in the documents.

Frequency of conjunction indicate continuity in someone's writing. On the other hand frequency of pronoun is indicative of particular pattern or style in writing.

Fig. 6 below shows the comparison of average number of conjunction and pronoun used per document by each writer.

It shows that average number of pronoun used by AH is higher than the others and in case of average number of conjunction is higher for EZ.

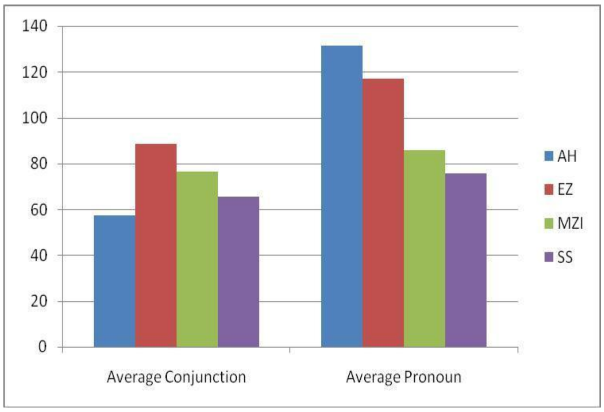

Figure 6. Average number of "Conjunction" & "Pronoun" per document.

8) Position wise word frequency:

Same words can be used in a different position in a sentence by different writers.

9) Writer wise unique words:

These words are especially used by one particular writer. These can be considered as recognizable words for a writer, which indicates the type or category of the text. In our analysis we found some recognizable words of three of the four writers.

In case of EZ, the "Recognizable words" are using different "Names."

Fig. 7 shows the frequency of recognizable words used by EZ comparing to other three writers.

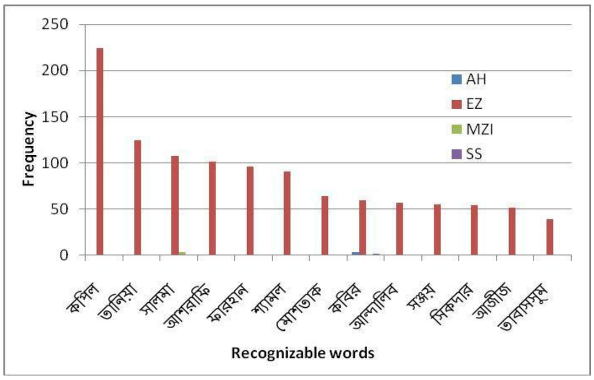

Figure 7. Comparison of "Recognizable words" used by EZ with others.

In case of MZI, the recognizable words are "Educational Words".

Fig. 8 shows the frequency of "Recognizable words" used by MZI comparing to other three writers.

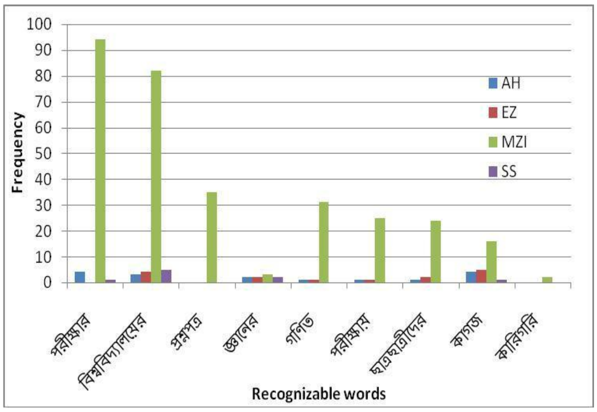

Figure 8. Comparison of "Recognizable words" used by MZI with others.

In case of SS the "Recognizable words" are "Political Words".

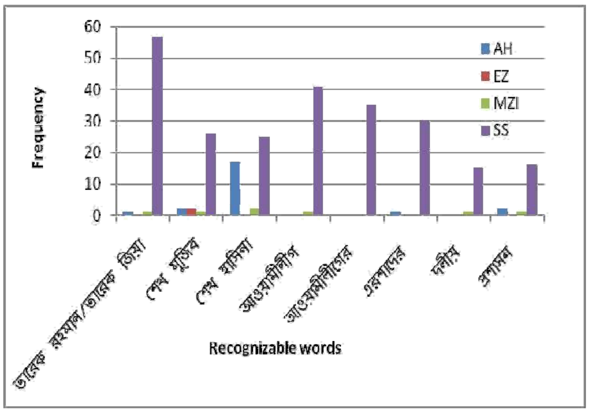

Figure 9. Comparison of "Recognizable words" used by SS with others.

