Shahjalal University of Science and Technology, Sylhet.
Department of Computer Science and Engineering

28th March 2015

Shahjalal University of Science and Technology
Department of Computer Science and Engineering

Stylogenetics

Student: (Name: Rishmita Tasmim. Reg. no: 2010331015 4/1,CSE
Name: Prapti Das. Reg. no: 2010331022 4/1, CSE)
Adviser: (Sabir Ismail, Lecturer, CSE)

30th March, 2015

Stylogenetics

A Thesis submitted to the Department of Computer Science and Engineering,
Shahjalal University of Science and Technology, in partial fulfillment of the requirements
for the degree of Bachelor of Science in Computer Science and Engineering.

Student: (Name: Rishmita Tasmim. Reg. no: 2010331015, 4/1,CSE
Name: Prapti Das. Reg. no: 2010331022, 4/1, CSE)
Advisor: (Sabir Ismail, Lecturer, CSE)

30th March, 2015

Recommendation Letter from Thesis Supervisor

These Students Rishmita Tasmim & Prapti Das whose thesis entitled “Stylogenetics” is under my supervision and agree to submit for examination.

Advisor :
Date :

Qualification Form of Bachelor Degree

Student Name : Rishmita Tasmim.
Prapti Das.
Thesis Title : Stylogenetics.

This is to certify that the thesis submitted by the student named above in 28th March, 2015. It is qualified and approved by the Thesis Examination Committee.

Head of the Dept. Chairman, Thesis Committee Supervisor

ABSTRACT

Every writer has a different style of writing of their own. We selected a few number of features in the writing of the writers. By analyzing various kinds of features we can identify and specify a characteristic in the writing of a writer which is known as Stylogenetics. In this paper we gathered Bengali blogs written by four different Bangladeshi writers.
Sometimes people claim someone else's writing as their own. Stylogentics will help us to identify the real writer. We did this research to verify a writer.
Our methodology is to analyze some features in their writings. For example, percentage of unique words, frequency of one word, frequency of two consecutive words, word length, sentence length, frequency of some parts of speech etc. By this analysis we gathered some statistical data for each writer. Then we compared the data with one another and represented it graphically. By analyzing those graphs we tried to find and choose the features that serve the best in finding the variation among writers and identify a specific writer.

Keywords :

Stylogenetics, unique word, word length, sentence length, parts of speech.

ACKNOWLEDGMENTS

Foremost, we would like to express our sincere gratitude to our advisor Sabir Ismail for the continuous support of our Bachelor Thesis study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped us in all the time of research and writing of this thesis. We could not have imagined having a better advisor and mentor for our Bachelor Thesis study.

Besides our advisor, we would like to thank the rest of our thesis committee: Prof. Dr. Mohammad Shahidur Rahman and Abu Naser, for their encouragement.

Finally we would like to thank our parents because all our academic achievements are the outcome of their sacrifice.

TABLE OF CONTENTS

Page

ABSTRACT…….4

ACKNOLEDGEDMENT……...5

TABLE OF CONTENTS……6

LIST OF TABLES……..8

LIST OF FIGURES……9

1 INTRODUCTION……11

1.1 Background…….12

1.2 Motivation……...13

2 BACKGROUND STUDY……13

2.1 Study of Different Thesis Papers……..13

3 METHODOLOGY……..16

3.1 Choosing Features……16

3.2 Feature Analysis…….16

4 EXPERIMENTAL STUDY……17

4.1 Raw Data Collection……..17

4.2 Calculating Word Frequency…….17

4.2.1 Frequency of one word……..17

4.2.2 Frequency of two Consecutive Words……...22

4.3 Calculating Word Length……...26

4.4 Calculating Sentence Length…….28

4.5 Type-Token Ratio……..30

4.6 Distribution of Parts-of-Speech…….31

5 RESULT ANALYSIS AND DISCUSSIONS……..35

5.1 Comparing Word Frequency……..35

5.2 Comparing word length……..36

5.3 Comparing sentence length……37

5.4 Comparing Type-Token Ratio……37

5.5 Comparing distribution of parts-of-speech…….38

6 FUTURE WORK……..39

6.1 Proposal 1……39

6.2 Proposal 2……39

6.3 Proposal 3……40

7 CONCLUSION……40

REFERENCES……41

APPENDIX……..41

List of Tables

		Page
Table 1	Top 20 most frequent words used by “Anisul Hoque”.	18
Table 2	Top 20 most frequent words used by “Dr. Muhammed Zafar Iqbal”.	19
Table 3	Top 20 most frequent words used by “Imon Zubair”.	20
Table 4	Top 20 most frequent words used by “Syed Shah Salim”	21
Table 5	Top 20 most frequent two consecutive words used by “Anisul Hoque”	22
Table 6	Top 20 most frequent two consecutive words used by “Dr. Muhammed Zafar Iqbal”.	23
Table 7	Top 20 most frequent two consecutive words used by “Imon Zubair”.	24
Table 8	Top 20 most frequent two consecutive words used by “Syed Shah Salim”	25

List of Figures

Figure 1	Graphical representation of Top 20 most frequent words used by "Anisul Hoque".	18
Figure 2	Graphical representation of Top 20 most frequent words used by "Dr. Muhammed Zafar Iqbal".	19
Figure 3	Graphical representation of Top 20 most frequent words used by "Imon Zubair".	20
Figure 4	Graphical representation of Top 20 most frequent words used by "Syed Shah Salim"	21
Figure 5	Graphical representation of Top 20 most frequent two consecutive words used by "Anisul Hoque"	23
Figure 6	Graphical representation of Top 20 most frequent two consecutive words used by "Dr. Muhammed Zafar Iqbal".	24
Figure 7	Graphical representation of Top 20 most frequent two consecutive words used by "Imon Zubair".	25
Figure 8	Graphical representation of Top 20 most frequent two consecutive words used by "Syed Shah Salim"	26
Figure 9	Top 10 most frequent words of different word length used by "Anisul Hoque".	26
Figure 10	Top 10 most frequent words of different word length used by "Dr. Muhammed Zafar Iqbal"	27
Figure 11	Top 10 most frequent words of different word length used by "Imon Zubair".	27
Figure 12	Top 10 most frequent words of different word length used by "Syed Shah Salim".	28
Figure 13	Top 10 most frequent sentences of different length used by "Anisul Hoque".	28
Figure 14	Top 10 most frequent sentences of different length used by "Dr. Muhammed Zafar Iqbal".	29

Figure 15	Top 10 most frequent sentences of different length used by “Imon Zubair”.	29
Figure 16	Top 10 most frequent sentences of different length used by “Syed Shah Salim”.	30
Figure 17	Comparison of Type-Token Ratio among four writers.	30
Figure 18	Frequency of Top 10 conjunction (অব্যয়) used by “Anisul Hoque”.	31
Figure 19	Frequency of Top 10 conjunction (অব্যয়) used by “Dr. Muhammed Zafar Iqbal”	31
Figure 20	Frequency of Top 10 conjunction **(**অব্যয়) used by “Imon Zubair”.	32
Figure 21	Frequency of Top 10 conjunction (অব্যয়)used by “Syed Shah Salim”.	32
Figure 22	Frequency of Top 10 pronoun (সর্বনাম) used by “Anisul Hoque”.	33
Figure 23	Frequency of Top 10 pronoun (সর্বনাম) used by “Dr. Muhammed Zafar Iqbal”.	33
Figure 24	Frequency of Top 10 pronoun (সর্বনাম) used by “Imon Zubair”.	34
Figure 25	Frequency of Top 10 pronoun (সর্বনাম) used by “Syed Shah Salim”.	34
Figure 26	Comparison of one word frequency used by four writers.	35
Figure 27	Comparison of consecutive two words frequency used by four writers.	36
Figure 28	Comparison of word length frequency used by four writers.	36
Figure 29	Comparison of sentence length frequency used by four writers.	37
Figure 30	Comparison of frequency of conjunction(অব্যয়) used by four writers	38
Figure 31	Comparison of frequency of pronoun (সর্বনাম) used by four writers	38

Chapter 1
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. It is a way of analyzing written texts to learn about the writer. Some things are often included unconsciously by the writer, like things that indicate gender, age, geographic location of the writer, personality characteristics, etc.
Stylogenetics helps to detect who is the actual author of a specific writing. Sometimes there are some writers who are frauds. They steal someone else's book or writing and claim it to be their own. Stylogenetics will analyze the writing and find a specific pattern or characteristic of that writing. Through which we can detect if that person really wrote that book.

Stylogenetics is mainly used to identify a writer. It helps to find out a special feature or pattern in a writers writing which is often included in the writing unconsciously. For example, a writer may use two specific consecutive words most frequently, a writer may start and end a sentence with specific parts of speech, a writer 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.

Our future plan is to analyze the writing of four or more Bangladeshi writers using Stylogenetics and find specific features that will help us to identify a writer precisely.
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 preposition etc.

1.1 Background

There have many work in the field of Stylometry which very similar to Stylogenetics. Stylometry is the application of the study of linguistic style, usually to written language. Stylometry is often used to attribute authorship to anonymous or disputed documents. It has legal as well as academic and literary applications, ranging from the question of the authorship of Shakespeare's works to forensic linguistics.

There has been much work covering different aspects of this field. For a comprehensive review we direct the reader to Holmes (1985). Many early attempts to quantify style relied on concordances, or inventories of the frequency of every word in a text. In 1901 T. C. Mendenhall reduced the concordances of Shakespeare and Bacon to distributions of word lengths and plotted these distributions as graphs. His so called "characteristic curves" serve as an early example of the use of graphics in distinguishing authorship. Mendenhall examined the differences in the shapes of the curves (such as the location of the mode) and concluded that Bacon probably did not write any of Shakespeare's works. C. B. Williams reproduced some of Mendenhall's curves and noted that he was mistaken in some of his conclusions and that there was little evidence for or against the theory that some works written by Shakespeare could have been written by Bacon (Williams, 1975). Brinegar (1963) also used word length distributions to determine if Mark Twain had written the Quintus Curtius Snodgrass (QCS) letters. He used X^2 tests and two-sample t-tests on the counts of 2, 3, and 4 letter words to check the agreement of the QCS letters with Twain's known writings. Thisted and Efron (1987) used the idea of vocabulary richness to determine the possibility of Shakespearean authorship of a newly discovered poem. They based their analysis of the poem on the rate of "discovery" of new words given the number of distinct words previously observed in the Shakespearean canon. Holmes (1992), in an example of the use of a standard multivariate analysis technique, used hierarchical cluster analysis to detect changes in authorship in Mormon scripture. He also used various measures of vocabulary richness to conduct his analysis.

There is no general agreement on the unit of analysis that should be used in authorship studies. In the previously mentioned examples, word length and vocabulary richness were the units used. Williams (1940) analyzed the sentence lengths of works written by Chesterton, Wells, and Shaw. He noticed that the log of the number of words per sentence appeared to follow a normal distribution. Morton (1965) also used sentence length in his analysis of ancient Greek texts. After initially using criteria such as word length and sentence length, Mosteller and Wallace (1963) focused on using function word counts to discriminate between the works of Hamilton and Madison in their seminal analysis of the Federalist Papers (see also Mosteller and Wallace, 1964). They found that Hamilton and Madison were "practically twins" with respect to the average sentence lengths in their writings. Therefore, they decided to use function words, which are words with very little contextual meaning. These words include conjunctions, prepositions, and pronouns. The logic behind using function words is that writers do not necessarily think about the way they use these words. Rather these words flow unconsciously from the mind to the paper. Therefore, the usage of function words should be invariant under changes of topic. Mosteller and Wallace (1963) successfully used the frequency distribution of a few function words to assign authorship to the unsigned Federalist Papers. Sarndal (1967) also used word counts in an interesting attempt to quantify type I and type II errors in authorship discrimination. He facilitated the analysis by assuming independent Poisson distributions for the word counts.

Mosteller and Wallace (1963) noted that in their study, the Poisson distribution did not fit the word count distributions particularly well, and that the negative binomial distribution provided a better fit because of its heavier tail.

1.2 Motivation

Stylogenetics is a new topic in the field of Bengali literature. There have been a few works on English literature. People have worked on famous English writers like Shakespeare, Jane Austen, Charles Dickens and so on. But there has not been any analysis on Bengali writers. There are many talented and skillful writers in our country. We often find interesting style and pattern in their writings. By analyzing their writing we can distinguish among the characteristics and personalities of the writers.

Bengali is our mother tongue and Bengali literature is quite famous all around the world. It is also one of the most enriched literatures. So we believe that in the field of Stylogenectics there should be proper research work on Bengali literature. That's why we decided to work with Bengali literature.

