Entropy Of Printed Bengali Language Texts

Subrata Pramanik

M. Sc., University of Rajshahi, Bangladesh, 1997

Thesis Submitted In Partial Fulfillment Of

The Requirements For The Degree Of

Master of Science

in

Mathematical, Computer, and Physical Sciences

(Computer Science)

The University Of Northern British Columbia

April 2008

Your file Votre référence
ISBN: 978-0-494-48804-1
Our file Notre référence
ISBN: 978-0-494-48804-1

Abstract

One of the most important sources of information is written and spoken human language. The language that is spoken, written, or signed by humans for general-purpose communication is referred as natural language. Determining the entropy of natural language text is a fundamentally important problem in natural language processing. The study and analysis of the entropy of a language can be a meaningful resource for researchers in linguistics and communication theory. For the purpose of this research we have taken printed Bengali language text as our source of natural language. We have collected a sufficient number of printed Bengali language text samples, and divided them into two classes, Newspaper and Literature. We have studied each class in order to come up with specific entropy for each category and analyzed their characteristics. As a separate study, we collected printed religious Bengali language texts, divided them into two classes, Islamic and Hindu, find out their entropy and studied and analyzed their characteristics. From our research, we have found the Zero and first-order entropy of Bengali language to be 5.52 and 4.55 respectively. And the language uncertainty and redundancy are 0.8242 and 17.58% respectively. These entropy and redundancy results of the language will be useful to the researchers to help find a better text compression method for Bengali language.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to express my deep gratitude to my Supervisor, Dr. Saif Zahir for his support throughout my master's study at UNBC. I am greatly indebted to him for his expert supervision during the time of my research and writing of this thesis. I am confident that I will benefit from his rigorous scientific approach and constructive thinking throughout my future career.

My special thanks goes to Dr. Charles Brown and Dr. Han Donker for their keen interest in this thesis project from the very beginning and their consent to serve as one of the members of my graduate committee. I'm greatly indebted to them for the effort they took to help me with valuable suggestion and constructive input.

I would like to thank my friends at UNBC, Sharmin, Maruf, Reza bhai, Alim, Zaman, Tareq bhai, Julius, Amit, Srinivas, Bharat, Jaison, Heath and Khaldoun for their co-operation, help and encouragement during my study period here.

I'm indebted to my parents and brother for their love, encouragement and support, which always acted as the driving force behind me. Whatever I have achieved in my life so far has been possible because of them. Last and most importantly, I would like to thank my wife for her love, encouragement and patience during my extended absence.

Chapter 1

Introduction

Throughout history, many have reflected on the importance of language. Language is not only a medium of expressing thoughts, perceptions, sentiments, and dreams; it also represents a fundamental expression of social identity. Language is very important in any culture. A language does far more than just enable people to communicate with each other. The language of one country is different from other country and it tells the features of the country which distinguish it from one country to another.

The language that is spoken, written, or signed by humans for general-purpose communication is referred as natural language. And determining the entropy of natural language text is a fundamentally important problem in natural language processing. Entropy is a measure of information. The study and analysis of the entropy of a language can be a meaningful resource for researchers in linguistics and communication theory. The results provide us deeper information about the language; help us in understanding its characteristics better and finding out ways to reduce its redundancy. We can also discover better ways to compress it for digital storage for information processing.

Entropy of a natural language text is the average number of bits per letter of the text that will be required to translate the language into binary bits [1]. From the entropy, the redundancy of a text can be calculated. Higher entropy indicates less redundant and lower entropy indicates a more redundant text. Once we have the redundancy of the language, it will be useful to the research to find out a better text compression method for the language.

1.1 Objective and Importance

We have taken printed Bengali texts as the source of our natural language. The objective of our research is to find out the entropy of printed Bengali language texts and study their characteristics.

Researchers have worked on English [2], Arabic [3], Russian [4] language to find out their entropy and thus paved the way for further linguistic and text compression research in those languages, which only helped in the various fields where those languages are used as a medium of communication. With 230 million native speakers, Bengali is one of the most widely spoken languages of the world (ranked seventh) [5] and when we found out that this kind of research has not been done with Bengali language, it inspired us to take this as our research problem. Applying statistical techniques related to information theory, it is possible to compute an estimate of the entropy rate of Bengali, thus investigating the redundancy and enabling further research for finding out the optimal compression of Bengali texts.

In natural languages, entropy is quite hard to measure. The languages whose entropy is calculated so far exhibits lower entropy, and lower entropy indicates higher redundancy. But this high level of redundancy is what allows us to only read the shape of words; what allows us to understand each other even in noisy environments. Computers which recognize human speech [6] try to use this low entropy in order to make sure they didn't misinterpret a sound; but this doesn't always work, as the redundancy only exists because a very large number of rules (phonetic and syntactic) come with the language and it is difficult to give instructions to a computer about all of them.

In fact, Computer Science is one area of applications where the low entropy of a language is a problem. When we are creating some text file or word document or other, it is not only a mere plain text, the text in fact follows the rules of some natural language (such as English), or some programming language (such as C++, Java, etc). Thus all of them contain much redundant information [7]. The compression algorithms try to utilize this redundant information; they take a file with low entropy and convert it into one which has higher entropy, thus saving space.

As stated previously, determining the entropy of natural language text is a fundamentally important problem in natural language processing. The ability to predict characters or words in text as well as a human, is equivalent to solving the artificial intelligence problem.

Entropy is also useful when defining the concept of unicity distance in cryptography. Unicity distance is a term used in cryptography referring to the length of an original ciphertext needed to break the cipher by reducing the number of possible spurious keys to zero in a brute force attack. When deciphering, if we know what cipher is used, and try each of the keys in turn to decode the secret message and if we know that the decoded message should be in English, then because of English's low entropy, it is highly unlikely there be more than one decoding which actually follows the rules of English [2]; which actually follows the unicity distance concept.

A table of the frequencies of letters in a given language is quite handy for breaking simple substitution ciphers or for devising static Huffman coding tables. Thus, entropy related research with Bengali language will help in Bengali language cryptography also.

1.2 Contributions

In our research work we have collected a sufficient number of printed Bengali language text samples (more than 50,000 characters), and divided them into three classes, Newspaper (15395 characters), Literature (14364 characters) and Religious (20388 characters) class. These characters are saved in 24-bit BMP image for high quality text so that we can process them in future. We have studied the characteristic of each class and compared the classes in order to come up with specific entropy for each class and analyzed their characteristic. For the religious class, we subdivided the samples into two religious classes, Islamic and Hindu. We then found out their entropy and redundancy and compared and analyzed their characteristics. We found out the entropy and redundancy of the total sample data and compared it with the entropies of each class. These entropy and redundancy results of the language will be useful to the researchers to find out a better text compression method for the language, and in other areas of information processing and linguistic research. We have created a table of frequencies of letters of Bengali language that will be helpful for breaking substitution ciphers and for devising static Hoffman coding tables. We have also conducted an analysis of the total sample and find out the frequency distributions of the consonant digraphs.

1.3 Thesis Organization

This thesis is organized as follows:

In Chapter 1, along with the introduction, the objective of our research and its objective and importance is discussed. Our contributions and the outline of the thesis are also presented.

In Chapter 2, theoretical background for entropy and language entropy is provided. Maximum and relative entropy, redundancy and previous works in entropy have also been discussed.

In Chapter 3, After an introduction to the Bengali language, A brief history and the characteristics of the language have been discussed. A detail of the language alphabet has been given at the end.

In Chapter 4, our data collection process, source of the data and the methodology of our research have been outlined. Then results of our work and discussions on the result are provided.

In Chapter 5, there is a conclusion emphasizing the significance of the research work accomplished in this thesis. At last, an outlining of the possible future work that can be based on this thesis is provided.

Chapter 2

Theoretical Background and Previous Work

2.1 Introduction

In information processing and transmission of information research field, one of the most important contributions was made by Claude Shannon through his introduction of mathematical theory of communication [9] and the introduction of a method for estimating the entropy and the redundancy of a language [2].

2.2 Entropy

"Entropy" is a widely used term in various research fields of science, where the term "entropy" is generally interpreted in three distinct, but semi-related ways, i.e. classical thermodynamics, statistical thermodynamics and an information theory viewpoint.

Entropy in information theory is a fundamentally different concept from thermodynamic entropy. The fundamental premise of information theory is that the generation of information can be modeled as a probabilistic process that can be measured in a manner that agrees with the intuition [10]. In information theory, there is a term called self information, which gives the information in a single outcome. But in most cases, it is much more interesting to know the average information content of a source. This average is given by the expected value of the self information with respect to the source's probability distribution. This average of self information is called the source entropy [11]. Generally speaking, the amount of self information attributed to any event E is inversely related to the probability of E. If $P(E) = 1$ (that is, the event always occurs), then $I(E) = 0$ and no information is attributed to it. That is, because no uncertainty is associated with the event, no information would be transferred by communicating that the event has occurred.

The entropy (average self information) of a discrete random variable X is a function of its probability mass function and is defined as:

$$\text{Entropy, } H(X) = -\sum_{i=1}^{N} P_x(x_i) \log_2 (P_x(x_i)) \tag{1}$$

where, N is the number of possible values of X and $P_x(x_i) = \Pr[X = x_i]$. The probability of x_i is defined as the number of occurrences favorable for the symbol x_i, over the number of total occurrences of all the symbols. When log is base 2 then the unit of entropy is bits per (source) symbol. Entropy is a measure of uncertainty in a random variable and a measure of information it can reveal [11].

Again, the word entropy conveys special meaning in different respective research fields of information theory. For example, the entropy of an image is quite different from the concept of language entropy. To understand the different approaches in concepts we start with a short description of Image entropy.

2.2.1 Image Entropy

Image entropy is a quantity which is used to describe the amount of information which must be coded for by a compression algorithm. In other words, it gives a measure of how much information content is there in a source image.

Low entropy images, such as those containing a lot of black sky, have very little contrast and large runs of pixels with the same or similar DN values. An image that is perfectly flat (only one color or intensity) will have entropy of zero. Consequently, they can be compressed to a relatively small size.

A Digital Number or DN is the value stored within a pixel or cell of an image. The DN of the pixel represents the amount of light intensity reflected back to the sensor. DN of 0 will appear as black, of 127 will appear gray, and of 255 will appear white.

On the other hand, high entropy images such as an image of heavily cratered areas on the moon have a great deal of contrast from one pixel to the next and consequently cannot be compressed as much as low entropy images.

For an image x, quantized to M levels, the entropy H_x is defined as [12]:

$$\text{Entropy, } H_x = -\sum_{i=0}^{M-1} P_i \log_2 P_i \tag{2}$$

where, P_i, i=0 to $M-1$, is the probability of the i^{th} quantizer level being used (often obtained from a histogram of the pixel intensities).

H_x represents the average number of bits per pixel with which the quantized image x can be represented using an ideal variable-length entropy code.

In a highly correlated image, the pixels tend to have equiprobable values, which results in *maximum entropy*. If the transformed pixels are de-correlated, certain pixel values become common, thereby having large probabilities, while others are much less. This results in small entropy [12].

An estimate, called the *first-order estimate*, of the entropy of the source can be computed with Eq^n. (2). Better estimates of the entropy of the gray-level source that generated the sample image can be computed by examining the relative frequency of pixel blocks [10] in the sample image, where a block is grouping of adjacent pixels.

For example, if we obtain the entropy by computing the relative frequencies of pairs of pixels, that estimate is called the *second-order estimate* of the source entropy, because it was obtained by computing the relative frequencies of 2-pixel blocks. Although higher order estimates provide even better approximations of source entropy, convergence of these estimates to the true source entropy is slow and computationally involved [10].

2.2.2 Language Entropy

The entropy is a statistical parameter which measures, in a certain sense, how much information is produced on the average for each symbol of a text in the language. If the language is translated into binary digits in the most efficient way, the entropy is the average number of binary digits required per letter of the original language [2].