10) Previous and Following words of a unique word:

There might be several words that are used frequently by all four writers. But analyzing the previous and following we can find some difference which could be an important feature.

11) Frequency of words related with "আমি", "করা", "পারা":

If the usage of words related with "আমি" is high in someone's writing, it can indicate that he is more likely to write in first person.

Fig. 10 shows the comparison of using words related with "আমি", "করা", "পারা" among four writers.

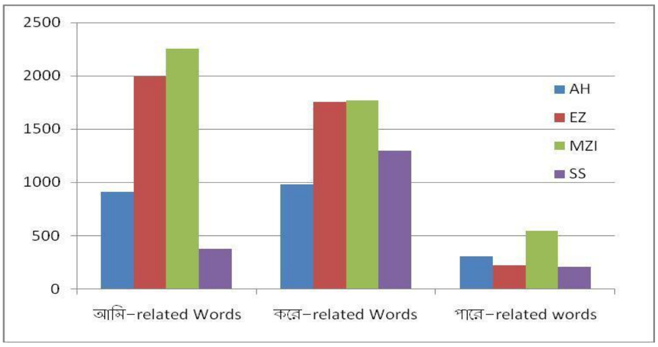

Figure 10. Usage of words related with "আমি", "করা" and "পারা".

We collected training data from these feature analysis. Then we compared the data with some random test data using statistical and machine learning approaches.

*A. Standard Deviation:*

It is a measure that is used to quantify the amount of variation or dispersion of a set of data values.

We counted the average number of conjunction per document for each writer from our training data. We did the same for test data. Then we calculated the standard deviation between a test data and each of the four writers' training data, e.g. we have test_data_1 and training data of four writers- AH, EZ, MZI and SS. If test_data_1 has the lowest standard deviation with training data of EZ, this indicates that test_data_1 might be the writing of EZ. Fig. 11 shows the standard deviation between test data and training data for average number of conjunction.

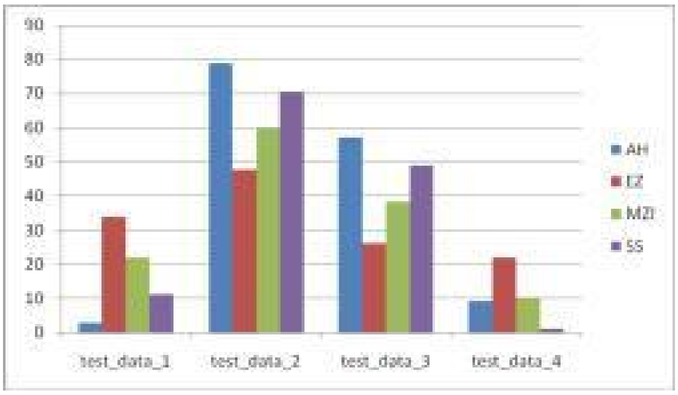

Figure 11. Standard Deviation between "Test data" & "Training data" for average number of "Conjunction".

We applied the same approach for average number of Pronoun.

Fig. 12 shows the standard deviation between test data and training data for average number of pronoun.

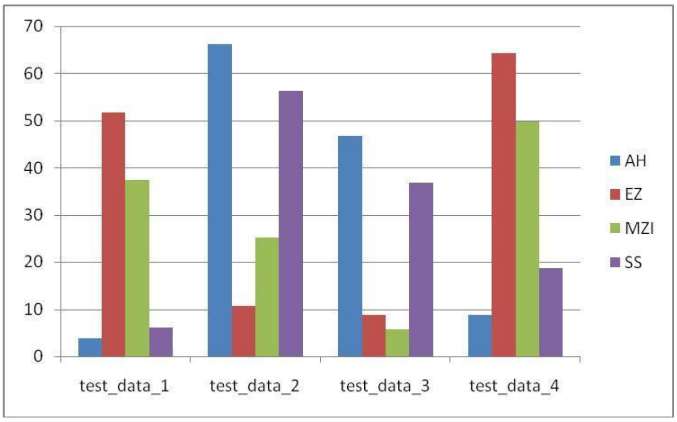

Figure 12. Standard Deviation between “Test data” & “Training data” for average number of “Pronoun”.