CHAPTER 2

BACKGROUND STUDY

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.

2.1 Study of Different Thesis Papers :

We studied a few papers related to Stylogenetics and Stylometry which will be discussed in this section.

In "**Stylogenetics: Clustering-based stylistic analysis of literary corpora**" paper by **Kim Luyckx, Walter Daelemans, Edward Vanhoutte** they worked on a Methodology which is borrowed from topic detection research - are

(i) Using more complex features than the simple lexical features suggested by traditional approaches.
(ii) Using authors or groups of authors as a prediction class.
(iii) Using clustering methods to indicate the differences and similarities between authors (i.e. Stylogenetics).

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. Combinations of syntactic, token-based and lexical features constitute a profile that characterizes the style of an author. The Stylogenetics methodology opens up new perspectives for literary analysis, enabling and necessitating close cooperation between literary scholars and computational linguists.

In their study four types of features that have been applied as style markers can be distinguished: token-level features (e.g. word length, readability), syntactic features (e.g. part-of-speech tags, chunks), features based on vocabulary richness (e.g. type-token ratio) and common word frequencies (e.g. of function words) (Stamatatos et al., 2001). While most Stylometric studies are based on token-level features, word forms and their frequencies of occurrence, syntactic features have been proposed as more reliable style markers since they are not under the conscious control of the author (Baayen et al., 1996; Diederich et al., 2000; Khmelev and Tweedie, 2001; Kukushkina et al., 2001; Stamatatos et al., 1999). Thanks to improvements in shallow text analysis, they extracted syntactic features to test their relevance in Stylogenetic research.

In **"Quantitative Analysis of Literary Styles"** paper by **Roger Peng Nicolas Hengartner** they presented an overview and brief history of the analysis of literary styles. In addition they used canonical discriminant analysis and principal component analysis to identify structure in the data and distinguish authorship.

In this study we also take groups of function words as the units of analysis. When analyzing word frequencies, one often makes the following assumptions:

(1) The style of an author remains the same throughout his/her life;

(2) Successive occurrences of function words are independent.

Neither assumption tends to hold in practice. The purpose of using function words in the first place is to deal with (1). Because function words have little contextual meaning, they thought of them abstractly as the "noise" of language.

They have shown in this paper that traditional multivariate techniques can be very useful for exploring and analyzing literary data. The data are inherently high dimensional and cannot be readily visualized or understood. Initially, principal component analysis was used to examine each individual author's function word counts. PCA proved useful for identifying unusual blocks and possible violations of the independence and uniformity assumptions.

Canonical discriminant analysis was used to provide dimension reduction and graphical displays of the differences between authors (canonical vector plots). Also, CDA was useful for identifying key function words which were most effective at discriminating between authors. The key words were identified by examining plots of the loadings for each function word.

Function words have proven to be effective instruments for accessing literary data. Function words were chosen as the unit analysis because they are highly variable between authors, abundant, and easy to count and identify.

In **"A Stylometric Analysis of Mormon Scripture and Related Texts"** paper by, **D. I. Holmes** they proposed a multivariate approach to measuring the richness of vocabulary of a literary text as a tool for problems of attribution of authorship. To obtain suitable quantitative indicators of richness of vocabulary, they believed that they should not only take into account the text length itself (N words) but also the number of different words in the text (V) and the structure of the vocabulary frequency distribution.

This paper puts forward a case for a multivariate approach to measuring richness of vocabulary, employing variables each of which reflects a different aspect of an author's vocabulary distribution. It then applies these techniques to a body of writings which are directly relevant to establishing the authenticity of the **Book of Mormon.**

In **"Practical Attacks Against Authorship Recognition Techniques"** by **Michael Brennan and Rachel Greenstadt** They seek to discover how robust current methods of Stylometry are in dealing with adversarial attacks. Until now the field has focused on creating new methods that attempt to classify existing unknown works sets of authors, with little attention being given to the question of what happens when an adversary tries to intentionally circumvent the classification system that has been established.

This study looks at three specific approaches and their resilience against two types of adversarial attacks. The first, which will be referred to as an obfuscation attack, is when an author attempts to write a document in such a way that their personal writing style will not be recognized. The second, which will be referred to as an imitation attack, is when an author attempts to write a document such that their writing style will be recognized as that of another specific author. The three methods of Stylometry investigated were chosen for their variety in both metrics and methodology. The study found that none of the methods performed better than chance in identifying the correct author in either of these attacks.

.

The three specific approaches are :

1. **Statistical Method using the Signature Stylometric System :**
   This approach analyzes the text, and then uses a basic statistical method for comparison. The features used for the analysis are word lengths, letter usage, and punctuation. On average this method correctly identified the original author 95% of the time.
2. **Neural Network Approach :**
   This approach is adapted from the approach outlined in Neural Network Applications in Stylometry (Tweedie, Singh, and Holmes 1996). This approach is one of the first and most widely used approaches of combining Stylometry with the field of Artificial Intelligence. The effectiveness of Method 2 was confirmed through repeated random sub-sampling validation.
3. **Synonym-Based Classifier :**
   The general approach of this method is to examine how each author chooses synonyms. The theory behind the method is that when a word has a large number of synonyms to choose from, the choice the author makes is significant in understanding his or her writing style.

CHAPTER 3

METHODOLOGY

For this paper we collected around 247 blogs written by four different Bangladeshi writers. Then we specified some special features to analysis the writings. The features we worked on are described in this section.

3.1 Choosing Features :

- **Word frequency:**

  Word frequency means how many times a word was used in a writers writing. This is used as a feature to find the most frequent words used by a writer.
- **Word length :**

  The distribution of words of different length has been used as a feature. We calculated the frequency of words of different length.
- **Sentence length :**

  The number of words present in a sentence is used as a feature. We defined it as sentence length.
- **Type-token ratio :**

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

3.2 Feature Analysis :

We analyzed the raw data according to these features and created some statistical measures for each feature. This statistical information is used in this research. We compared these data of each writer with one another.

We are planning on collecting more blogs from more writers. And analyze their writings and gather more statistical information. After that we plan on applying Dimension Reduction on this data and try to find variations among the writing style of the writers. This will help us to identify a writer by their writing.

CHAPTER 4

EXPERIMENTAL STUDY

4.1 Raw Data Collection :

In this paper we did our analysis on blogs written by four different Bangladeshi writers. The raw data for this study was obtained from different Internet Websites. Multiple works for each author were downloaded in text format and processed. The four writers we worked on are :

a) Anisul Hoque (69 Blogs)
b) Dr. Muhammed Zafar Iqbal (79 Blogs)
c) Imon Zubair (50 Blogs)
d) Syed Shah Salim (49 Blogs)

The links of the websites from which the blogs are collected are given in Appendix A.

4.2 Calculating Word Frequency :

A writer may use some specific words most frequently. By calculating word frequency we can rank those frequent words. Thus we can find the words that a writer uses the most.

4.2.1 Frequency of one word :

We used a Java code and run on all the blogs written by the four writers. We calculated the frequency of every word.

We chose top 20 most frequent words used by each writer and represented them in graphs.

Table 1: Top 20 most frequent words used by “Anisul Hoque”.

Words	Frequency
না	1139
এই	700
করে	643
একটা	584
আর	576
আমাদের	545
আমরা	443
হবে	429
হয়	412
কোনো	376
থেকে	375
যে	368
আছে	337
আমি	318
কথা	306
তো	296
করতে	294
তিনি	290
কিন্তু	286
আমার	286

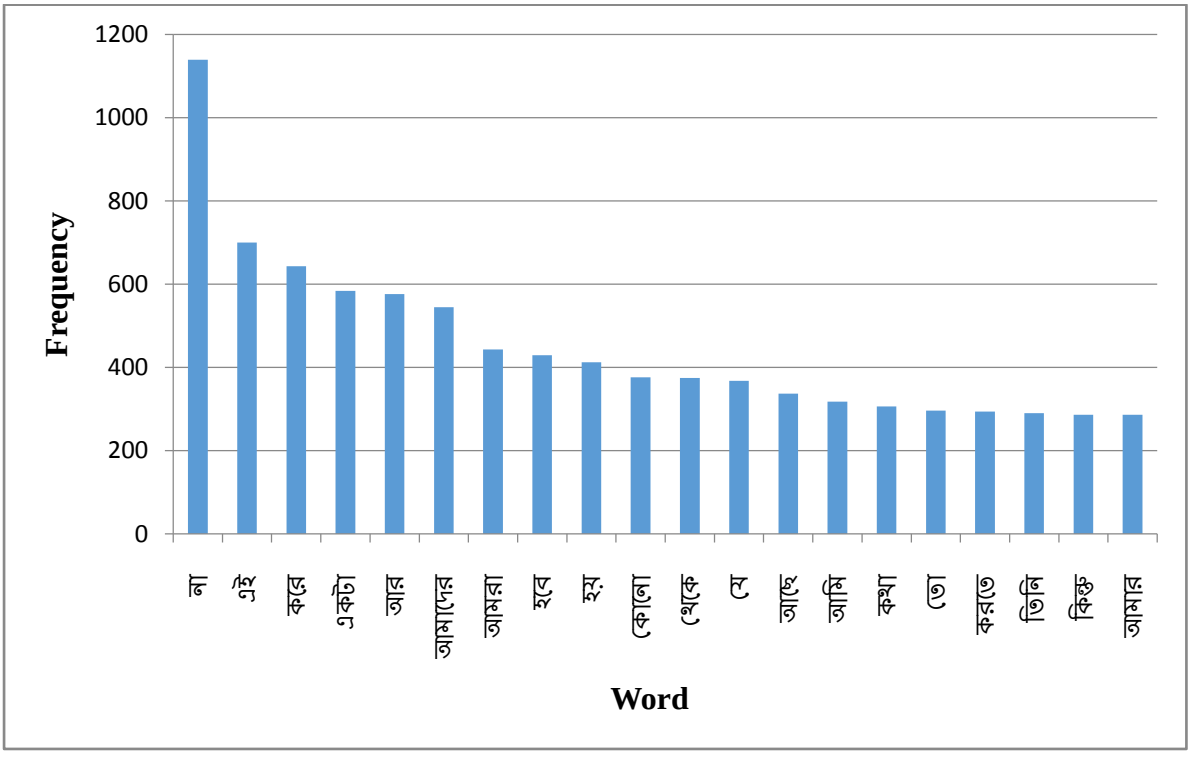

Figure 1: Graphical representation of Top 20 most frequent words used by “Anisul Hoque”.

Table 2: Top 20 most frequent words used by “Dr. Muhammed Zafar Iqbal”.

Word	Frequency
করে	1777
না	1687
এই	1322
আমি	1289
একটা	1076
আমার	1009
আমাদের	887
তাদের	837
সেই	781
দেশের	739
তার	712
যে	665
থেকে	658
জন্য	654
আমরা	643
তারা	614
হয়	607
করতে	604
নিয়ে	601
কিন্তু	582

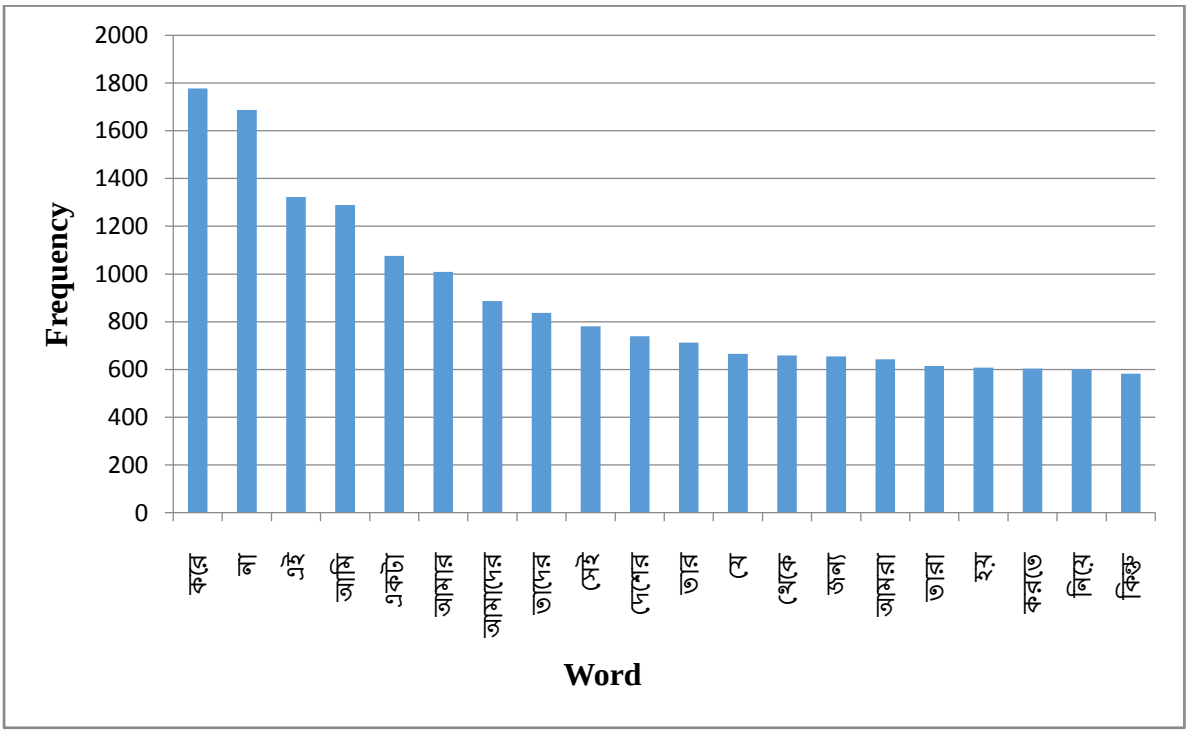

Figure 2: Graphical representation of Top 20 most frequent words used by “Dr. Muhammed Zafar Iqbal”.