We assume that a single symbol a_i occurs in some translated data with probability P_i. The probability P_i is defined as the number of occurrences favorable for the i-th symbol, over the number of total occurrences of all the symbols. The entropy of symbol a_i is defined as

($-P_i \log_2 P_i$), where P_i is the probability of occurrence of a_i in the data. The entropy of a_i is the smallest numbers of bits needed to represent symbol a_i.

So, if there are n symbols, then the entropy of the data is $-\sum_1^n P_i \log_2 P_i$

The entropy of the data depends on the individual probabilities P_i, and is largest when all n probabilities are equal [12].

Let us assume that we have a language alphabet containing X number of characters. With this, we can write words and arrange them into sentences. Now, if we take a string of these characters of length y, how many possible strings exist? X^y (because for each character we have X numbers of choices). Now, all of these strings possible are not necessarily legal in a given language. The higher the proportion of legal strings, the higher the entropy.

Binary code again shows us how to measure the entropy of our random language; because it takes $log_2 n$ bits to code a n-character alphabet, the random language has an entropy per character of 4.8. That means that a 28-character alphabet can encode a maximum of 4.8 bits (or pieces) of information with each character [7].

An interesting observation was made by Shannon, he found that the capacity to guess the next letter of a text is a measure of the redundancy of the text. The more guessable the text, more redundant it is and lower the entropy. If the test is perfectly guessable that means zero information and zero entropy [9].

2.2.2.1 Maximum Entropy

The maximum entropy concept has a long history. Laplace [13] may be rightly considered the father of maximum entropy, having introduced the underlying theme 200 years ago in his "Principle of Insufficient Reason":

> *"When one has no information to distinguish between the probabilities of two events, the best strategy is to consider them equally likely."*

As E. T. Jaynes [14], a more recent pioneer of maximum entropy put it:

> *"The fact that a certain probability distribution maximizes entropy subject to certain constraints representing our incomplete information is the fundamental property which justifies use of that distribution for inference; it agrees with everything that is known, but carefully avoids assuming anything that is not known. It is a transcription into mathematics of an ancient principle of wisdom."*

According to the properties of entropy H [15], maximum entropy is reached when the probabilities of occurrence of the n symbols of a particular sequence are equal, i.e., when $p_1 = p_2 = \ldots = p_n = 1/n$ and therefore:

$$H\_maximum = -n\,(1/n)\log_2(1/n) = \log_2 n \qquad (3)$$

Actually, the symbols in a natural language are far from being equiprobable like this, and therefore the first-order entropy H, defined on the base of unconditional probabilities of separate symbols is considerably less than the maximum entropy.

2.2.2.2 Relative Entropy and Redundancy

We define the relative entropy or the relative uncertainty as follows:

$$\text{H\_relative} = \frac{\text{H\_actual}}{\text{H\_maximum}} \tag{4}$$

Where, H_actual refers to the first-order estimate of the entropy and H_maximum is the maximum entropy.

The redundancy can be defined as the difference between H_maximum and H_actual expressed as a fraction of H_maximum [15]:

$$\text{Redundancy} = \frac{\text{H\_maximum} - \text{H\_actual}}{\text{H\_maximum}} \tag{5}$$

$$= 1 - \text{H\_relative} \tag{6}$$

2.3 Previous Works

In information theory, Harvard linguistic professor George Kingsley Zipf [16] first introduced the principle of "least effort" and used this concept to study the economy of words. However, a method for estimating the entropy of language text was first introduced by Claude Shannon [9]. Shannon originally devised the use of entropy to study the amount of information in a transmitted message. According to the definition of information entropy it is expressed in terms of a discrete set of probabilities p_i. In the case of transmitted messages, these probabilities were the probabilities that a particular message was actually transmitted, and the entropy of the message system was a measure of how much information was in the message.

Shannon estimated the entropy of written English in 1950 by having human subjects guess successive characters in a string of text selected at random from various sources. In one experiment, random passages were selected from *Jefferson the Virginian* by Dumas Malone. The subject was shown the previous 100 characters of text and asked to guess the next character until successful. The text was reduced to 27 characters (A-Z and space). Subjects were allowed to use a dictionary and character frequency tables (up to trigram) as aids.

Many researchers continued the research started by Shannon and analyzed his empirical results for English text. Grignetti [17] recalculated Shannon's estimate of the average entropy of words in English text, Treisman [18] commented on contextual constraints in language, White [19] used a dictionary encoding technique to achieve compression for printed English and Jamission [20], Wanas [03] discuss entropy of other languages and Miller [21] discussed the effects of meaningful patterns in English text on the entropy calculations.

Some efforts were made to estimate the higher-order entropies of languages by use of word statistics. Kuepfmueller [22] used the statistics of words and syllables (not including the spaces), but his approach suffers from neglecting the combinations of letters belonging to different neighboring words.

Major progress in developing statistical approaches to entropy evaluation for long symbol sequences (not necessarily texts in natural languages) has been achieved by Grassberger [23] who introduced efficient methods based on modifications of the Lempel-Ziv universal coding algorithms [24].

2.3.1 Shannon's Prediction method

Shannon [2] proposed a unique method, which overcomes the limitations of the statistical approach, to obtain upper and lower bounds of the conditional entropy for higher order estimates of entropies.

The conditional entropy expresses the uncertainty of a symbol following a set of length L (an L-gram). An L-gram is a string of L consecutive letters all belonging to a standard alphabet. If the entropy is close to zero, it means that the symbol can be predicted almost for certain when the previous L-gram is given. The larger the entropy, the more difficult it is to predict the following symbol. This means, in particular, that the predicted symbol can be wrong, and that more than one attempt may be needed to obtain the right result. It suggests to us that it is possible to go in the opposite direction and to extract information about the value of the entropy from the results of a prediction experiment.

Shannon suggested using a human being – a person experienced in the language – as a predictor for an experiment of this sort. The reason for this in his own words [9]:

> *"The new method of estimating entropy exploits the fact that anyone speaking in a language possesses, implicitly, an enormous knowledge of the statistics of the language. Familiarity with the words, idioms, clichés and grammar enables him to fill in missing or incorrect letters in proof-reading or to complete an unfinished phrase in the conversation."*

One of the most important factor of Shannon's method in comparison with the direct statistical approaches is that the probabilities of L-grams for large L cannot be obtained from statistical data not only because of computational difficulties but also because the total amount of texts in any language is also limited. Only a small part of all the possible meaningful larger L-grams can be found in published texts. The total length of all the existing texts is not enough in order to find probabilities of long L-grams. However, a human guesser can usually suggest several different continuations of a given possible text which is much larger than the totality that actually exists. In fact, a human being possesses such an enormous variety of possible texts because of his knowledge; and this knowledge, mostly intuitive, brings the guesser by reading a previous text into a state which is close to the state of the source itself, and that enables him to predict efficiently the continuation. [1]

Chapter 3

Bengali Language

3.1 Introduction

Bengali or Bangla is an Indo-Aryan language of the eastern Indian subcontinent, evolved from the Magadhi Prakrit, Pāli and Sanskrit languages. Its immediate predecessor was 'Magadhi Apabhransha' [5]. From this emerged the three languages - Bengali, Oriya and Assamese. Of these three, Oriya has separate script, while Bengali and Assamese share similar script (except for very few differences).

Bengali is the national language of Bangladesh and one of 18 languages listed in the Indian Constitution. It is the administrative language of the Indian states of Tripura and West Bengal, as well as one of the administrative languages of Kachar district, Assam.

With nearly 230 million native speakers, Bengali is one of the most widely spoken languages in the world, making it the seventh language after Chinese, English, Hindi-Urdu, Spanish, Russian and Arabic. It is perhaps the only language on the basis of which an independent state was created. Bengali is the main language spoken in Bangladesh; in

India, it is ranked as the second most spoken language. Along with Assamese, it is geographically the most eastern of the Indo-Iranian languages. [5]

3.2 History:

Bengali emerged as a new Indo-Aryan language by 900-1000 AD through Magadhi Apabhramsa and Abahattha, two stages of Magadhi Prakrit (600 BC - 600 AD), along with two other Indo-Aryan languages, Oriya and Assamese (Fig. 1). Until the 14th century, there was little linguistic difference between Bengali and Assamese.

The evolution of Bengali may be divided into three historical phases [25]:

1) **Old Bengali** (900/1000 CE–1400 CE): Texts include *Charyapada*, devotional songs; emergence of pronouns *Ami*, *tumi*, etc; verb inflections *-ila*, *-iba*, etc. Oriya and Assamese branch out in this period.
2) **Middle Bengali** (1400–1800 CE): Major texts of the period include Chandidas's *Srikrishnakirtan*; elision of word-final *ô* sound; spread of compound verbs; Persian influence. Some scholars further divide this period into early and late middle periods.
3) **New Bengali** (since 1800 CE): Shortening of verbs and pronouns, among other changes (e.g. *tahar* → *tar* "his"/"her"; *koriyachhilô* → *korechhilo* he/she had done).

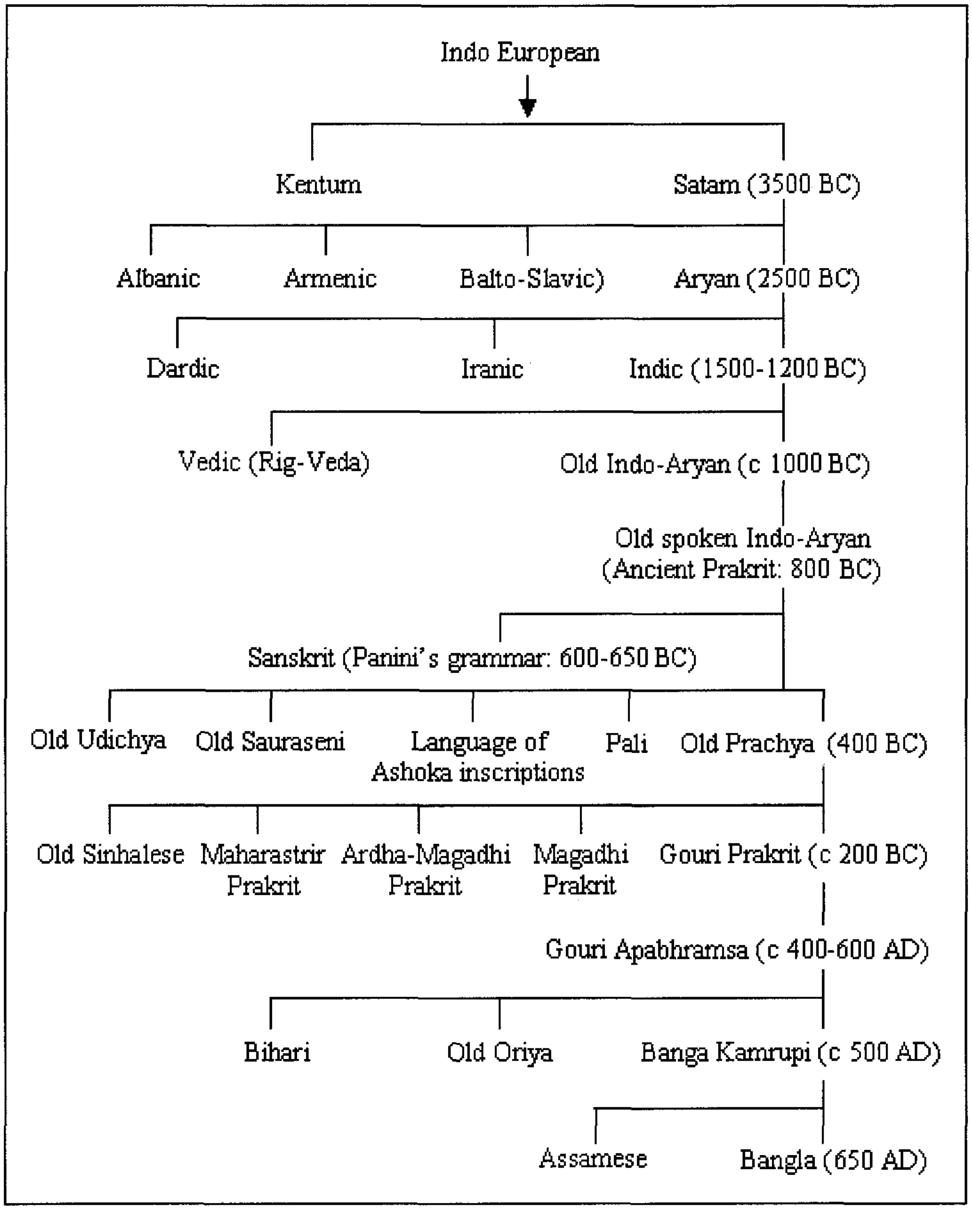

Figure 1: Evolution of Bengali language [5]

Historically closer to Pali, Bengali saw an increase in Sanskrit influence during the Middle Bengali (Chaitanya era), and also during the Bengal Renaissance. Of the modern Indo-European languages in South Asia, Bengali and Marathi retain a largely Sanskrit vocabulary base while Hindi and others are more influenced by Arabic and Persian.