*B. Jaccard Similarity:*

Jaccard Similarity is used to find similarity between two data sets. We applied this method for the feature writer wise unique word. We created a vector space model for the training data of unique words of each of the four writers and also for test data. Then we calculated the Jaccard similarity between test and training data, e.g. if test_data_1 has highest similarity value with AH, it indicates test_data_1 is most likely to be written by AH. Fig. 13 shows the Jaccard Similarity between test data and training data for writer wise unique words.

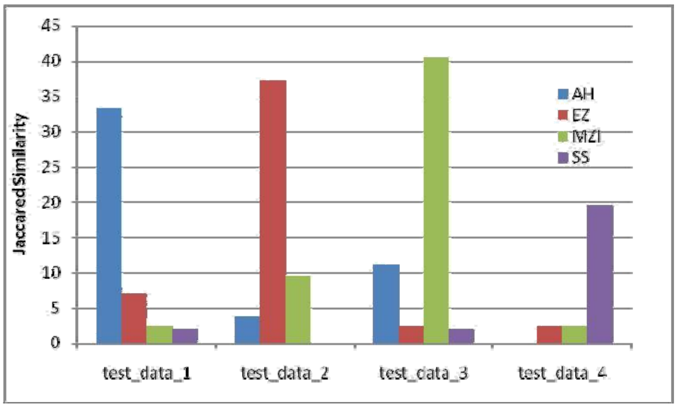

Figure 13. Jaccard Similarity between test data and training data for writer wise unique words.

## IV. CONCLUSION

Stylogenetics provides an interesting venue for motivating and demonstrating many standard multivariate statistical techniques. This can be very useful for exploring and analyzing literary data. This paper attempts to recognize different authors based on their style of writing. As our first approach in this paper we analyzed twelve features according to the style of Bangla literature. Both standard deviation and Jaccard similarity can be one of the approaches to find some significant Stylometric features of a writer.

There have been no previous works on Stylometry for Bangla literature. Machine learning methods like Cosine similarity or Clustering can also be used to find some Stylometric features of a writer.

Stylogenetics is quite a new topic in the field of science and literature, especially Bangla literature. We hope our work will inspire others to work with Bangla literature in future.

## REFERENCES

[1] Kim Luyckx,Walter Daelemans and Edward Vanhoutte, “Stylogenetics: Clustering-based stylistic analysis of literary corpora” , University of Antwerp Faculty of Arts Universiteitsplein 1, B-2610 Antwerp, Belgium.

[2] Roger Peng and Nicolas Hengartner, “Quantitative Analysis of Literary Styles”, The American Statistician 56.3 (2002), 175-185.

[3] Michael Brennan and Rachel Greenstad,“ Practical Attacks Against Authorship Recognition Techniques”, Dept. of Computer Science, Drexel University, 3175 JFK Blvd Room 140 Philadelphia, PA 19104.

[4] David. I. Holmes, “A Stylometric Analysis of Mormon Scripture and Related Texts”, Journal of the Royal Statistical Society. Series A (Statistics in Society), Vol. 155, No. 1,(1992), 91-120.

[5] Holmes, David I, “The Analysis of Literary Style: A Review,” Journal of the Royal Statistical Society, Series A (General), (1985), 328-341.

[6] Ramyaa, Congzhou He, Khaled Rasheed, “Using Machine Learning Techniques for Stylometry”, Artificial Intelligence Center, Proceedings of International Conference on Machine Learning, 2004.

[7] Jan Rygl, “Automatic Adaptation of Author’s Stylometric Features to Document Types”, Text, Speech and Dialogue, Springer International Publishing, 2014, 53-61.

[8] Bangla Blog, http://blog.priyo.com/blogs, Last accesses: 11th May, 2015.

[9] Bangla Blog, http://www.somewhereinblog.net/blog, Last accesses: 11th May, 2015

[10] Bangla Blog, https://shadashidhekothaarchive.wordpress.com, Last accesses: 11th May, 2015.