Table 3: Top 20 most frequent words used by “Imon Zubair”.

Word	Frequency
না	997
করে	988
আমি	825
আমার	787
আর	707
এই	629
বলল	555
বলে	516
থেকে	498
কী	490
একটা	447
ছিল	443
যে	409
এর	408
বললেন	406
হয়ে	406
ও	398
মনে	396
এক	395
হল	393

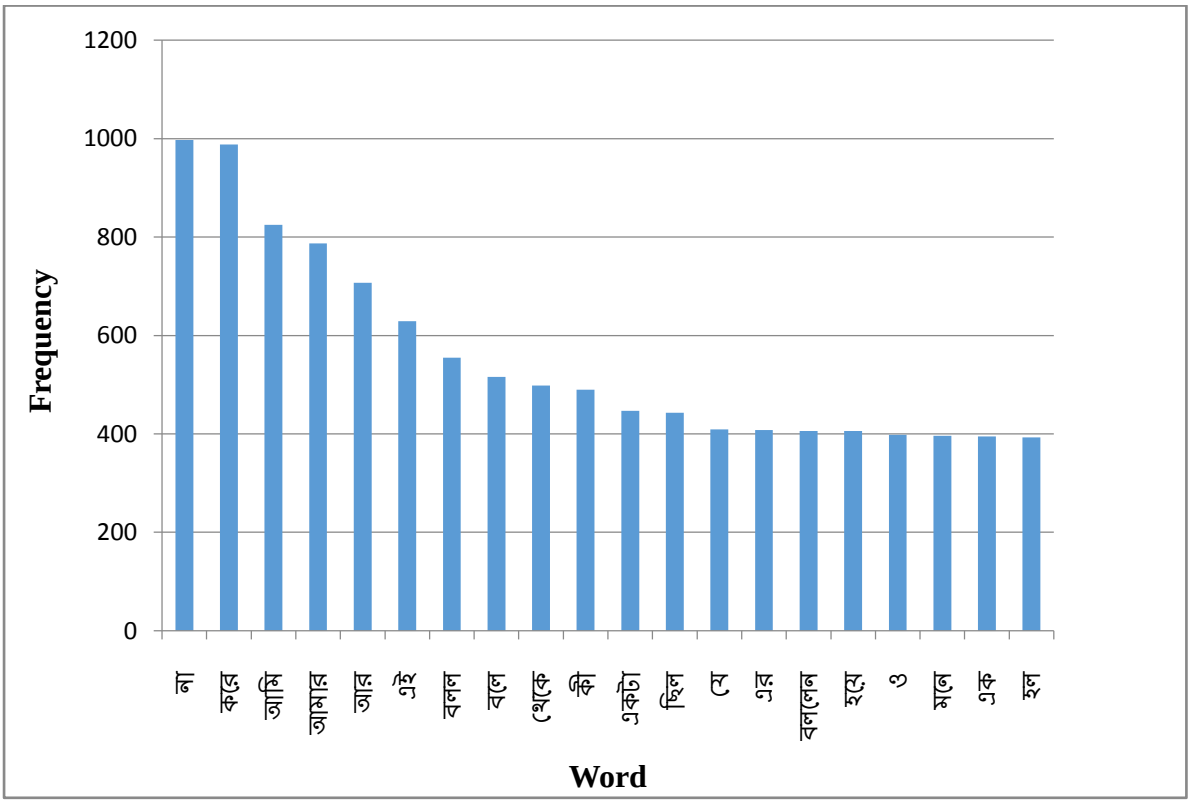

Figure 3: Graphical representation of Top 20 most frequent words used by “Imon Zubair”.

Table 4: Top 20 most frequent words used by “Syed Shah Salim”

Words	Frequency
করে	770
ও	724
এই	600
আর	579
এবং	472
সাথে	340
কোন	337
এক	286
থেকে	278
যে	266
সেই	248
জন্য	231
তার	221
এর	213
সব	201
তাদের	190
বা	183
কি	181
আমাদের	169
না	162

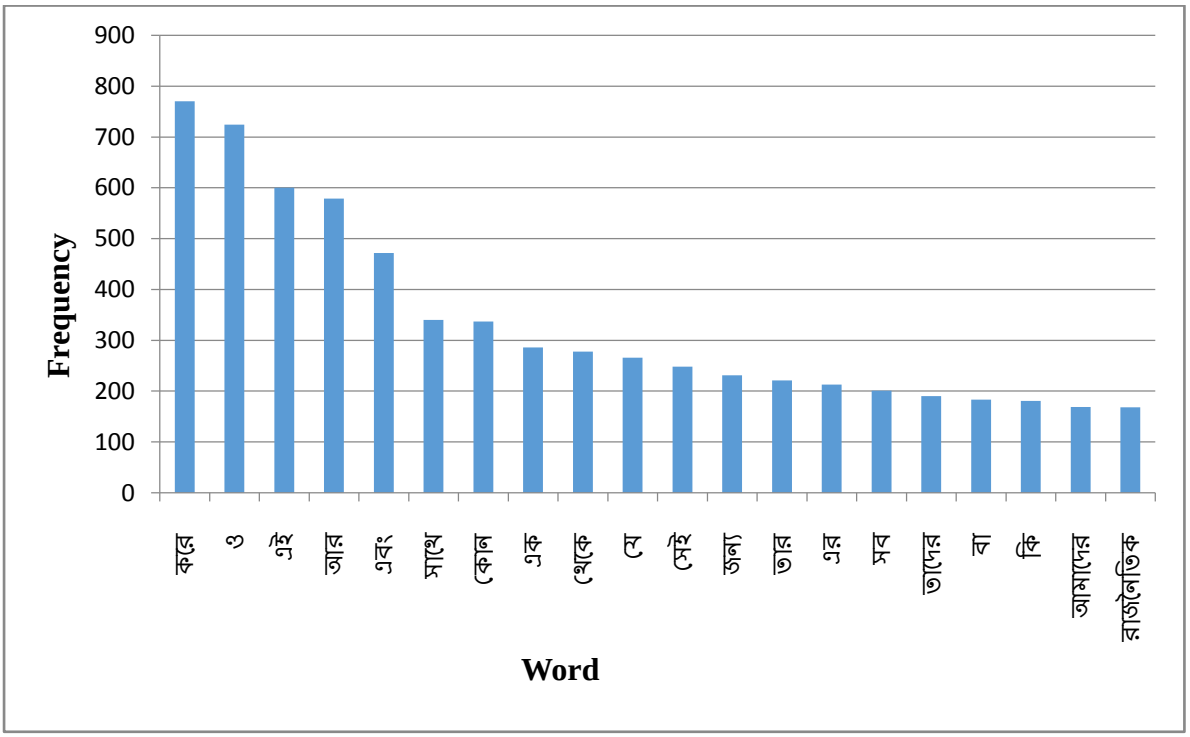

Figure 4: Graphical representation of Top 20 most frequent words used by “Syed Shah Salim”

4.2.2 Frequency of two Consecutive Words :

Sometimes writers often use two specific words consecutively. We calculated the frequency of two consecutive words written by each writer. We used a Java code and run it on all the blogs.

We chose top 20 most frequent two consecutive words used by each writer and represented them in graphs.

Table 5: Top 20 most frequent two consecutive words used by "Anisul Hoque"

Words	Frequency
হয় না	61
আওয়ামী লীগ	49
করতে হবে	46
মনে হয়	44
হবে না	43
হতে পারে	41
পারে না	36
এই দেশে	36
না এই	36
করতে পারে	34
এই দেশের	33
দেশের মানুষ	32
আওয়ামী লীগের	31
কি না	31
এই রকম	30
যায় না	28
করার জন্য	28
পদ্মা সেতু	27
যাবে না	27
না কিন্তু	26

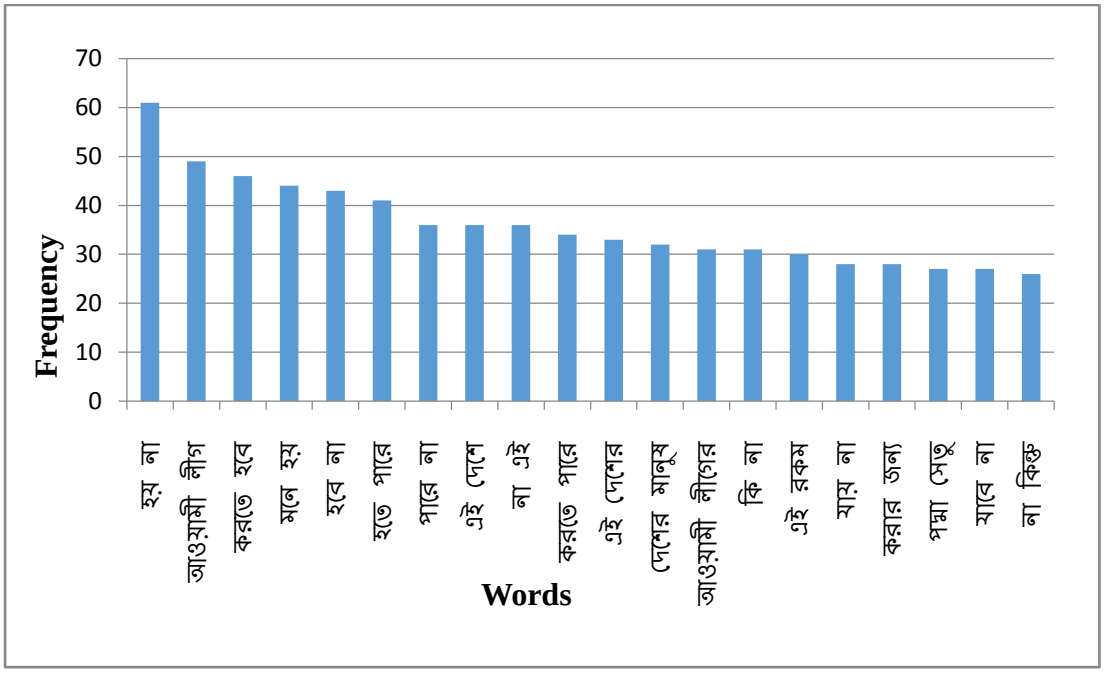

Figure 5: Graphical representation of Top 20 most frequent two consecutive words used by "Anisul Hoque"

Table 6: Top 20 most frequent two consecutive words used by "Dr. Muhammed Zafar Iqbal".