Until the 18th century, there was no attempt to document the grammar for Bengali. The first written Bengali dictionary/grammar, *Vocabolario em idioma Bengalla, e Portuguez dividido em duas partes*, was written by the Portuguese missionary Manoel da Assumpcam between 1734 and 1742 while he was serving in Bhawal. Nathaniel Brassey Halhed, a British grammarian, wrote a modern Bengali grammar (*A Grammar of the Bengal Language (1778)*) that used Bengali types in print for the first time. Raja Ram Mohan Roy, the great Bengali Reformer, also wrote a "Grammar of the Bengali Language" (1832) [26].

During this period, the *Choltibhasha* form, using simplified inflections and other changes, was emerging from *Shadhubhasha* (older form) as the form of choice for written Bengali.

Bengali was the focus, in 1951–52, of the Language movement (*Bhasha Andolon*) in what was then East Pakistan (now Bengalidesh). Although Bengali speakers were more numerous in the population of Pakistan, Urdu was legislated as the sole national language. On February 21, 1952, protesting students and activists walked into military and police fire in Dhaka University and three young students and several others were killed. Subsequently, UNESCO has declared 21 February as International Mother Language Day [26].

3.3 Characteristics of the language

3.3.1. Influence of other languages

Bengali has been greatly influenced by two non-Aryan languages [5]:

Dravidian and Kol.

Their influence is evident not only in the vocabulary but also in the construction of sentences. A large number of onomatopoeic words, repetitive words and conjunctive verbs in Bengali reveal non-Aryan influence; for example, words such as *ghoda-toda* (horses etc), *kapad-chopad* (clothes etc), *tuk-tuk*, *khatkhat*, *khankha*, *dhandha*, *basiya pada* (sitting down), *lagiya thaka* (to persevere), etc. There are plenty of Dravidian and other non-Aryan words in Bengali, especially in place names, indicating that Bengali passed through many stages and was influenced by various other languages.

One of the main influences on Bengali was that of Sanskrit as this language was the vehicle of literature and culture for almost the whole of the subcontinent since the beginning of the Christian era. (The religious discourses of the Buddhists and the Jains were carried on in Pali and Ardhamagadhi respectively.) In the days of old Bengali, many Bengalis used to write poetic works in Sanskrit. Even after the evolution of Bengali, many well-known Bengali poets, such as Jaydev, Umapatidhara and Govardhan Acharya, continued to compose their literary works in Sanskrit. The result was that many pure Sanskrit words entered Bengali from the very early stages.

Following the establishment of Muslim rule in Bengal in the 13th century, Bengali came under the influence of Arabic, Persian and Turkish. Persian was the language of the court during Muslim rule in the 14th and 15th centuries. Because of this special status as well as other cultural influences, Bengali picked up many Persian words at this time. In the 16th century, with the Portuguese inroads, several Portuguese words entered Bengali; for example, words such as *anaras* (pineapple), *ata* (custard-apple) and *tamak* (tobacco).

A table comparing the similarities between Bengali, Sanskrit, English & Latin words						
Language	**Word**					
English	month	mother	new	night	nose	three
Latin	mensis	mater	novus	nox	nasus	tres
Sanskrit	mãs	matar	nava	nakt	nãs	trayas
Bengali	maash	mata	nobo	ratri	naakh	tin

Table 1: Comparing the similarities between Bengali, Sanskrit, English & Latin language words [27]

From the 17th century, the Dutch, French and English started arriving in Bengal. As a result, words from these languages started entering Bengali vocabulary; for example, from the French: *cartouche*, coupon, *depot*; English: table, chair, lord/*lat*, general/*jadrel*, etc. During the 17th and 18th centuries effective use of Bengali prose began through the efforts of Christian missionaries.

With the start of British rule in the 18th century and the spread of English education, Bengali started absorbing increasing numbers of English words. Following the establishment of the Bengali Department at Fort William College in Calcutta in 1801, the efforts of its head, William Carey, and his associate Bengali scholars, made Bengali fit for fine prose.

During the 19th century, the efforts of Bengali writers contributed to the further growth of the language. Among them were Raja Rammohan Roy, Bhabanicharan Bandyopadhyay, Iswar Chandra Vidyasagar, Bankimchandra Chattopadhyay, Michael Mdhusudan Dutt, and Mir Mosharraf Hossain. The 20th century witnessed the elevation of colloquial Bengali to a written literary medium through the work of many talented writers such as Rabindranath Tagore (Nobel Loriete) and Pramatha Chowdhury.

3.3.2 Dialects

Regional variation in spoken Bengali constitutes a dialect continuum. Linguist Suniti Kumar Chatterjee grouped these dialects into four large clusters - Radh, Banga, Kamarupa and Varendra. The south-western dialects (Radh) form the basis of standard colloquial Bengali, while Bangali is the dominant dialect group in Bengalidesh. In the dialects prevalent in much of eastern and south-eastern Bengal (Barisal, Chittagong, Dhaka and Sylhet divisions of Bengalidesh), many of the stops and affricates heard in West Bengal are pronounced as fricatives. The influence of Tibeto-Burman languages on the phonology of Eastern Bengali is seen through the lack of nasalized vowels. Some variants of Bengali, particularly Chittagonian and Chakma-Bengali, have contrastive tone; differences in the pitch of the speaker's voice can distinguish words [5].

Rajbangsi, Kharia Thar and Mal Paharia are closely related to Western Bengali dialects, but are typically classified as separate languages. Similarly, Hajong is considered a separate language, although it shares similarities to Northern Bengali dialects.

During the standardization of Bengali in the late 19th and early 20th century, the cultural center of Bengal was its capital Kolkata (then Calcutta). What is accepted as the standard form today in both West Bengal and Bengalidesh is based on the West-Central dialect of Nadia, a district located near Kolkata. There are cases where speakers of Standard Bengali in West Bengal will use a different word than a speaker of Standard Bengali in Bengalidesh, even though both words are of native Bengali descent. For example, *nun* (salt) in the west corresponds to *lôbon* in the east.

Even in Standard Bengali, vocabulary items often divide along the split between the Muslim populace and the Hindu populace. Due to cultural and religious traditions, Hindus and Muslims might use, respectively, Sanskrit-derived and Perso-Arabic words. Some examples of lexical alternation between these two forms are:

- hello: nômoshkar (S) corresponds to assalamualaikum/slamalikum (A)
- invitation: nimontron/nimontonno (S) corresponds to daoat (A)
- paternal uncle: kaka (S) corresponds to chacha (S/Hindi)
- water : jol (D) corresponds to pani (S)

(here, S = derived from Sanskrit, D = deshi; A = derived from Arabic)

3.3.3 Forms of Written Language

Written Bengali has two forms [5]:

a) **Sadhu** or chaste and

b) **Chalita** or colloquial or spoken.

The two differ basically in verbs and pronouns. The verbs and pronouns get shortened in the colloquial form. For example: করিয়া (kariya; to do) করে (kare); তাহার (tahar; his/hers) তার (tar). The importance of the colloquial form arose at the beginning of the 20th century but the use of chaste Bengali did not disappear totally. Chaste language continued to be used in contemporary newspapers, works of documentation and in statements by the government and on matters of serious import. Colloquial Bengali was the language of the Calcutta gentry, a considerable number of whom used the colloquial form to write literary works.

The parallel currents of chaste and colloquial streams created a unique phenomenon of diglossia in Bengali. Although the main peculiarity of the colloquial stream is the shortened form of verbs and pronouns, their real difference is in temperament. The mix of sadhu and chalita, as used in poetry, has been on the wane since World War II, giving way to the chalita form only. Since March 1965, many Bengali newspapers have adopted the chalita form, discarding the sadhu one. "The ITTEFAQ", which had retained the sadhu form, has also started using the chalita form since 2001.

Hindus and Muslims differ in their ways of using the language, and even West Bengalis and Bengalideshis differ somewhat in their practices. The Muslim rule in Bengal prior to the British rule led to an extensive development of Bengali and a plentiful influx of Arabic, Persian and Turkish vocabulary. Towards the end of the 18th century, even high-caste Hindus used to cultivate the court language, Persian, allowing their Bengali to be influenced by it. Even today over 2,000 Arabic and Persian words relating to war, taxation, legal and cultural matters, and crafts are in use in Bengali.

Such words and their impact increased substantially in the language of the Muslim rural masses of East Bengal prior to the partition of India in 1947. A major difference exists in the language used by Hindus and Muslims in respect of words that refer to relatives or food. Hindus use Sanskrit and Bengali words, while Muslims use Urdu and Arabic words, eg *kaka/chacha* (uncle), *ma/amma* (mother), *baba/abba* (father), *didi/bubu* (sister), *dada/bhaiya* (brother), *jal/pani* (water), *mangsa/gosht (meat).*

At the same time, it should be noted that Muslims in the Jessore area also use the so-called 'Hindu terms' of *didi* and *dada*. Although the written language of West Bengal and Bengalidesh is more or less similar, spoken Bengali differs widely. There are also many regional Bengali dialects. Some dialects, such as those of Sylhet, Noakhali and Chittagong, differ so greatly from each other and standard Bengali, which people of one region can hardly communicate with people of the other.

3.4 Bengali Alphabet

3.4.1 Notable features

- To start with Bengali does not have the concept of upper and lower case letters. It is written from left to right and top to bottom as in English.
- In Bengali alphabet each consonant has an inherent vowel and has two different pronunciations, the choice of which is not always easy to determine and which is sometimes not pronounced at all [27].
- Vowels can be written independently or by using specific diacritical marks.
- When two consonants occur together in clusters, they change their original shape and take a special form.

3.4.2 The alphabet

There are 11 vowels (Fig. 2), 35 consonants (Fig. 3) and 5 modifier symbols (Fig. 4) in Bengali language alphabet. A selection of conjunct consonants and Numerals are also shown in Fig. 5 and 6 respectively [27].