Words	Frequency
এই দেশের	228
করার জন্য	115
এই দেশে	100
হতে পারে	100
এ রকম	91
হয় না	88
মনে হয়	83
আমাদের দেশের	80
আমার কাছে	74
করতে হবে	73
কথা বলতে	71
আমি জানি	70
হবে না	69
ভর্তি পরীক্ষা	66
না আমি	65
জানি না	64
করা হয়	63
পারবে না	61
দেশের মানুষ	60
করতে পারে	58

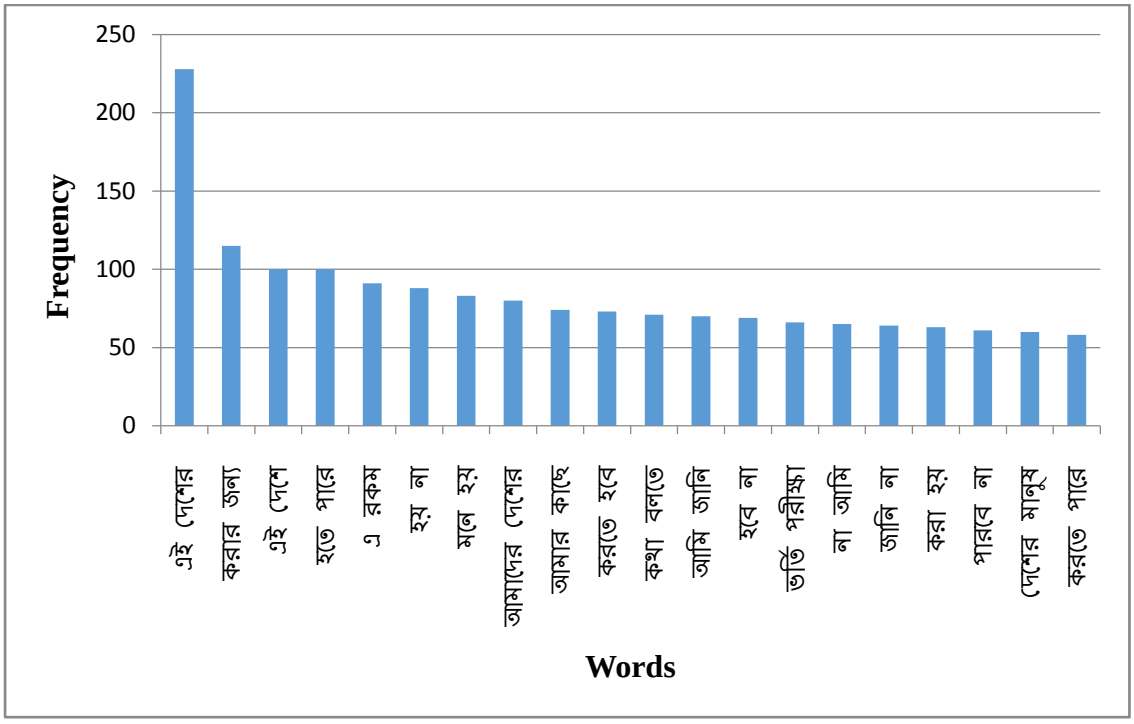

Figure 6: Graphical representation of Top 20 most frequent two consecutive words used by “Dr. Muhammed Zafar Iqbal”.

Table 7: Top 20 most frequent two consecutive words used by “Imon Zubair”.

Word	Frequency
মনে হল	130
অধ্যাপক মোশাররফ	67
চুপ করে	58
সাংবাদিক মোশতাক	58
প্রফেসর আশরাফি	55
মনে হয়	52
মোশতাক আন্দালিব	52
একটু পর	50
বলে মনে	47
জিজ্ঞেস করে	46
টেবিলের ওপর	43
প্রোফেসর আশরাফি	42
তাজউদ্দীন আহমদ	40
কালো রঙের	37
ছিল না	37
কী যেন	36
হয় না	35
এই মুহূর্তে	35
গায়ের রং	35
তানিয়া তাবাসসুম	34

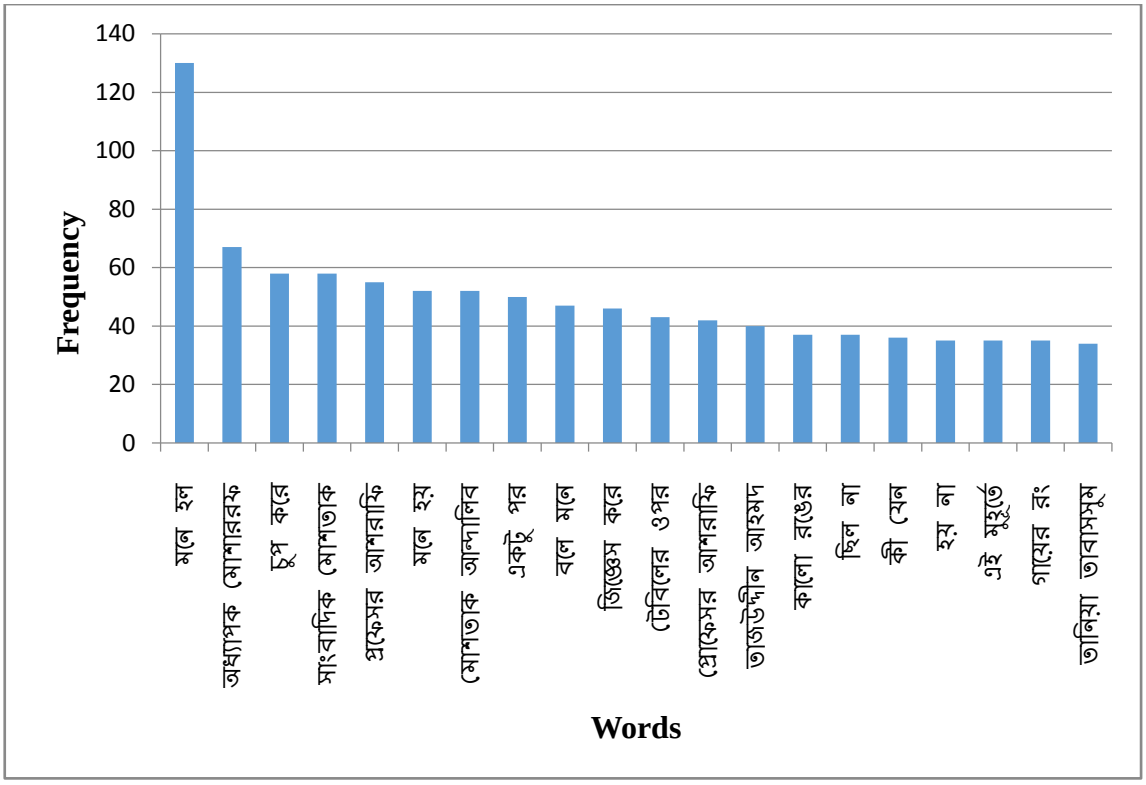

Figure 7: Graphical representation of Top 20 most frequent two consecutive words used by "Imon Zubair".

Table 8: Top 20 most frequent two consecutive words used by "Syed Shah Salim"

Words	Frequency
সাল্লাল্লাহু আলাইহি	59
তারেক রহমান	54
সেই সব	43
এই সব	40
একই সাথে	32
তারেক রহমানের	28
আলাইহি ওয়াসাল্লাম	28
আলাইহি ওয়াসাল্লামের	27
কাছ থেকে	26
শেখ মুজিব	26
শেখ হাসিনা	25
এই রকম	23
রাসূলুল্লাহ সাল্লাল্লাহু	22
শেখ হাসিনার	22
করে চলেছেন	22
যেমন করে	22
করার জন্য	21
একের পর	21
পর এক	20
কি করে	19

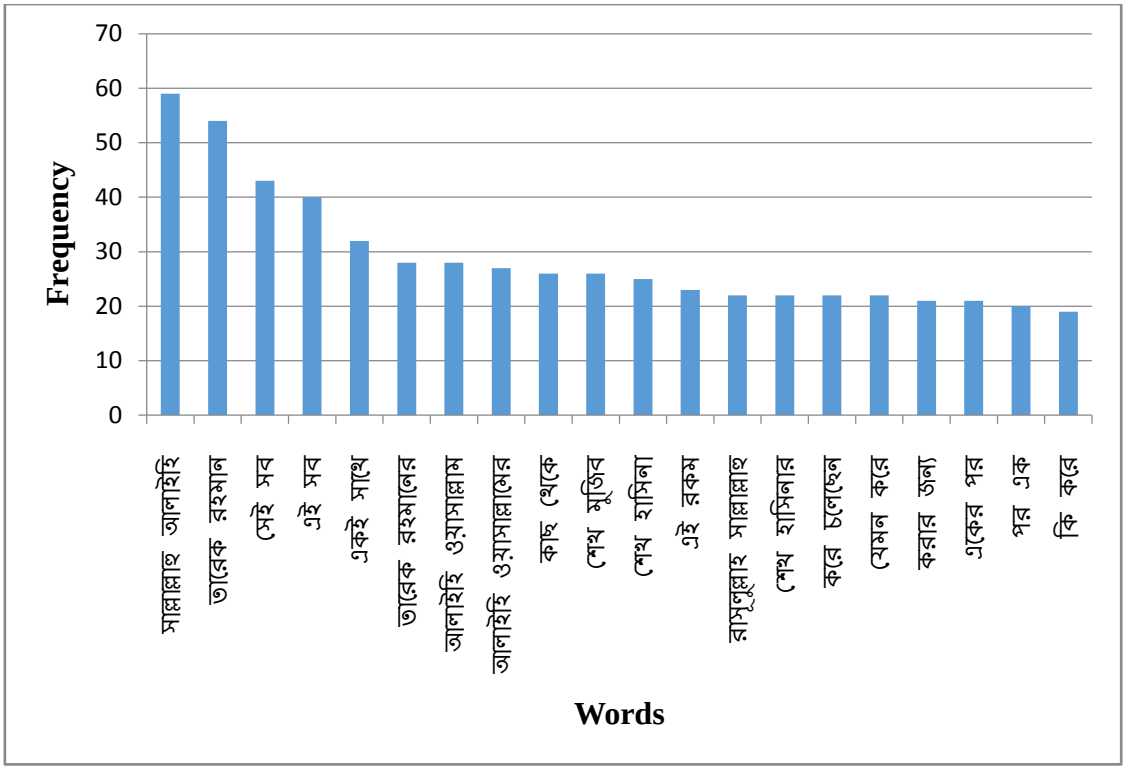

Figure 8: Graphical representation of Top 20 most frequent two consecutive words used by "Syed Shah Salim"

4.3 Calculating Word Length :

The number of letters that creates a word is known as the word length. We calculated the number of words of different lengths written by each writer.

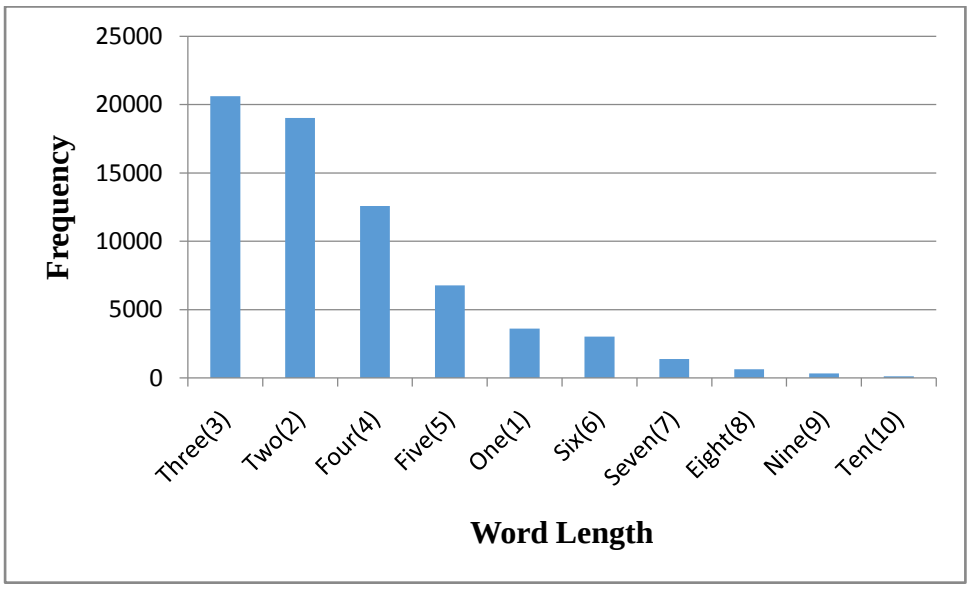

Figure 9: Top 10 most frequent words of different word length used by "Anisul Hoque".