অ	আ	ই	ঈ	উ	ঊ	ঋ	এ	ঐ	ও	ঔ
a	ā	i	ī	u	ū	ṛ	e	ai	o	au
[ɔ, o]	[ɑː]	[i, e]	[i]	[u, o]	[u]	[ri]	[e, æ]	[oj]	[o]	[ow]
ক	কা	কি	কী	কু	কূ	কৃ	কে	কৈ	কো	কৌ
ka	kā	ki	kī	ku	kū	kṛ	ke	kai	ko	kau

Figure 2: Vowels and vowel diacritics of Bengali alphabet

ক	ka	[kɔ]	খ	kha	[kʰɔ]	গ	ga	[gɔ]	ঘ	gha	[gʱɔ]	ঙ	ṅa	[ŋɔ]
চ	ca	[ʧɔ]	ছ	cha	[ʧʰɔ]	জ	ja	[ʤɔ]	ঝ	jha	[ʤʱɔ]	ঞ	ña	[nɔ]
ট	ṭa	[ʈɔ]	ঠ	ṭha	[ʈʰɔ]	ড	ḍa	[ɖɔ]	ঢ	ḍha	[ɖʱɔ]	ণ	ṇa	[ɳɔ]
ত	ta	[t̪ɔ]	থ	tha	[t̪ʰɔ]	দ	da	[d̪ɔ]	ধ	dha	[d̪ʱɔ]	ন	na	[nɔ]
প	pa	[pɔ]	ফ	pha	[pʰɔ]	ব	ba	[bɔ]	ভ	bha	[bʱɔ]	ম	ma	[mɔ]
য	ya	[jɔ]	র	ra	[rɔ]	ল	la	[lɔ]						
শ	śa	[ʃɔ/sɔ]	ষ	ṣa	[ʃɔ]	স	sa	[ʃɔ/sɔ]	হ	ha	[ɦɔ]			
য়	ya	[ʤɔ]	ড়	ṛa	[ɽɔ]	ঢ়	ṛa	[ɽɔ]						

Figure 3: Consonants of Bengali alphabet

ক্ক	kka	ক্ট	kṭa	ক্ত	kta	ক্ব	kba	ক্ম	kma	ক্র	kra	ক্ল	kla	ক্ষ	kṣa	ক্ষ্ম	kṣma
ক্স	ksa	গ্ধ	gdha	গ্ন	gna	গ্ব	gba	গ্ম	gma	গ্ল	gla	ঘ্ন	ghna	ঙ্ক	ṅka	ঙ্ক্ষ	ṅkṣa
ঙ্খ	ṅtha	ঙ্গ	ṅga	ঙ্ঘ	ṅgha	ঙ্ম	ṅma	চ্ছ	ccha	চ্ছ্ব	cchba	চ্ঞ	cña	জ্জ	jja	জ্জ্ব	jjba
জ্ঝ	jjha	জ্ঞ	jña	জ্ব	jba	ঞ্চ	ñca	ঞ্ছ	ñcha	ঞ্ঝ	ñjha	ট্ট	ṭṭa	ট্ব	ṭba	ণ্ট	ṇṭa
ণ্ঠ	ṇṭha	ণ্ড	ṇḍa	ণ্ণ	ṇṇa	ণ্ম	ṇma	ত্ত	tta	ত্ত্ব	ttba	ত্থ	ttha	ত্ন	tna	ত্ব	tba
ত্ম	tma	ত্র	tra	দ্দ	dda	দ্ধ	ddha	দ্ব	dba	দ্ভ্র	dbhra	ন্ট	nṭa	ন্ড	nḍa	ন্ত	nta
ন্ত্ব	ntba	ন্ত্র	ntra	ন্দ	nda	ন্ধ	ndha	ন্ন	nna	ন্ব	nba	ন্স	nsa	প্ট	pṭa	প্ত	pta
প্ন	pna	প্প	ppa	প্ল	pla	প্স	psa	ফ্ল	phla	ভ্র	bhra	ভ্ল	bhla	ম্ন	mna	ম্ফ	mpha
ম্ব	mba	ম্ল	mla	ল্ট	lṭa	ল্ড	lḍa	ল্ব	lba	ল্ল	lla	শ্চ	shcha	ষ্ক	ṣka	ষ্ট	ṣṭa
ষ্ণ	ṣṇa	স্ক্র	skra	স্ত	sta	স্ত্র	stra	স্ব	sba	হ্ন	hna	হ্ম	hma	হ্ব	hba	হ্ল	hla

Figure 4: A selection of conjunct consonants

◌্	hasanta - mutes inherent vowel	ক্	k [k]
ৎ	khanda-ta - final unaspirated dental	কৎ	kat [kɔt]
◌ং	anusvāra - final velar nasal	কং	kaṁ [kɔŋ]
◌ঃ	visarga - adds voiceless breath after vowel	কঃ	kaḥ [kɔh] / [kɔ]
◌ঁ	chandra-bindu - nasalises vowels	কঁ	= kñ [kɲ]

Figure 5: Modifier symbols of Bengali alphabet

০	১	২	৩	৪	৫	৬	৭	৮	৯	১০
শূন্য	এক	দুই	তিন	চার	পাঁচ	ছয়	সাত	আট	নয়	দশ
śūnya	ek	dui	tin	cār	pñāc	chay	sāt	āṭ	nay	daś
0	1	2	3	4	5	6	7	8	9	10

Figure 6: Numerals of Bengali alphabet.

3.5 Bengali Language Entropy:

According to Shannon, capacity to guess the next letter of an English text is a measure of the redundancy of English. More guessable the text, it is more redundant and lower the entropy. If the test is perfectly guessable that means zero information and zero entropy.

Example: 4 symbols, অ, আ, র, এ

অ =00, আ =01, র =10, এ =11

In general, with n symbols, codes need to be of length $\log_2 n$, rounded up. For Bengali text, 46 letters + space = 47 symbols, length = 6 since $2^5 < 47 < 2^6$ (replace all punctuation marks by space).

3.5.1 First-Order model of Bengali

We know that, if a symbol S has probability p, its *self-information* is

$$H(S) = \log_2(1/p) = -\log_2 p$$

For example:

S	অ	আ	র	এ
p	.25	.25	.25	.25
$H(S)$	2	2	2	2
p	.6	.2	.1	.1
$H(S)$	.74	2.32	3.32	3.32

Table 2: Two sample examples for determining entropy

Now, from Eq^n (1),

First-Order Entropy of Source

= Weighted Average Self-Information

$$= -\sum_{k=1}^{L} P_k \log_2 P_k$$

So, for the above examples:

S	অ	আ	র	এ
p	.25	.25	.25	.25
-logp	2	2	2	2
-plogp	.5	.5	.5	.5
p	.6	.2	.1	.1
-logp	.74	2.32	3.32	3.32
-plogp	.444	.464	.332	.332

Table 3: Calculations for determining entropy of the Table 2 sample examples.

First Order Entropy for the first case

$= -\sum p\log p = 2$

First Order Entropy for the second case

$= -\sum p\log p = 1.572$

Chapter 4

Simulation and Results

4.1 Data Collection

For the purpose of this research we have collected a sufficient number of printed Bengali language text samples containing more than 50,000 numbers of characters. We collected samples representing three classes:

- Newspaper class,
- Literature class and
- Religious class.

Our objective was to find out the entropy of these classes separately and compare them with one another and also with the total sample. We also sub-divided the religious class into two sub-classes, namely Islamic and Hindu class, in order to find out their entropy separately and study and analyze the characteristics of these texts.

The newspaper samples that we used are available on the internet. They are taken from "*Prothom Alo*" and "*Shamokaal*", two major daily newspapers of Bangladesh, and from "*Anandabazar Patrika*", which is the main Bengali daily newspaper of India. Newspaper topics include a wide range of applications such as general news, science, health, sports, politics, travel, weather, etc.

As for the literature, we scanned them from book pages. For literature samples, we chose the material from well known books by distinguished author of Bengali language. Samples are taken from the following books:

- Banaphuler Shrestha Galpa – By Balaichand Mukherjee [28]
- Detective – By Rabindranath Tegore [29]
- Anukul – Satyajit Ray [30]
- Debi – Humayun Ahmed [31]
- Elebele – Humayun Ahmed [32]
- Sanket – Shirshendu Mukhopadhyay [33]
- Maa – Anisul Haque [34]
- Krishnapakhsa – Humayun Ahmed [35]

Samples of Islamic and Hindu religious texts are taken from two sources. From scanned pages of religious books, such as Al-Kuran, Srimadbhagbad Geeta, Saraswati bratakatha, etc; and from Islamic and Hindu religious columns published in newspapers and journals.

Sample texts obtained for Newspaper class contain more than 15000 characters, Literature class contain more than 14000 characters and from Religious class contain more than 20000 characters. The exact number of total characters is 50247. These samples are saved in 24-bit BMP image for high quality text.

The untagged version of this corpus is uploaded in the following location so that researchers can process them in future:

http://rapidshare.com/files/104424388/Untagged_Corpus_2.zip

There are two versions, one is the main untagged corpus and in the other version each Bengali character is separated by a space for ease of understanding and reading. The untagged file is compressed into a zip file and is password protected, password is "UNBC".

4.2 Methodology

For our research we have calculated the entropy of the Bengali religious texts using the basic entropy calculation method described in section 2.2. As there is no previous research work available on finding out the entropy of Bengali language, we stayed with the basic method for entropy calculation. For the Bengali alphabets, we have taken into account all 11 vowels and 35 consonants, a total of 46 characters, and have not considered the modifiers or the spaces in between the words.

For each class text we have:

1) Calculated the total number of characters in the sample texts.

2) Calculated the number of occurrences of each letter of the Bengali alphabet in the sample texts.

3) Generated the frequency table for occurrence of Bengali alphabet.

4) Calculated the probability of each letter in the sample texts.

5) Calculated the entropy of each character in the sample texts.

6) Calculated the entropy of full sample texts.

To calculate the entropy of the Bengali language, total samples are considered and the following is performed:

1) Calculated the total number of characters in total sample texts.

2) Calculated the number of occurrence of each letter of the Bengali alphabet in total sample texts.

3) Generated the frequency table for occurrence of Bengali alphabet characters and study the behavior

4) Calculated the probability of each letter in total sample texts.

5) Calculated the entropy of each character in total sample texts.

6) Calculated the entropy of total sample texts.

For the calculation of the occurrence of characters, we have not used the Unicode system, as there is a basic problem using Unicode for Bengali alphabet. As stated previously in section 3.4.1, the consonants of Bengali alphabet have an inherent vowel and have two different pronunciations, the choice of which is not always easy to determine and which

Letter	English equivalent	Letter	English equivalent	Letter	English equivalent
অ	a	উ	u	ঐ	oi
আ	aa	ঊ	uu	ও	o
ই	i	ঋ	ri	ঔ	ou
ঈ	ii	এ	e		

(a) English equivalent used for 11 Vowels

Letter	English equivalent	Letter	English equivalent	Letter	English equivalent
ক	k	ড	d	ম	m
খ	kh	ঢ	dh	য	y
গ	g	ণ	nn	র	r
ঘ	gh	ত	ta	ল	l
ঙ	ng	থ	tha	শ	sh
চ	ch	দ	da	ষ	shh
ছ	chh	ধ	dha	স	s
জ	j	ন	n	হ	h
ঝ	jh	প	p	য়	ya
ঞ	na	ফ	ph	ড়	rr
ট	t	ব	b	ঢ়	rrr
ঠ	th	ভ	bh		

(b) English equivalent used for 35 Consonants

Table 4: Template of equivalent English characters used for (a) Vowels and (b) Consonants to convert the Bengali texts into its English equivalent.

is sometimes not pronounced at all. Again, some conjunct consonants take an entirely different shape when occurring together in a word. These are not possible to distinguish in Unicode format separately. For this shape shifting problem and inherent vowel we haven't used the Unicode system; we have implemented the Bengali texts into English text files and carried out the character count by hand. The template of equivalent English characters used for Bengali characters are shown in Table 4. The calculation related programming is implemented in Matlab 7.0 and carried out on a Pentium IV Windows XP workstation.

4.3 Results and Discussion

4.3.1 Literature Class

The total data samples collected for newspaper class contain 14364 characters. The 4 most frequently occurring vowels and consonants and their frequency rates for the Literature class are shown in Tables 5 and 6 respectively.