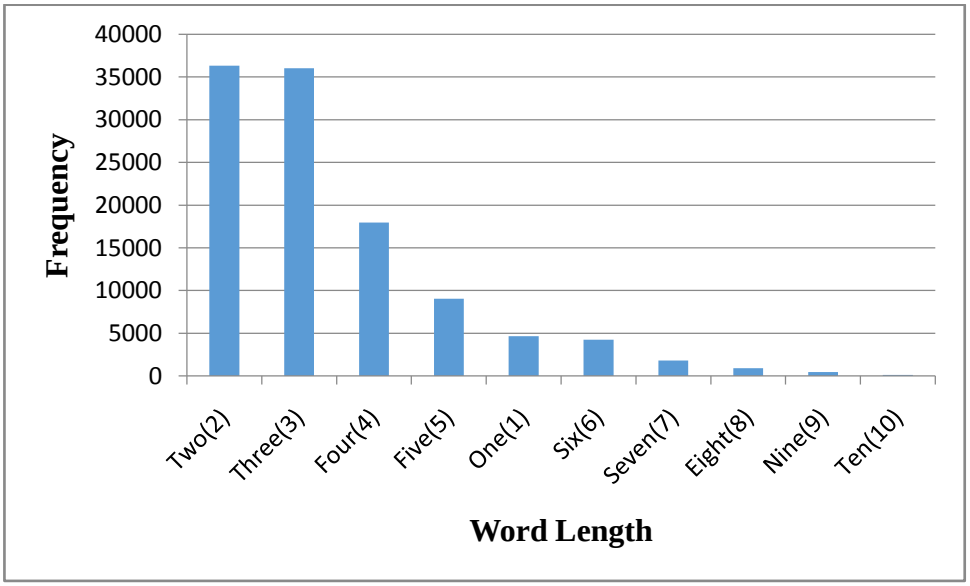

Figure 10: Top 10 most frequent words of different word length used by "Dr. Muhammed Zafar Iqbal"

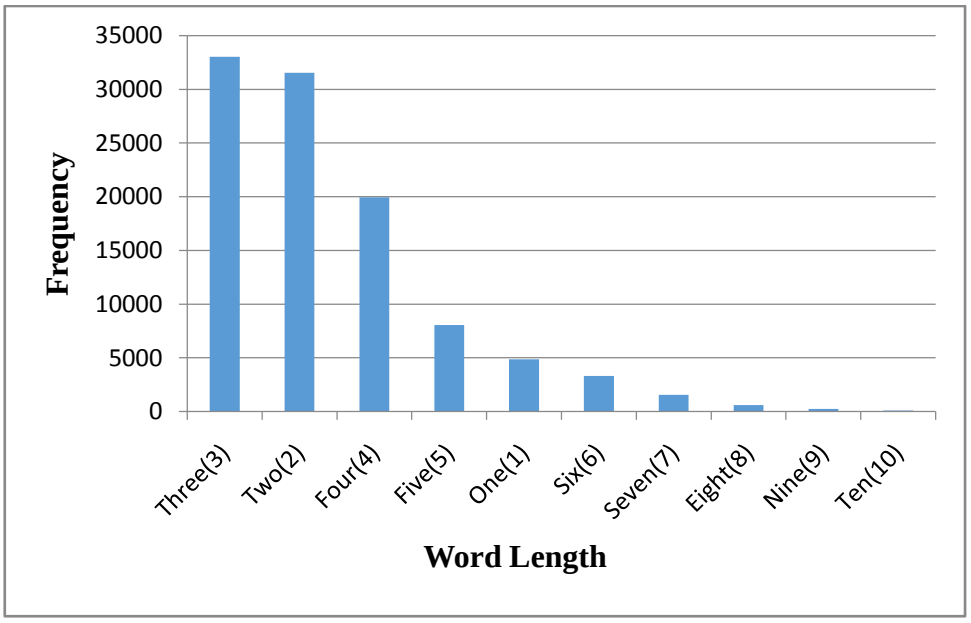

Figure 11: Top 10 most frequent words of different word length used by "Imon Zubair".

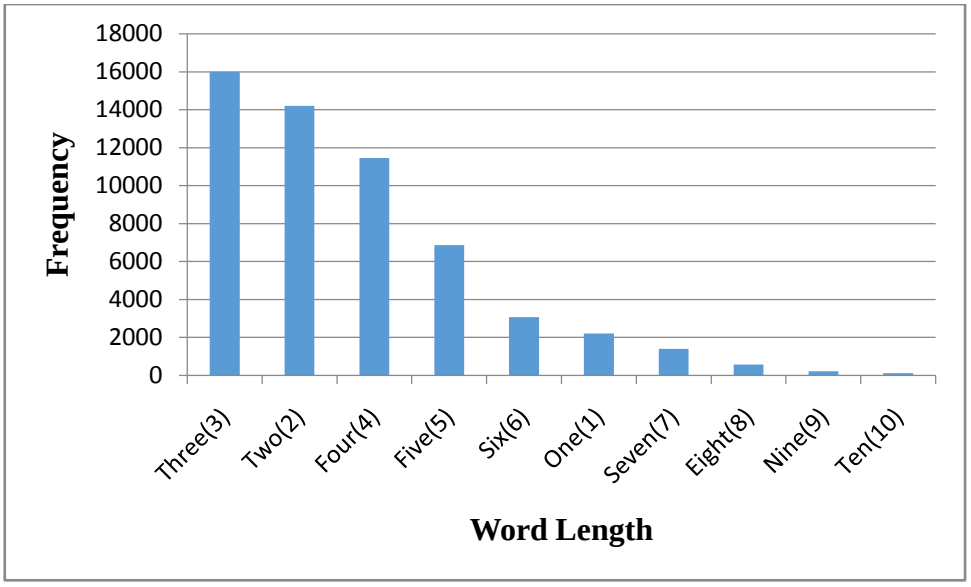

Figure 12: Top 10 most frequent words of different word length used by "Syed Shah Salim".

4.4 Calculating Sentence Length :

Sentence length means the number of words that exist in that sentence. We considered it as a feature. Different writers use sentences of different length. So we might find some variations in this case. We calculated the number of sentences of different lengths written by each writer. For example, we counted the number of sentences of length n1, n2, n3…etc.

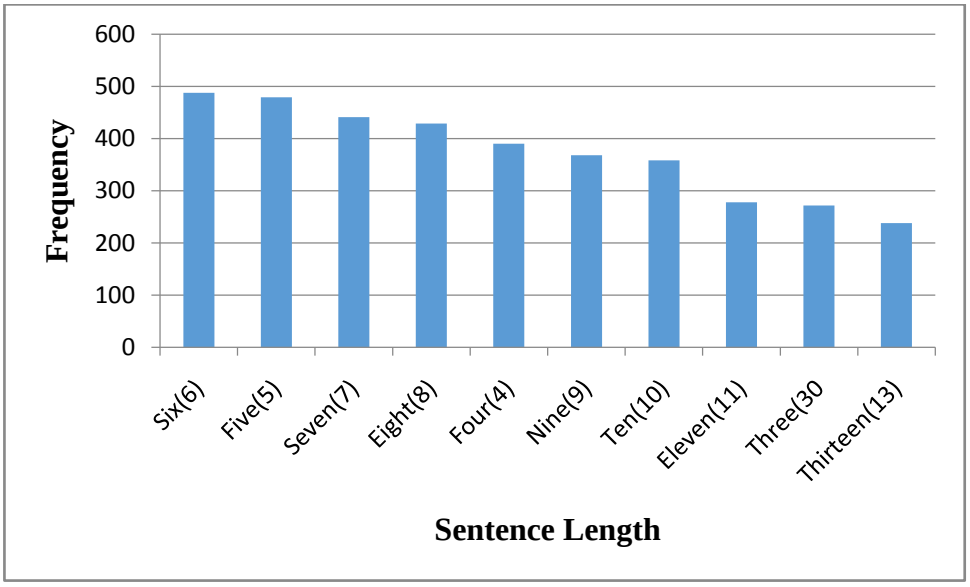

Figure 13: Top 10 most frequent sentences of different length used by "Anisul Hoque".

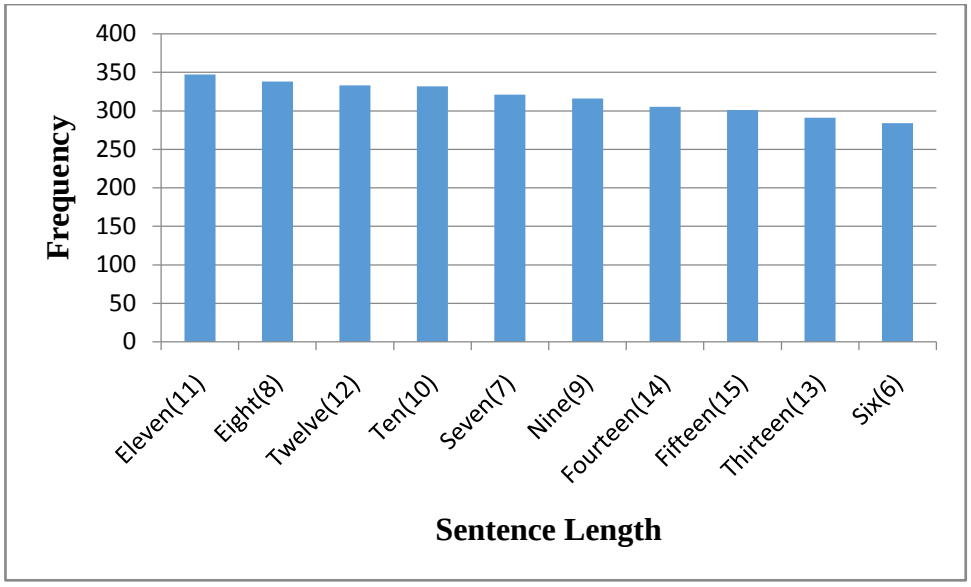

Figure 14: Top 10 most frequent sentences of different length used by "Dr. Muhammed Zafar Iqbal".

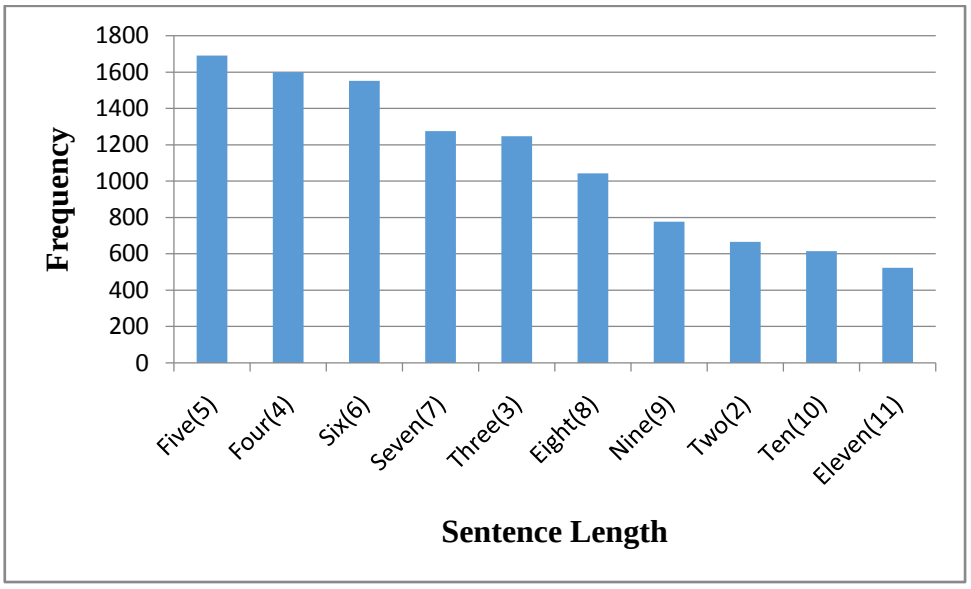

Figure 15: Top 10 most frequent sentences of different length used by "Imon Zubair".

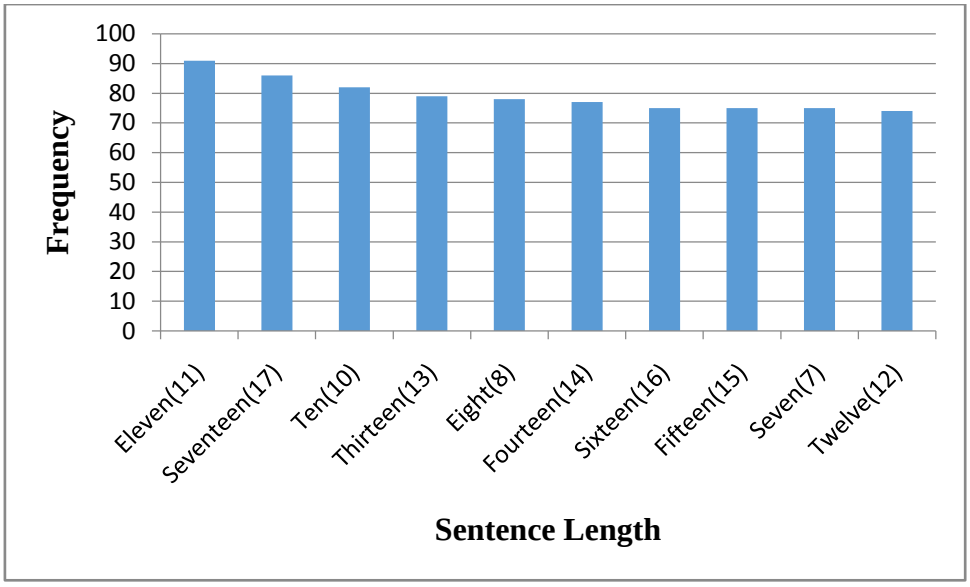

Figure 16: Top 10 most frequent sentences of different length used by "Syed Shah Salim".

4.5 Type-Token Ratio :

Type-token ration is the ratio of number of unique words and total number of words present in a document.

Type-token ratio = *N/V*, here, N= number of unique words
V= number of 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.

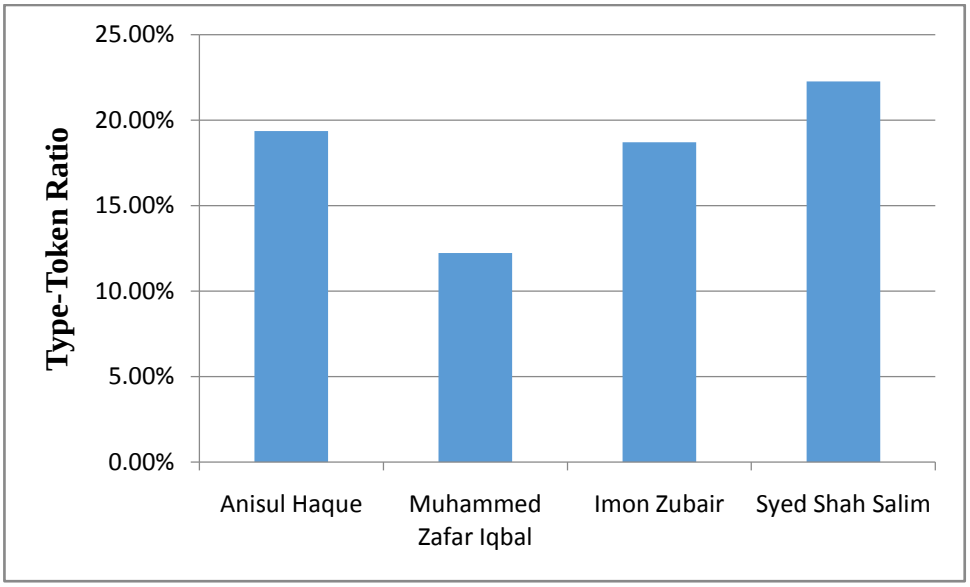

Figure 17: Comparison of Type-Token Ratio among four writers.

4.6 Distribution of Parts-of-Speech :

Distribution of parts-of-speech is an important feature. Using different types of parts-of-speech says a lot about the characteristic of the text. We calculated the number of conjunction and pronoun in the documents.

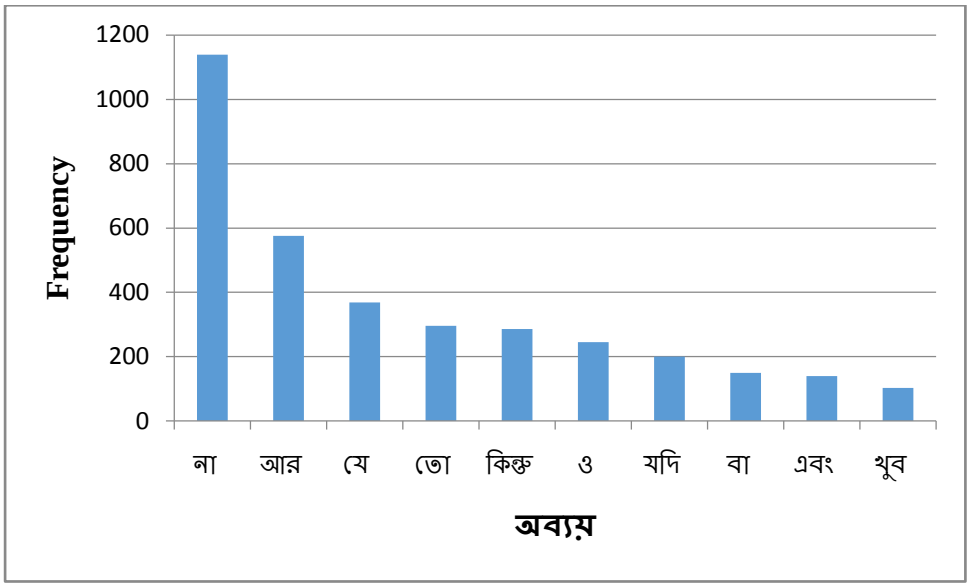

Figure 18: Frequency of Top 10 conjunction (অব্যয়) used by "Anisul Hoque".

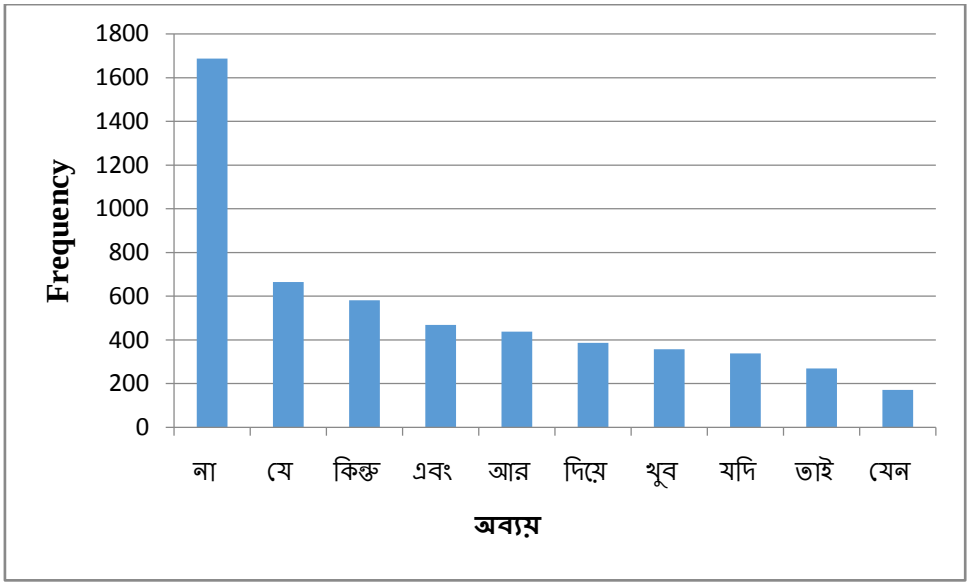

Figure 19: Frequency of Top 10 conjunction (অব্যয়) used by "Dr. Muhammed Zafar Iqbal"

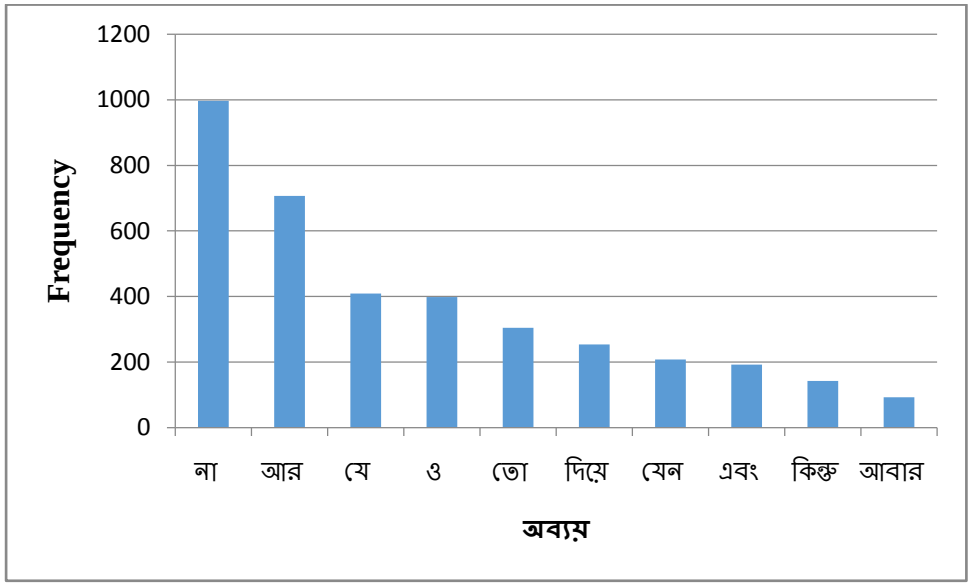

Figure 20: Frequency of Top 10 conjunction **(**অব্যয়**)** used by "Imon Zubair".

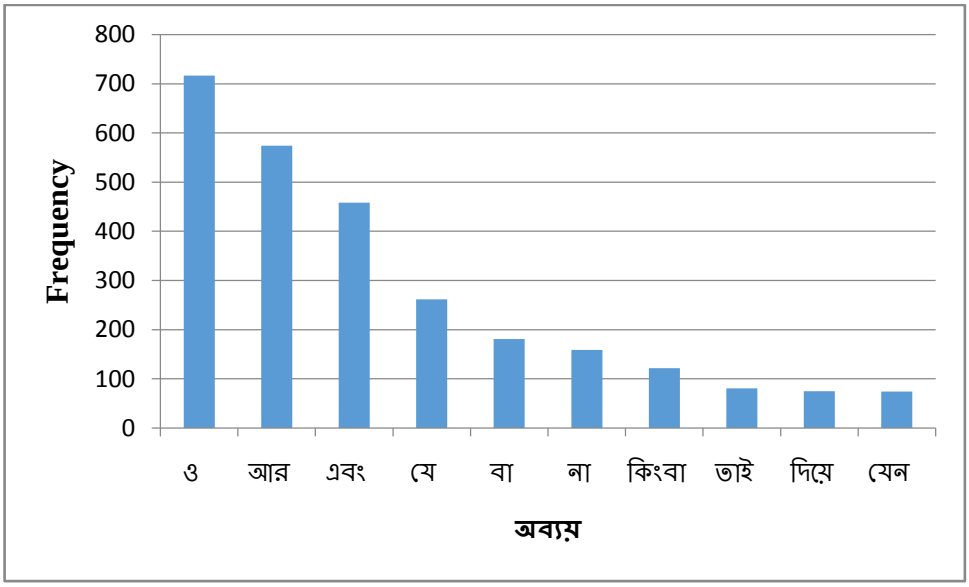

Figure 21: Frequency of Top 10 conjunction (অব্যয়) used by "Syed Shah Salim".

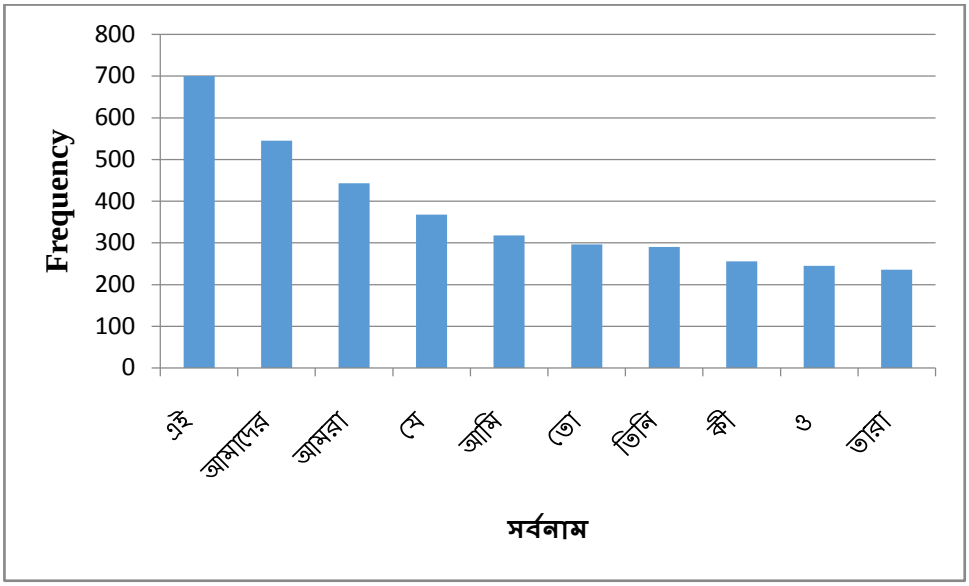

Figure 22: Frequency of Top 10 pronoun (সর্বনাম) used by "Anisul Hoque".