Letter	Occurrence	Total char	Frequency (%)
আ	1786	14364	12.70
এ	1558	14364	12.29
অ	1476	14364	8.80
ই	936	14364	5.64

Table 5: 4 Most Frequently Occurred Vowels of Literature Texts and Their Frequency Rates

Letter	Occurrence	Total char	Frequency (%)
র	981	14364	6.83
ক	681	14364	4.74
ন	664	14364	4.62
ব	564	14364	3.93

Table 6: 4 Most Frequently Occurred Consonants of Literature Texts and Their Frequency Rates

The top 20 most frequently occurred characters are plotted in a graph (Fig. 7) in the following order (ordered in terms of their frequency):

আ, এ, অ, র, ই, ক, ন, ব, ম, ত, ল, স, উ, ও, প, য়, দ, হ, ট, ছ.

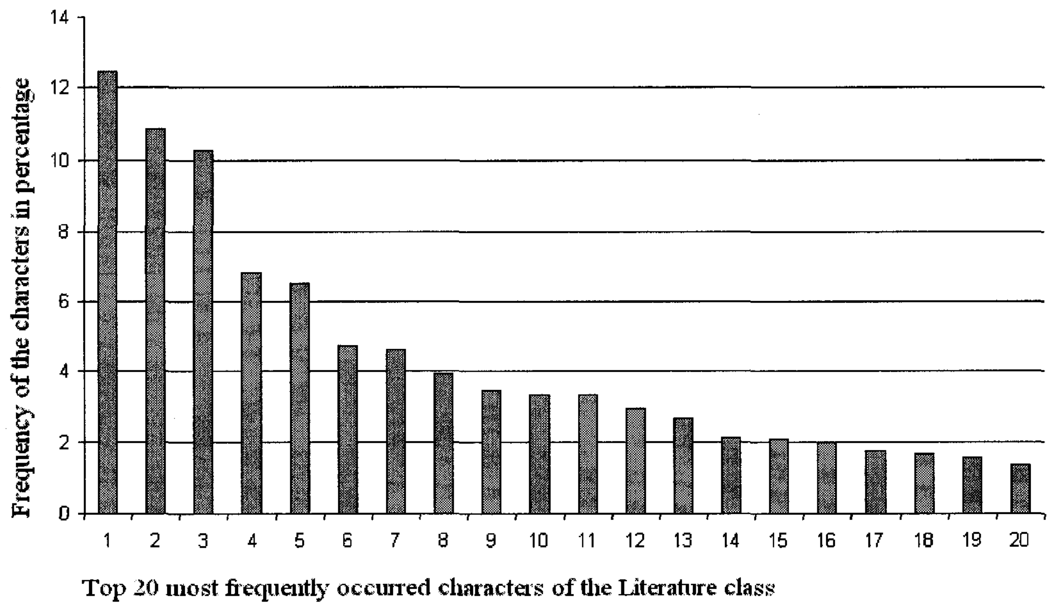

Figure 7: Histogram of the top 20 characters of literature class.

A more detailed table containing their number of occurrence and frequency is shown in Table 7:

Letter	Letter class	Occurrence	Total char	Frequency (%)
আ	Vowel	1786	14364	12.43
এ	Vowel	1558	14364	10.85
অ	Vowel	1476	14364	10.28
র	Consonant	981	14364	6.83
ই	Vowel	936	14364	6.52
ক	Consonant	681	14364	4.74
ন	Consonant	664	14364	4.62
ব	Consonant	564	14364	3.93
ম	Consonant	497	14364	3.46
ত	Consonant	480	14364	3.34
ল	Consonant	478	14364	3.33
স	Consonant	422	14364	2.94
উ	Vowel	383	14364	2.67
ও	Vowel	305	14364	2.12
প	Consonant	299	14364	2.08
য়	Consonant	284	14364	1.98
দ	Consonant	250	14364	1.74
হ	Consonant	242	14364	1.68
ট	Consonant	224	14364	1.56
ছ	Consonant	195	14364	1.36

Table 7: Top 20 Most Frequently Occurred Character and Their Frequency (%) of literature texts.

As evident from Table 7, vowels occupy 4 of the top 5 most frequently occurred characters list. Of all the characters, vowels constitute 45.84% of the total literature data sample and consonants constitute 54.16%. The top 20 characters constitute 88.46% of the data sample.

Now H-maximum refers to the theoretical maximum entropy, this being achieved only when the letters of the alphabet are all equiprobable [15]. From (3),

$$\begin{aligned} H\_maximum &= \log_2 46 \\ &= 5.5236 \end{aligned}$$

From our calculation, the actual entropy or 1st order entropy that we got for literature class is:

$$H\_actual = 4.5232$$

Hence, from eq^{n} (4) and (6), for literature class:

$$\begin{aligned} H\_relative &= 4.5232 / 5.5236 \\ &= 0.81889 \\ Redundancy &= 1 - 0.81889 \\ &= 0.18111 \end{aligned}$$

So, the relative uncertainty is 81.89% for Literature texts and the redundancy is 18.11%. We will compare it later with the redundancy found for other classes.

4.3.1.1 Observation on character frequencies:

As has been mentioned before, in the Bengali alphabet each consonant has an inherent vowel and has two different pronunciations. Most of the time this inherent vowel is “অ”, and this character is used in Bengali words widespread and can be found occurring in many of the commonly used Bengali words in Literature sample, such as অনেক, অতি, সময়, হল, অপর, কম, নয়, খবর, যখন, তখন, আসল, নকল, etc.

The second most frequently occurred character “আ” is also found in numerous most commonly used words, such as মা, বাবা, আমি, আমার, তোমার, তারা, আপনাকে, সকাল, পা, রাখা, চাকা, হাসপাতাল, নাম, আশা, etc. In many of these words the character when used with consonants changes its shape from “আ” to “◌া” and the second shape never occurs alone, occurs only after a consonant.

Likewise, other vowels ই, ঈ, উ, ঊ, ঋ, এ, ঐ, ও and ঔ, when they occur with consonants, they change their shapes into “◌ি”, “◌ী”, “◌ু”, “◌ূ”, “◌ৃ”, “◌ে”, “◌ৈ”, “◌ো” and “◌ৌ”. The dotted circle in the script means that a consonant has to be placed in that space.

Among the rest of the characters of the top 20 table, “এ” is found in the commonly used words like ছেলে, মেয়ে, যেতে, খেতে, আছে, করে, গিয়ে, পেয়ে, করেছে, গিয়েছে, পেয়েছে, আছে, নেবে, etc. “র” is found in common words like আমার, তোমার, আমরা, তোমরা, করা, করে, করতে, তার, পর, শহর, বছর, রিকশা, etc.

These words are very frequently used in common texts. As a result, the characters associated with them are also found to occur more frequently. Similarly,

“ই” is found in words like নেই, তাই, চাই, একই, ছিল, তিনি, এই, বই, কবি, কমিটি, সমিতি, etc.

“ক” is found in words like কি, কে, এক, একা, একক, করা, করে, করতে, করেছে, কম, চাকা, etc.

“ন” is found in words like নিয়ে, নুতন, এখন, অনেক, মনে, নয়, যখন, তখন, নকল, তিনি, etc.

“ব” is found in words like বাবা, সব, খবর, বই, কবি, নেবে, বন, সুবিধা, হবে, বাকি, বিষয়, etc.

“ম” is found in words like মা, আমি, আমার, আমরা, তোমার, সময়, মতো, মন, কম, মেয়ে, etc.

“ত” is found in words like তুমি, তোমার, তাদের, তার, মতো, তিন, হাত, তখন, যেতে, খেতে, etc.

“ল” is found in words like সকাল, সকল, দল, কাল, জল, আসল, নকল, ছেলে, করছিল, হয়েছিল, etc.

“স” is found in words like সব, সকল, সময়, সেরা, আসল, সমিতি, সুবিধা, সকাল, সহ, হাসপাতাল, etc.

“উ” is found in words like তুমি, দুই, সুবিধা, উচিত, কিছু, খুব, খুশি, মুখ, সুখ, চুপ, নুতন, etc.

“ও” is found in words like ওর, ওরা, যাওয়া খাওয়া, মতো, ওঠে, কোনো, তোমার, তোমরা, ছোট, etc.

“য়” is found in words like মেয়ে, সময়, হয়, গিয়ে, যাওয়া, খাওয়া, নিয়ে, দিয়ে, যায়, হয়েছে, etc.

“দ” is found in words like দিয়ে, দুই, দেখা, দেখে, দল, তাদের, দিন, আদর, বেদনা, দুনিয়া, etc.

“হ” is found in words like হয়ে, হয়, হয়েছে, হবে, হল, শহর, হাসপাতাল, হাসি, মহিলা, মহল, etc.

“ট” is found in words like এটা, সেটা, একটি, দুটি, সাইট, কাটা, কাটিয়ে, টিকে, টানা, টেনে, ছোট, etc.

“ছ” is found in words like ছেলে, আছে, হয়েছে, করেছে, গিয়েছে, পেয়েছে, ছিল, বছর, ছোট, কিছু, etc.

4.3.2 Newspaper Class

The total data samples collected for newspaper class contain 15395 characters. The 4 most frequently occurring vowels and consonants, and their frequency rates for Newspaper class are shown in Tables 8 and 9 respectively.

Letter	Occurrence	Total char	Frequency (%)
অ	1721	15395	11.18
আ	1663	15395	10.80
এ	1518	15395	9.86
ই	1054	15395	6.85

Table 8: 4 Most Frequently Occurred Vowels of Newspaper texts and their frequency rates

Letter	Occurrence	Total char	Frequency (%)
র	1302	15395	8.46
ক	785	15395	5.10
ন	738	15395	4.79
ত	575	15395	3.73

Table 9: 4 Most Frequently Occurred Consonants of Newspaper texts and Their Frequency Rates

The top four most frequently used vowels and consonants are almost same as the one we have found for the literature class before, just with an exception of "ত" occupying the fourth most occurred consonant place in space of "ব". But, if we look at a more detailed table (Table 10), there is not much difference between the occurrence of "ত" and "ব", and "ব" is the 5th most frequently occurred consonant here.

A detailed table containing the number of occurrence and frequency of top 20 characters are shown in Table 10:

Letter	Letter class	Occurrence	Total char	Frequency (%)
অ	Vowel	1721	15395	11.18
আ	Vowel	1663	15395	10.80
এ	Vowel	1518	15395	9.86
র	Consonant	1302	15395	8.46
ই	Vowel	1054	15395	6.85
ক	Consonant	785	15395	5.10
ন	Consonant	738	15395	4.79
ত	Consonant	575	15395	3.73
ব	Consonant	544	15395	3.53
স	Consonant	477	15395	3.10
প	Consonant	399	15395	2.59
ম	Consonant	380	15395	2.47
ল	Consonant	381	15395	2.47
য়	Consonant	376	15395	2.44
উ	Vowel	372	15395	2.42
ও	Vowel	340	15395	2.21
দ	Consonant	282	15395	1.83
ট	Consonant	223	15395	1.45
য	Consonant	220	15395	1.43
হ	Consonant	208	15395	1.35

Table 10: Top 20 Most Frequently Occurred Character and Their Frequency (%) of newspaper texts.

As evident from Table 9, vowels occupy 4 of the top 5 most frequently occurred characters list. Of all the characters, vowels constitute 44.57% of the total newspaper data sample and consonants constitute 55.43%. And the top 20 characters constitute 88.06% of the total newspaper sample.

The top 20 most frequently occurred characters are plotted in a graph (Fig. 8) in the following order (ordered in terms of their frequency):

অ, আ, এ, র, ই, ক, ন, ত, ব, স, প, ম, ল, য়, উ, ও, দ, ট, য, হ.

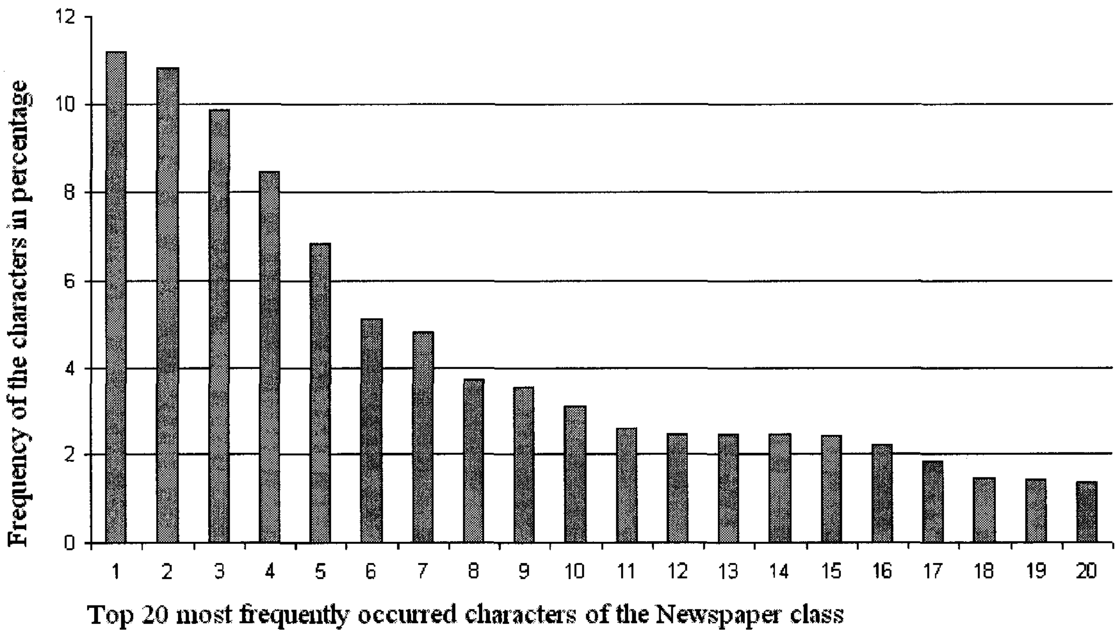

Figure 8: Histogram of the top 20 characters of newspaper class.

When we compare our findings for the newspaper class with the literature class, we find that among the top 20 characters 19 are the same, the different ones are "ছ" for the Literature class and "য" for the newspaper class. Other differences are অ, আ, and এ which are the top three most frequently occurred characters, they change their positions.

In literature class আ and এ were top two characters, the main reason of this being the use of more casual words and pronouns being used in the literatures such as আমি, আমার, আমরা, তুমি, তোমার, তোমরা, তারা, তাদের, তাহার, হয়েছে, করেছে, গিয়েছে, পেয়েছে, etc. All of these words have these two characters আ and এ occurring in abundance, whereas অ occurring very less here. One more difference in the newspaper class is that here many more words from other languages are used, and also full names of local and foreign people and various organizations. Some foreign words that we found to occur frequently are Cricket, Coach, One-day, Test, Security, Police, Computer, Windows, Internet, Microsoft, Copy, Service, Tool, TV, Audio, Video, Class, etc.

Now H-maximum refers to the theoretical maximum entropy, this being achieved only when the letters of the alphabet are all equiprobable [15]. From eq^n (3),

$$\text{H\_maximum} = \log_2 46$$

$$= 5.5236$$

From our calculation, the actual entropy or 1st order entropy that we got for newspaper class is:

$$\text{H\_actual} = 4.5454$$

Hence, from eq^n (4) and (6), for literature class:

$$\text{H\_relative} = 4.5454 / 5.5236$$

$$= 0.82291$$

$$\text{Redundancy} = 1 - 0.82291$$

$$= 0.17709$$

So, the relative uncertainty is 82.29% for newspaper texts and the redundancy is 17.71%. We will compare it later with the redundancy found for other classes.

4.3.3. Religious Class

We have subdivided the religious data into two classes, Islamic and Hindu religious texts class. 4 most frequently occurred vowels and consonants and their frequency rates for Hindu and Islamic texts are shown in Tables 11, 12 and 13, 14 respectively.

Letter	Occurrence	Total char	Frequency (%)
অ	1619	10107	16.02
আ	885	10107	8.76
এ	709	10107	7.01
ই	478	10107	4.73

Table 11: 4 Most Frequently Occurred Vowels of Hindu Text and their frequency rates

Letter	Occurrence	Total char	Frequency (%)
র	851	10107	8.42
ব	505	10107	5.00
ন	501	10107	4.96
ত	447	10107	4.42

Table 12: 4 Most Frequently Occurred Consonants of Hindu Text and Their Frequency Rates

Letter	Occurrence	Total char	Frequency (%)
আ	1318	10381	12.70
অ	1276	10381	12.29
এ	914	10381	8.80
ই	586	10381	5.64

Table 13: 4 Most Frequently Occurred Vowels Of Islamic Texts and Their Frequency Rates

Letter	Occurrence	Total char	Frequency (%)
র	924	10381	8.90
ন	516	10381	4.97
ত	477	10381	4.59
ক	396	10381	3.81

Table 14: 4 Most Frequently Occurred Consonants of Islamic Texts and Their Frequency Rates

Among top 20 most frequently occurred characters 18 are the same ones, and 4 characters are different taking two from each class. These 18 characters are plotted in a graph (Fig. 9) in the following order (ordered in terms of their frequency in Islamic texts):

আ, অ, র, এ, ই, ন, ত, ক, ব, ম, দ, ল, স, হ, য, উ, প, ঈ.

The different ones are ও and য় for Islamic texts and শ and গ for Hindu texts (Table 14). Actually these characters can also be found in the chart if we just take 25 top chars into account.

And if we have a look at the comparative frequencies of the characters in two classes we find that except 2 characters (char 1 and 2, namely আ, অ) almost all of the rest of them occurs at close frequencies.

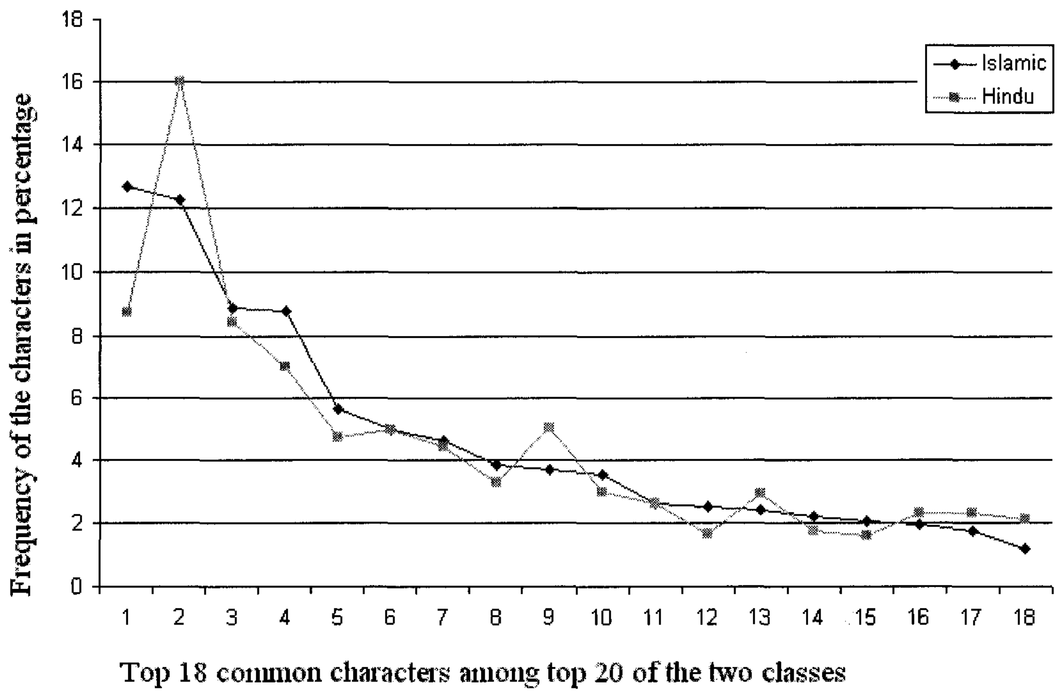

Figure 9: Graph depicting comparative frequency of the top 18 characters of two classes.

A more detailed table containing the frequency of occurrences (in percentage) of the characters that are common in the top 20 list of the Hindu and Islamic religion class are shown in table 15.

Islamic Text	
Letter	Frequency (%)
আ	12.70
অ	12.29
র	8.90
এ	8.80
ব	5.64
ন	4.97
ই	4.59
ত	3.81
ক	3.68
ম	3.53
স	2.62
দ	2.62
উ	2.50
প	2.38
ঈ	2.18
শ	2.04
হ	1.96
ল	1.72
গ	1.68
য	1.18

Hindu Text	
Letter	Frequency (%)
অ	16.02
আ	8.76
র	8.42
এ	7.01
ই	5.00
ন	4.96
ত	4.73
ক	4.42
ব	3.27
ম	2.95
ও	2.90
দ	2.59
ল	2.31
স	2.31
হ	2.08
য	1.81
উ	1.75
প	1.64
য়	1.60
ঈ	1.59

Table 15: Top 20 Most Frequently Occurred Character and Their Frequency (%) of religious texts

আ and অ are top 2 most frequently occured characters in both the classes, but there is a significant difference between the occurence of them in two classes. For the Islami ones their frequency are very close to each other, but in Hindu texts অ is much higher, in fact about double of the second frequent character. Another interesting aspect is that the total frequency of these two characters are about same for both classes (24.99% for Islami texts and 24.78% for Hindu texts). So, from this observation we can colclude that the character অ occurs more frequently in Hindu religious texts compared to Islamic ones, but at the same time frequency of আ decreases.

	Number of occurrences	
Letter	Islamic texts	Hindu texts
অ	1276	1619
আ	1318	885
ঈ	123	210
ঋ	13	61
এ	914	709
ও	272	148
ঐ	5	24
গ	95	162
ঔ	5	52
ফ	23	4
ব	382	505
ভ	49	114
ল	260	166

Table 16: Characters with notable differences in occurrence (between the Islamic and Hindu text classes).

After studying the character frequency table we found some interesting characteristics of the texts of two religious classes. The notable differences of the occurrence of characters can be seen in Table 16.

First thing to notice here is the occurrence of অ and আ in the two classes of texts. One of the main reasons of অ occurring more frequently and আ occurring less in Hindu texts than the Islamic one is the usage of Sanskrit words in abundance in Hindu religious texts. The Sanskrit words and also many Hindu religious words have a tendency of using of অ more often after each consonant and especially at the end position of each word. And as has been stated previously, it is evident from study that in Hindu texts whereas the number of occurrences of অ increases, occurrences of আ are found in decreased numbers than the regular usage of the two letters.

Then, the letter ঈ occured more frequently in Hindu texts. One significant reason of this is the usage of a certain word শ্রী, which is used to address anyone respectfully and is used many times in the Hindu religious books.

There is no specific Islamic word usage that clarifies the occurrence of এ and ও more times in Islamic texts, but we can conclude that the general Bengali words containing these characters are more used in these texts, like তোমাকে, তোমরা, বলো, থেকে, গেছে, etc.

Three letters ঋ, ঐ and ঞ occur much less not only in Islami texts, but in fact these letters or words containing these lettesr are very seldom used in Bengali language. The reason of them happening more frequently in Hindu texts is due to some specific words happening lot of times in Hindu texts, such as কৃষ্ণ, বৈষ্ণব, বিষ্ণু, বৃন্দ, অদ্বৈত, চৈতন, বৃন্দাবন, etc.

ফ is another character that is very less used in words of Bengali language, the reason of happening of this character more in Islami texts is due to the usage of some specific religious words found in them, like ফিতর, ফিকর, মারূফ, etc.

আললাহ and ভগবান ("Allah" & "Bhagaban") are the two specific words that means God in Islamic and Hindu religion respectively. As a result these two words are found more frequently when these religion's texts are studied. As these two words happen more frequently in their respective texts, some characters found in them are also found in increased numbers in their occurrence table. Like ল in Islami texts and ভ, গ and ব in Hindu texts.