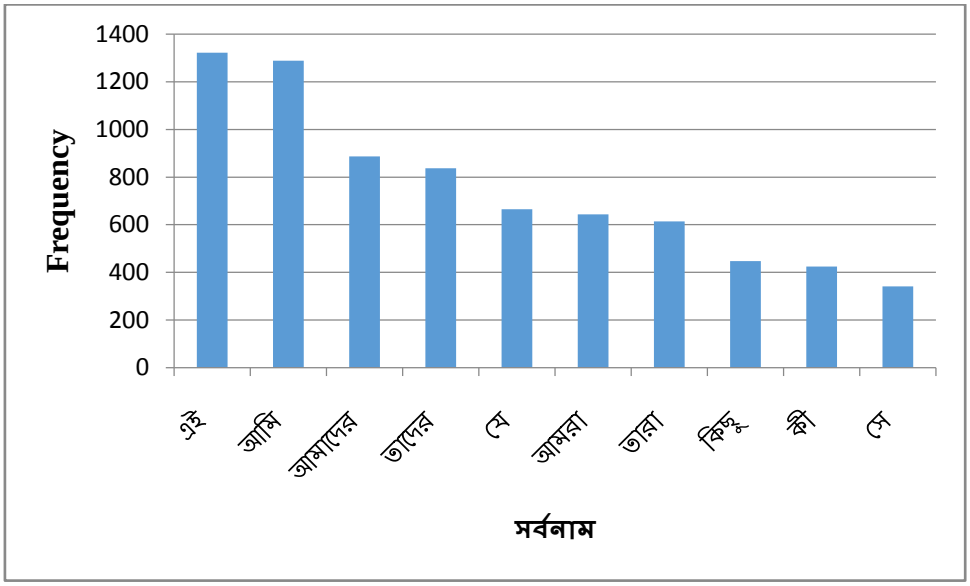

Figure 23: Frequency of Top 10 pronoun (সর্বনাম) used by "Dr. Muhammed Zafar Iqbal".

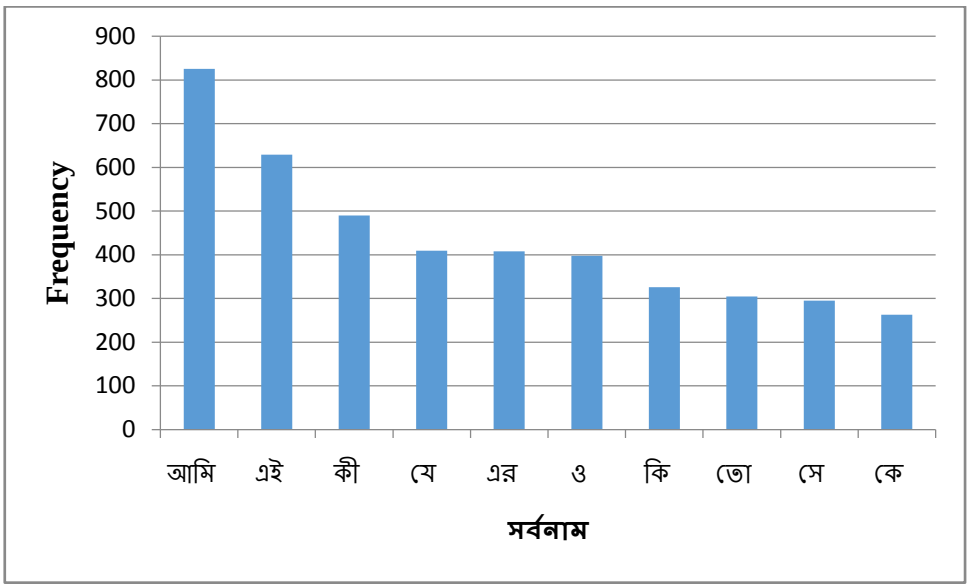

Figure 24: Frequency of Top 10 pronoun (সর্বনাম) used by "Imon Zubair".

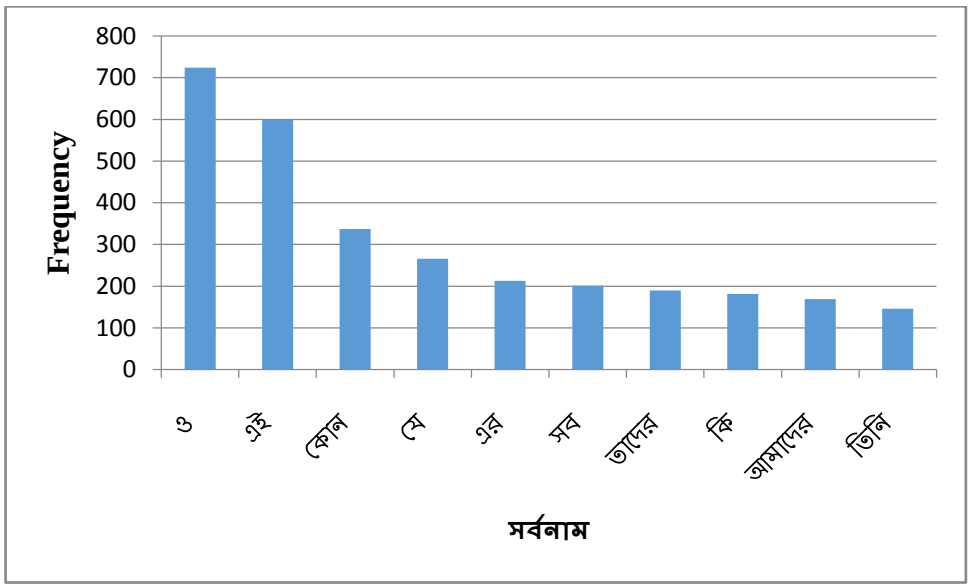

Figure 25: Frequency of Top 10 pronoun (সর্বনাম) used by "Syed Shah Salim".

CHAPTER 5

RESULT ANALYSIS AND DISCUSSION

We gathered some information of all four writers about each feature. Then we compared that information among the writers.

5.1 Comparing word frequency :

We counted the frequency of each word of the writers. Then we compared all the data with one another. We also counted the frequency of two consecutive words and compared the data.

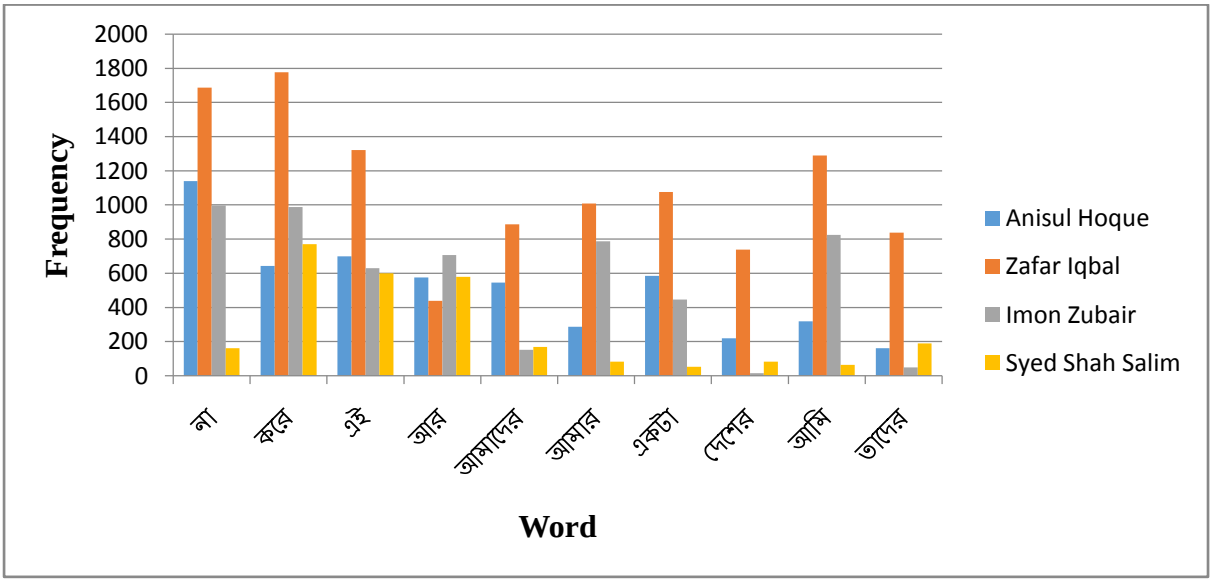

Figure 26: Comparison of one word frequency used by four writers.

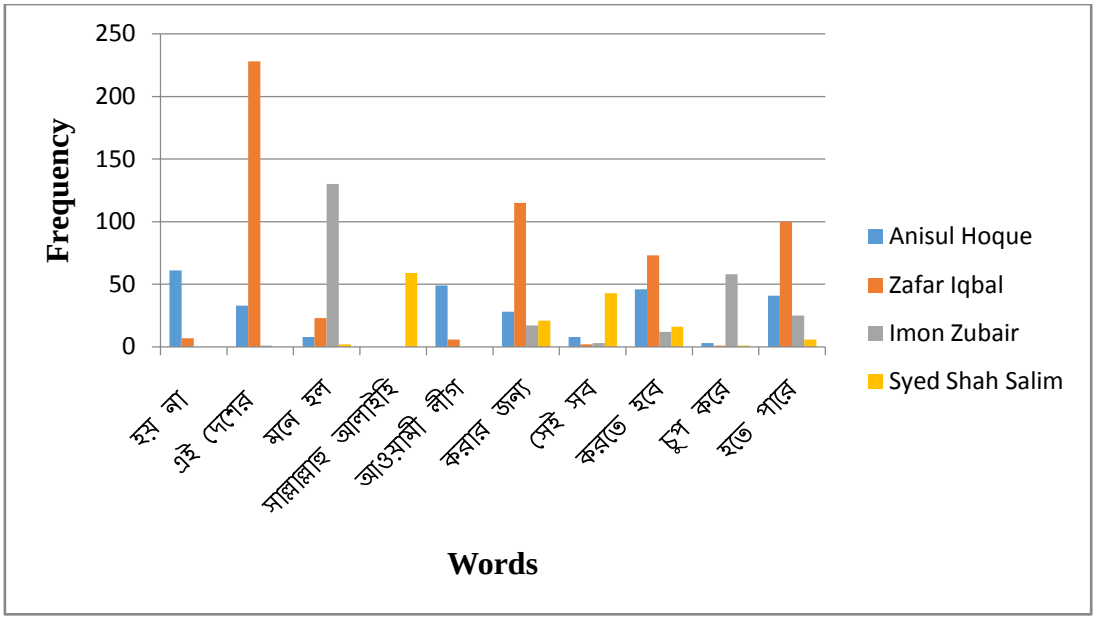

Figure 27: Comparison of consecutive two words frequency used by four writers.

5.2 Comparing word length :

A word is made of a number of letters that is known as the word length. We calculated the number of words of different lengths written by each writer and compared the data of all four writers.

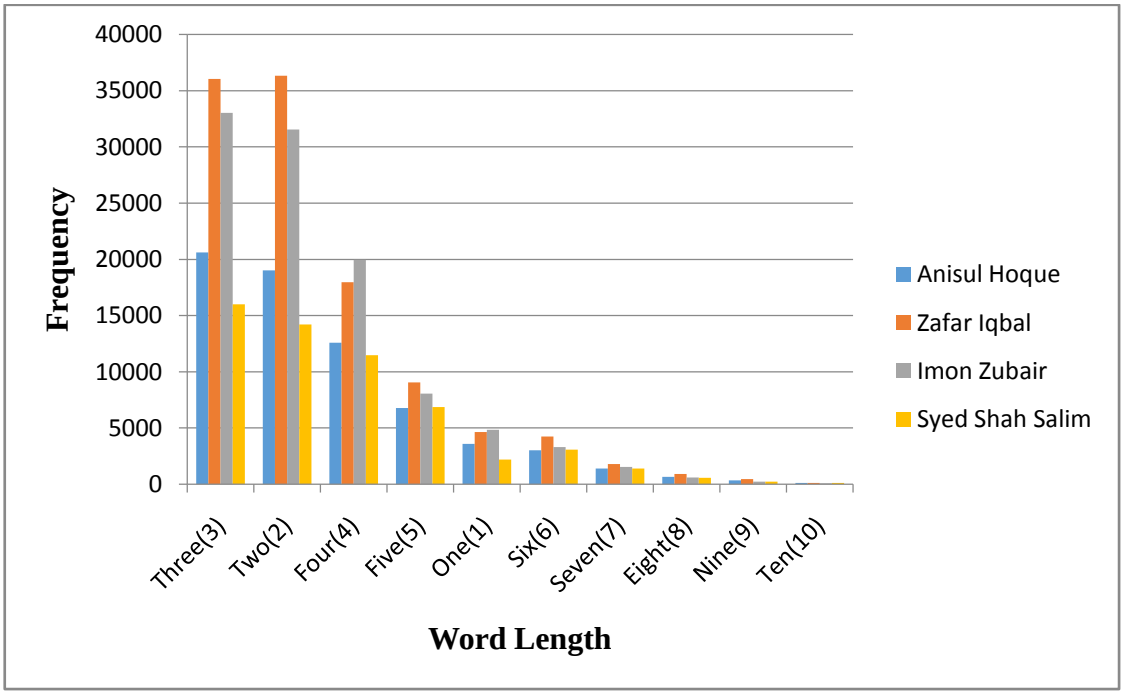

Figure 28: Comparison of word length frequency used by four writers.