Two foreign languages have great influences on these two classes of religious Bengali language texts, Arabic language on Islami texts and Sanskrit on Hindu texts to be specific. Many words from these languages are used without any alteration in Bengali language. Some of these words have now become a part of Bengali language and no other words can be found in the language to represent the same meaning.

Islamic religious words	আললাহ, ঈমান, রসূল, নবী, নামায, জিহাদ, জাননাত, মুমিন, উমমত, হাদীস, আয়াত, এবাদত, ঈদ, রোজা, ঈদ-উল-ফিতর, রমজান, যাকাত, আদব, etc.
Hindu religious words	ভগবান, কৃষঞ, বিষঞু, রাম, দূর্গা, সরসবতী, পূজা, দেবী, ধর্ম, বেদ, শ্রী, গীতা, মনত্র, ঠাকুর, পরসাদ, মূর্তি, পরণাম, অদবৈত, বৃনদাবন, চৈতন, বৈষঞব, মহামায়া, শ্রীমদভাগবদগীতা, etc

Table 17: Commonly used Islamic and Hindu religious words

We calculated the Spearman's rank correlation coefficient (nonparametric test) in order to measure the correlation between the ranks of Islamic and Hindu text characters.

We set our hypothesis as:

H_0: ρ=0 (no correlation between character ranks)

H_a: ρ>0 (positive correlation between character ranks)

Spearman's rank correlation is calculated as,

$$\rho = 1 - \frac{6\sum_{i=1}^{n}[R(\mathrm{I}_i) - R(\mathrm{H}_i)]^2}{n(n^2 - 1)}$$

where,

ρ	Spearman's rank correlation coefficient
$R(\mathrm{I}_i)$	Ranks of Islamic characters
$R(\mathrm{H}_i)$	Ranks of Hindu characters
n	Number of characters ($n=46$)

We found that the characters ranks of the Islamic texts are positive correlated ($\rho=0.995$) with the character ranks of the Hindu texts. The results are significant at the 1% level. Based on our sample, we conclude that the use of characters is identical in both languages.

Now H-maximum refers to the theoretical maximum entropy, from eq^n (3),

$$H\_maximum = \log_2 46 = 5.5236$$

From our calculation, the actual entropy or 1st order entropy that we got for each class, and the relative entropy and redundancy are as follows:

	Islamic texts	Hindu texts
H-maximum	5.5236	5.5236
H-actual	4.5000	4.4000
H-relative	0.8147	0.7966
Redundancy	0.1853	0.2034

Table 18: Entropy and redundancy of printed religious Bengali texts.

Therefore, the average number of bits per letter required to translate the language into binary are 4.5 and 4.4 for Islamic and Hindu texts respectively.

The relative uncertainty in this ensemble is 81.47 for Islamic texts and 79.66 for Hindu texts. Therefore, the redundancy is 18.53% for Islamic texts and 20.34% for Hindu texts.

4.3.3. Total sample

The total data samples collected for all classes contain 50247 characters. 4 most frequently occurring vowels and consonants, and their frequency rates for total samples are shown in Tables 19 and 20 respectively.

Letter	Occurrence	Total char	Frequency (%)
অ	6092	50247	12.12
আ	5652	50247	11.25
এ	4699	50247	9.35
ই	3054	50247	6.08

Table 19: 4 most frequently occurred vowels of total sample texts and their frequency rates

Letter	Occurrence	Total char	Frequency (%)
র	4058	50247	8.08
ন	2419	50247	4.81
ক	2192	50247	4.36
ব	1995	50247	3.97

Table 20: 4 most frequently occurred consonants of total sample texts and their frequency rates

A detailed table containing the number of occurrence and frequency of top 20 characters are shown in Table 21:

Letter	Letter type	Occurrence	Total char	Frequency (%)
অ	Vowel	6092	50247	12.12
আ	Vowel	5652	50247	11.25
এ	Vowel	4699	50247	9.35
র	Consonant	4058	50247	8.08
ই	Vowel	3054	50247	6.08
ন	Consonant	2419	50247	4.81
ক	Consonant	2192	50247	4.36
ব	Consonant	1995	50247	3.97
ত	Consonant	1979	50247	3.94
ম	Consonant	1541	50247	3.07
স	Consonant	1439	50247	2.86
ল	Consonant	1285	50247	2.56
উ	Vowel	1191	50247	2.37
প	Consonant	1110	50247	2.21
ও	Vowel	1065	50247	2.12
দ	Consonant	1066	50247	2.12
য়	Consonant	976	50247	1.94
হ	Consonant	853	50247	1.70
য	Consonant	778	50247	1.55
গ	Consonant	610	50247	1.21

Table 21: Top 20 most frequently occurred characters and their frequency (%) of total sample texts.

As evident from Table 21, vowels occupy 4 of the top 5 most frequently occurred characters list. Of all the characters, vowels constitute 45.01% of the total data sample and consonants constitute 54.99%. And the top 20 characters constitute 87.67% of the total sample.

When we compare the top 20 most frequently occurred characters of total sample with the separate literature, newspaper, Islamic and Hindu religion classes, we found 15 characters are same in all the top 20 table, they are:

অ, আ, এ, র, ই, ন, ক, ব, ত, ম, স, ল, উ, প, হ

The uncommon ones are ও, দ, য়, য, গ, হ, ট, ছ, ঈ, শ.

So, in all top 20 tables we have actually 25 characters. It happens actually for specific characteristics of the words used in a class, which has been discussed in the respective sections before.

From the findings of total data sample we have the classified the characters into following groups according to their frequency of occurrence:

Group	Characters	Frequency (%)
Vowels	অ, আ, ই, ঈ, উ, ঊ, ঋ, এ, ঐ, ও, ঔ	45.01
High-Frequency Consonants	র, ন, ক, ব, ত, ম, স, ল, প, দ	37.98
Medium-Frequency Consonants	য়, হ, য, গ, শ, ট, ছ, জ, চ, থ	12.37
Low-Frequency Consonants	ষ, ভ, ধ, খ, ণ, ড়, ড, ঠ, ফ, ঞ, ঘ, ঙ, ঝ, ঢ, ঢ়	4.64

Table 22: Classification of characters according to their frequency of occurrences

Now we have the maximum entropy, from eq^n (3),

$$H\_maximum = \log_2 46$$

$$= 5.5236$$

From our calculation, the actual entropy or 1st order entropy that we got for total data is:

$$H\_actual = 4.5524$$

Hence, from eq^n (4) and (6), for literature class:

$$H\_relative = 4.5524 / 5.5236$$

$$= 0.82417$$

$$Redundancy = 1 - 0.82417$$

$$= 0.17583$$

The actual entropy, relative entropy (or uncertainty) and redundancy of all the classes are shown in the following table:

Class	Maximum Entropy	Actual Entropy	Relative Entropy	Redundancy
Literature	5.5236	4.5232	0.8189	18.11 %
Newspaper	5.5236	4.5454	0.8229	17.71 %
Islamic	5.5236	4.5000	0.8147	18.53 %
Hindu	5.5236	4.4000	0.7966	20.34 %
Total	5.5236	4.5524	0.8242	17.58 %

Table 23: Entropy calculation results for total and all the separate classes

From Table 22, we can conclude that the texts of the Hindu class are more redundant, that means they are more guessable than the other classes. This result is statistically significant. From Shannon's calculation of Entropy of English (eq^n 5), we have the redundancy of English language as,

$$\text{Redundancy} = 1 - (4.14/4.7) = 1 - 0.88085 = 0.11915 \text{ or } 11.91\%$$

So, we can find that Bengali language is more redundant compared to the English language. This result is also statistically significant. In fact, Redundancy in a language is actually useful at times, for how else we can discern what is said in a noisy room? The redundancy allows one to understand what is said when only part of a message is available or comes across. But, on information processing point of view, this redundant data is an overhead. This redundant data should be compressed in a way to ensure better storage and transmission of data and save storage space and processing time.

Next, we can have a look at the zero and first order entropies of different languages with the one we calculated for Bengali language in Table 24:

Language	F_0	F_1
English[2]	4.7	4.14
German[22]	4.7	4.1
Russian[4]	5.00	4.35
Arabic[3]	5.00	4.2
Bengali	5.52	4.55

Table 24: Zero and First order entropies of different languages

For the zero-order entropy, we know that it is the maximum entropy of that language and the entropy is at a maximum when all of the states have equal probabilities. And this maximum entropy increases with the increase of number of states (here the number of characters in alphabet). For English language, the one used in the table was originally calculated by Shannon with 26 characters of English alphabet. When spaces are also taken into account the maximum entropy increases to 4.76. As Bengali language alphabet has 46 characters, this explains the much bigger maximum entropy for the language. And from the first-order entropy Bengali language, we can conclude that the average number of bits per letter required transmitting or store Bengali language is 4.55 bits. We haven't tested the statistical significance of the difference between first-order entropy of English and Bengali.

4.3.4 Consonant Bigrams

Bigrams are groups of two successive written letters or two symbols and are very commonly used as the basis for simple statistical analysis of text. They are used in one of the most successful language models for speech recognition [36].

There are 35 consonants in Bengali language alphabet, and when Bigrams of all consonants (all Bigrams which consists only consonants) are taken into account they make 1225 different Bigrams. We have calculated all of them, and from our total data sample of 50247 characters we have found 3652 occurrences of consonant Bigrams. Among the 1225 probable Bigrams only 266 different bigrams are found to occur at least once. The top 50 Bigrams, their occurrences and frequency of occurrence are shown in table 24. Of the 1225 bigrams, these top 50 most frequently occurred ones constitute 68.45% of the total occurrences of Bigrams.

Bigram	Occurrence	Frequency (%)
পর	192	5.26
সত	105	2.88
নত	104	2.85
রব	102	2.79
কষ	100	2.74
তর	94	2.57
নদ	90	2.46
লল	80	2.19
নয	78	2.14
বয	73	2.00
রত	73	2.00
ঙগ	68	0.68
রম	66	1.81
কত	65	1.78
কট	64	1.75
চছ	63	1.73
ষট	59	1.62
সব	54	1.48
তয	53	1.45
শর	49	1.34
রক	47	1.29
সথ	45	1.23
দধ	41	1.12
মব	39	1.07
মর	38	1.04

Bigram	Occurrence	Frequency (%)
শব	38	1.04
তত	36	0.99
ধয	36	0.99
গর	35	0.96
দর	34	0.93
রয	34	0.93
দব	33	0.90
তব	31	0.85
রণ	31	0.85
কর	30	0.82
সট	29	0.79
ষঞ	29	0.79
দয	27	0.74
সক	26	0.71
গঞ	25	0.68
রদ	24	0.66
লত	24	0.66
লপ	24	0.66
নধ	23	0.63
নন	23	0.63
রথ	23	0.63
শয	22	0.60
টর	21	0.58
বর	21	0.58
ষঠ	21	0.58

Table 25: Top 50 Consonant Bigrams, their occurrence and frequency of occurrence.

A statistics of number of Bigrams starting with a specific consonant (having at least one occurrence), their occurrences and frequency of occurrence are shown in Table 26:

Letter	No. of Bigrams	Occurrence	Frequency (%)
ক	19	353	9.67
খ	06	46	1.26
গ	12	102	2.79
ঘ	03	03	0.08
ঙ	03	70	1.92
চ	09	80	2.19
ছ	01	01	0.03
জ	10	37	1.01
ঝ	03	05	0.14
ঞ	03	29	0.79
ট	10	54	1.48
ঠ	02	08	0.22
ড	04	07	0.19
ণ	07	36	0.99
ত	07	231	6.33
থ	04	08	0.22
দ	11	157	4.30
ধ	06	49	1.34
ন	19	421	11.53
প	10	235	6.43
ফ	03	06	0.16
ব	14	138	3.78
ভ	03	19	0.52
ম	20	200	5.48
য	01	02	0.05
র	26	573	15.69
ল	16	186	5.09
শ	11	138	3.78
ষ	11	144	3.94
স	12	314	8.60

Table 26: Bigrams starting with a specific consonant, their occurrences and frequency of occurrence

From Table 26 we find that ক, ত, ন, প, ম, র and স are the top 7 consonants that make most of the Bigrams that have at least one occurrence. Together they have 113 different Bigrams occurring at least once, and have a total of 2327 occurrences, which is the 63.73% of the total occurrence.