5.3 Comparing sentence length :

A sentence consists of number of words which is called sentence length. We counted the frequency of sentences of different lengths. Then we compared the data of all four writers.

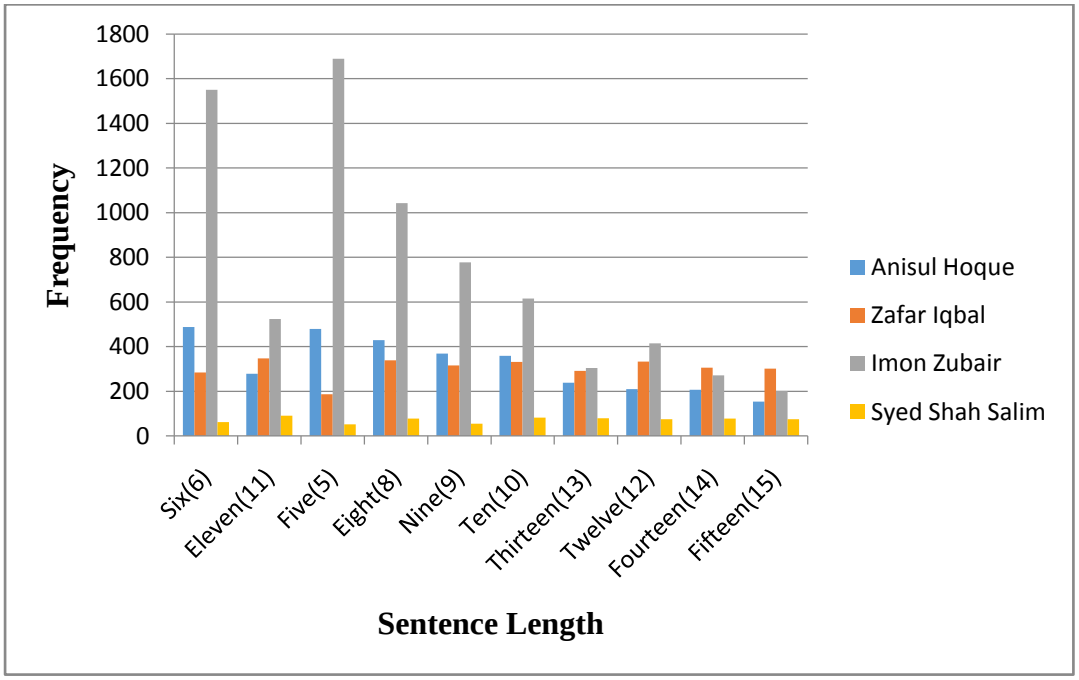

Figure 29: Comparison of sentence length frequency used by four writers.

5.4 Comparing Type-Token Ratio :

We calculated the type-token ration (*N/V)* of each of the writers. Type-token ration indicates to the fact that how rich the vocabulary of a text is. So we compared the ratio of four writers to compare the richness of their vocabulary.

5.5 Comparing distribution of parts-of-speech :

We calculated the frequency of conjunction and pronoun in those texts. We compared some of the frequent words used by almost all writers.

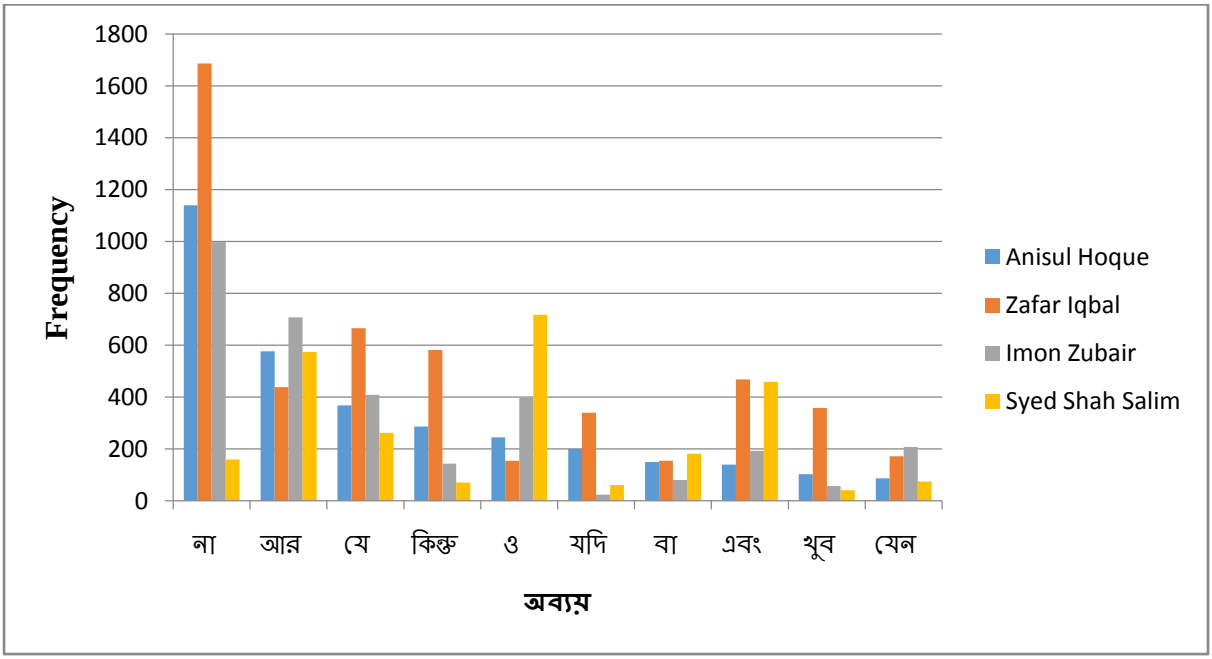

Figure 30: Comparison of frequency of conjunction(অব্যয়) used by four writers

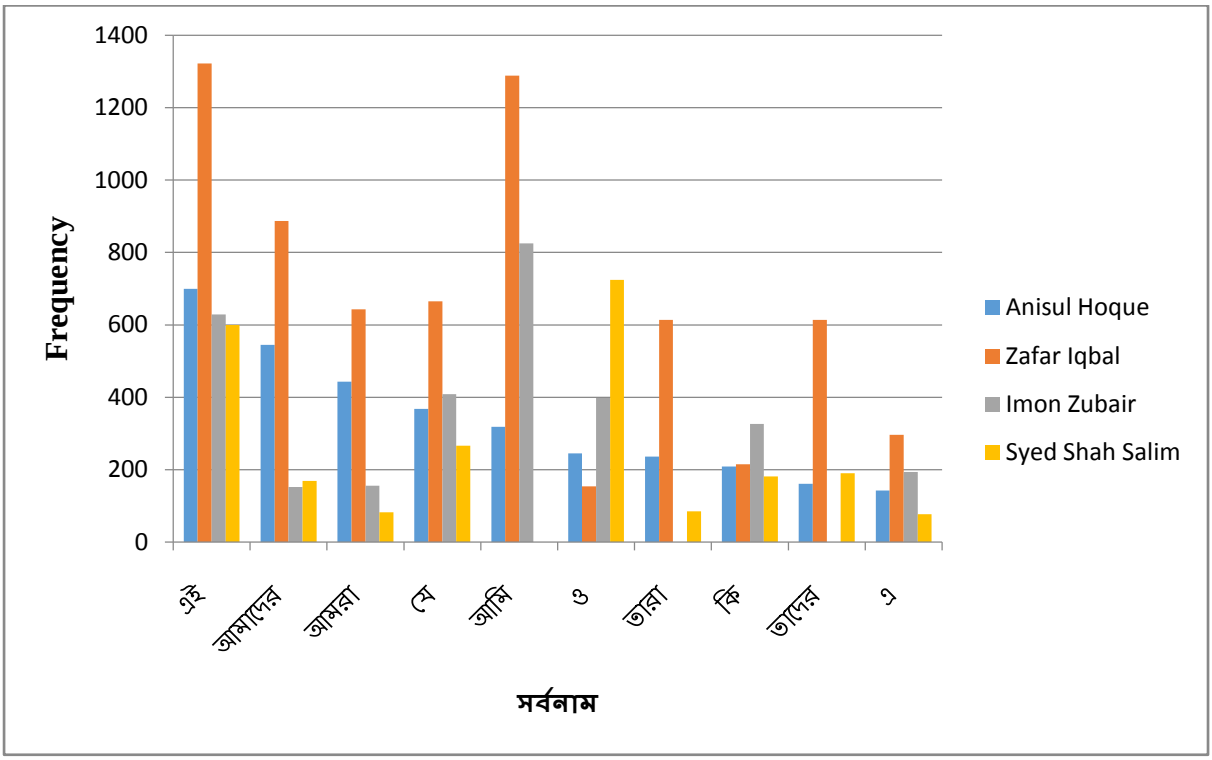

Figure 31: Comparison of frequency of pronoun (সর্বনাম) used by four writers

CHAPTER 6

FUTURE WORK

So far in our study we did analysis on some specific features of Stylogenetics. Using these features we gathered some statistical information about the writers. Our future plan is to work with this statistical information to find the variation among the writers and to find the features that will help us most to identify a writer. For this we plan to use "Dimension Reduction" & "Principal Component Analysis".

6.1 Proposal 1:

We plan to analyze our data using Dimension Reduction. Through this we can find those specific features that we help us to find the difference among writers and identify a writers writing.

Dimension Reduction :

In machine learning and statistics, Dimension Reduction is the process of reducing the number of random variables under consideration, and can be divided into feature selection and feature extraction.

Feature selection also known as variable selection, attribute selection or variable subset selection, it is an approach to try to find a subset of the original variables (also called features or attributes). In some cases, data analysis such as regression or classification can be done in the reduced space more accurately than in the original space. Feature selection, is the process of selecting a subset of relevant features for use in model construction. The central assumption when using a feature selection technique is that the data contains many redundant or irrelevant features. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context.

Feature extraction transforms the data in the high-dimensional space to a space of fewer dimensions. Feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative, non redundant, facilitating the subsequent learning and generalization steps, in some cases leading to better human interpretations.

6.2 Proposal 2 :

Using Principle Component Analysis can help us to reveal the internal structure of the data in a way that best explains the variance in the writings of each writer.

Principal Component Analysis :

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to (i.e., uncorrelated with) the preceding components. The principal components are orthogonal because they are the eigenvectors of the covariance matrix, which is symmetric. PCA is sensitive to the relative scaling of the original variables

6.3 Proposal 3 :

Collecting writings of more writers :

In future we plan on collecting more documents written by some new writers. By adding a few more writers we hope to get a more precise and efficient result in indentifying the characteristic of a writer. Working on a larger amount of corpus will help us to get a better result.

Specially we wish to work on the writing of "Humayun Ahmed".

.

CHAPTER 7

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.

In this paper we worked with blogs written by four different writers. We analyzed with five different features and gathered statistical data. We compared this data of four writers with one another.

In future we will apply Dimension Reduction and Principal Component Analysis to find which feature is most important to find the proper variance in the writings of the writers.

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

REFERENCES :

1) Kim Luyckx,Walter Daelemans and Edward Vanhoutte, “Stylogenetics: Clustering-based stylistic analysis of literary corpora”
2) Roger Peng and Nicolas Hengartner, “Quantitative Analysis of Literary Styles”, 1974;
3) Michael Brennan and Rachel Greenstad,“ Practical Attacks Against Authorship Recognition Techniques”.
4) D. 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), pp. 91-120.
5) Holmes, D. I. (1985), \The Analysis of Literary Style: A Review," Journal of the Royal Statistical Society, Series A, 148, 328{341.
6) Williams, C. B. (1940), \A Note on the Statistical Analysis of Sentence-Length as a Criterion of Literary Style," Biometrika, 31, 356{361.

Appendix :
Raw Data collection:

1) The Blogs of Anisul Haque were downloaded from http://blog.priyo.com/blogs/anisulhaque

2) The Blogs of Imon Zubair were downloaded from http://www.somewhereinblog.net/blog/benqt60

3) The Blogs of Muhammed Zafar Iqbal were downloaded from https://shadashidhekothaarchive.wordpress.com/

4) The Blogs of Syed Shah Salim Ahmed were downloaded from http://blog.priyo.com/blogs/syed-shah-salim-ahmed