An n-gram is a sub-sequence of n items from a given sequence. The items can be letters, words or base pairs according to the application. An n-gram of size 1 is a "unigram"; size 2 is a "bigram" (also called a "digram"); size 3 is a "trigram"; and size 4 or more is simply called an "n-gram". For, 46 characters of Bengali alphabet bigram gives 2116 possible character combinations, trigram gives 97336 combinations and n-gram gives 4477456 combinations, whose frequency count has first to be calculated in order to proceed with further entropy related calculation for all these combinations. It wasn't possible to do this kind of vast amount of calculations within our limited scope and time constraints. Automated software should be built to carry out this kind of huge calculation and that will further enhance research in this field.

Chapter 5

Conclusion & Future Directions

5.1 Conclusion

Natural Language is a very important aspect of human life and culture, and researches addressing natural language is not only helpful in information processing of that language but also helps researchers and others understand the language better and provide results that enables them to have a deep insight into the language.

From our research, we have found the Zero and first-order entropy of Bengali language to be 5.52 and 4.55 respectively. And the language uncertainty and redundancy are 0.8242 and 17.58% respectively. We have also studied three different data classes, namely Literature, Newspaper and Religious class in order to come up with specific entropy for each of them and analyzed their characteristics. There are no studies in the literature to compare our results with, in regarding to Bengali language, which makes our results as a new reference for researchers in this field. But, our results show that the entropy and redundancy values of are larger compared to other languages, such as English (Table 23). This should mean that there is scope for researchers to alleviate this redundancy. We hope that these results will help the researchers in future to further investigate in this field and find better compression methods to alleviate the redundancy of the language.

5.2 Future Directions

There are many probable directions of extending the current research work presented in this thesis. Some possible directions of future work are outlined below:

- We haven't included spaces and some modifier symbols that are character look-alike in our alphabet count, further research can be done including these in total character count.
- A very large number of text databases can be made to further investigate the entropy of Bengali language, which was not possible for us due to time constraints. Moreover, higher order entropies should be calculated and compare these results with other languages.
- We have calculated only consonant Bigrams; research can be done to find out the Bigrams of all the characters. The total number of Bigrams in that case will be 2116, and when vowels are included there will be a lot more occurrences to work with.
- As the redundancy of the language is found out, research can be done to find out how to alleviate this redundancy for a better storage and transmission of Bengali language.
- Due to the problems mentioned in section 4.2, we couldn't use the Unicode system. This problem and the unavailability of a corpus restricted us from working with more extensive data. An automated program or software should be built to solve this problem and ensure future research with more extensive data.

Bibliography

[1] Hamid Moradi, Jerzy W. Grzymala-Busse and James A. Roberts, *"Entropy of English text: experiments with humans and a machine learning system based on rough sets"*, Information Sciences: an International Journal archive, Vol. 104, Issue 1-2 (Jan, 1998), pp. 31-47.

[2] C. E. Shannon, *"Prediction and entropy of printed English"*, Bell Sys. Tech. J. 30, (1951), pp. 50–64.

[3] M. A. Wanas, A. I. Zayed, M. M. Shaker and E. H. Taha, *"First, second and third-order entropies of Arabic text"*, IEEE transactions, Information theory, IT-22, 1978.

[4] V. A. Garmash, N. E. Kirillov and D. S. Lebedev, *"An experimental study of statistical properties of message sources"*, Problems of Information Transmission, Issue 5, Moscow.

[5] *"Bangalipedia: Bengali Language"*, http://Bengalipedia.search.com.bd/HT/B_0137.htm, Nov 2007.

[6] Paul A. Booth, *"An Introduction to Human-Computer Interaction"*, Psychology press, 1989.

[7] *"The Entropy of a Language", http://www.everything2.com/index.pl?node_id=14845,* Nov 2007.

[8] *"English language letter frequencies"*,

http://www.everything2.com/index.pl?node_id=14845, Nov 2007.

[9] C. E. Shannon, "*A Mathematical theory of communications*", Bell Syst. Tech. J. 27, pp. 379-423 (1948)

[10] Rafael C. Gonzalez and Richard E. Woods, "*Digital image processing*", Addison-Wesley, 1992.

[11] Anders Gjendemsjo and Behnaam Aazhang, "*Entropy*",

http://cnx.org/content/m11839/latest/, Nov 2007.

[12] David Salomon, "*Data Compression: The Complete Reference (3rd Ed.)*", Springer-Verlag New York, Inc., Secaucus, NJ, 2004.

[13] Silviu Guiasu and Abe Shenitzer, *"The principle of maximum entropy"*, The Mathematical Intelligencer, 42-48, 1985.

[14] E.T. Jaynes, *"Notes on present status and future prospects"*, In: W.T. Grandy and L.H. Schick Jr., Editors, *Maximum Entropy and Bayesian Methods*, Kluwer, Dordrecht, The Netherlands (1990), pp. 1–13.

[15] E.J. Yannakoudakis and G. Angelidakis, "*An insight into the entropy and redundancy of the English dictionary*", Transactions on Pattern Analysis and Machine Intelligence, Vol. 10, Issue 6, Nov 1988, pp. 960-970.

[16] G. K. Zipf, "*Human Behavior and the Principle of Least Effort*", Addison-Wesley Ltd., 1949.

[17] M. Grignetti, "*A note on the entropy of words in printed English,*" Inform. Contr., Vol. 7, pp. 304-306, 1964.

[18] A. Treisman, "*Verbal responses and contextual constraints in language*", Journal of Verbal Learning Behaviour 4, pp. 118-128, 1965.

[19] H. E. White, "*Printed English compression by dictionary encoding*, "Proc. IEEE, vol. 55, no. 3, pp. 390-396, Mar. 1967.

[20] D. Jamison and K. Jamison, "*A note on the entropy of partially known languages*," Inform. Contr., vol, 12, pp. 164-167, 1968.

[21] G. A. Miller and J. A. Selfridge, "*Verbal context and the recall of meaningful material*", Amer. J. Psych., Vol. 63, pp. 176-185, 1950.

[22] K Kepfmueller, *"The entropy of the German Language"*, Fernmeldetechnische Zeitschrift, (J. Telecommunication) VII, pp. 265-272.

[23] P Grassberg, "*Estimating the information content of symbol sequences and efficient codes*", IEEE Transaction, Information Theory, IT-24, 413-421, 1978.

[24] Lev B. Levitin and Zeev Reingold, "*Entropy of natural languages: Theory and experiment*", Chaos, Solitons & Fractals, Vol. 4, Issue 5, May 1994, pp. 709-743.

[25] "*Wikipedia: Bengali language*", http://en.wikipedia.org/wiki/Bengali_language, Nov 2007.

[26] *Bhattacharya, T, "Bangla (Bengali)", in* Gary, J. and Rubino. C., *"Encyclopedia of World's Languages: Past and Present", WW Wilson, New York, 2000.*

[27] "*Omniglot: Bengali alphabet, pronunciation and language*", http://www.omniglot.com/writing/bengali.htm, Nov 2007.

[28] Banaphul, "*Banaphuler Chhotogolpo Samagra*", Banishilpo Prokashoni, Kolkata, January 2003.

[29] Rabindranath Tagore, "*Detective*", short story from Book: "*Galpaguchchha",* 1964, Visva-Bharati, Shantiniketan, Kolkata.

[30] Satyajit Ray, "*Anukul*", short story from Book: "*Aaro Baro*", Ed. 1, 1981, Ananda Publishers, Kolkata

[31] Humayun Ahmed, "*Debi*", Abasar Publishers, Dhaka, June, 2004.

[32] Humayun Ahmed, "*Elebele*", Samay Publishers, Dhaka, January, 1990.

[33] Anisul Haque, "*Maa*", Samay Publishers, Dhaka, February, 2003.

[34] Shirshendu Mukhopadhyay, "*Sanket*", Purnima, Eid-ul-Fitr edition, 2006.

[35] Humayun Ahmed, "*Krishnapaksha*", Mawla Brothers Publishers, January, 2006.

[36] Michael Collins, "*A new statistical parser based on bigram lexical dependencies*", Proceedings of the 34th Annual Meeting of the Association of Computational Linguistics, Santa Cruz, CA., 1996, pp.184-191.

[37] Shahjahan Munir, "*Bangla Byakoron*" 7th ed, 1989, Students Publications, Bangladesh

[38] Haralaal Ray, "*Byakoron o Rochona*", Ed. 11, Puthighar Publications Ltd, 2002.

[39] "*Banglapedia* (Bangladesh National Encyclopedia)", Bangladesh Asiatic Society, March 2003.

[40] V. Van de Laar, B. Kleijn and E. Deprettere, "*Perceptual entropy rate estimates for the phonemes of American English*", Proceedings of the 1997 IEEE International

Conference on Acoustics, Speech, and Signal Processing (ICASSP '97), Vol. 3, 1997, pp. 1719-1722

[41] R. Merkyte, "*The information content of the Lithuanian language*", Lithuanian Mathematical Journal, Vol. 18, No. 3, July 1978, pp. 384-389.

[42] "*Bengali language and script*", http://www.isical.ac.in/~rc_Bengali/Bengali.html, Nov 2007.

[43] "*Bengali language: A brief introduction*", http://www.i3pep.org/archives/2003/11/03/bengali-language-a-brief-introduction/, Nov 2007.

[44] "*Introduction of Bengali language*", http://www.Bengali-online.info/BengaliLanguage/IntroductionOfBengaliLanguage.htm, Nov 2007.

[45] "*Bengali: The official language of Bengalidesh*", http://www.betelco.com/bd/Bengali/Bengali.html, Nov 2007.

[46] Dmitriy Genzel and Eugene Charniak, "*Entropy rate constancy in text*", Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 199-206.

[47] Zhihui Jin and Kumiko Tanaka-Ishii, "*Unsupervised segmentation of Chinese text by use of branching entropy*", Proceedings of the COLING/ACL 2006 Main conference poster sessions, Sydney, Australia, July 2006, pp. 428-435

[48] Adam L. Berger, Vincent J. Della Pietra and Stephen A. Della Pietra, "*A maximum entropy approach to natural language processing*", Computational Linguistics, Vol. 22, Issue 1, March 1996, pp. 39-71.

[49] Anil K. Jain, "*Fundamentals of Digital Image Processing*", Prentice – Hall, Englewood Cliffs, N.J., 1989.

[50] W. J. Teahan and J. G. Cleary, "*The entropy of English using PPM-based models*", Proceedings of the Conference on Data Compression, 1996, pp. 53-62

[51] Jan Hajič, "*Morphological tagging: data vs. dictionaries*", Proceedings of the first conference on North American chapter of the Association for Computational Linguistics, Seattle, Washington, 2000, pp. 94-101.

[52] Wong Ping Wai and Yang Yongsheng, "*A maximum entropy approach to HowNet-based Chinese word sense disambiguation*", International Conference On Computational Linguistics COLING-02, 2002, pp. 1-7.

[53] Victoria Fossum, "*Entropy, Compression, and Information Content*", http://www.eecs.umich.edu/~vfossum/pubs/entropy_explanation.pdf

[54] Debra A. Lelewer and Daniel S. Hirschberg, "*Data Compression*", ACM Computing Surveys (CSUR), Vol. 19, No. 3, September 1987, pp. 261-296

[55] Timothy Bell, Ian H. Witten and John G. Cleary, "*Modeling for text compression*", ACM Computing Surveys (CSUR), Vol. 21, No. 4, December 1989, pp. 557-591

[56] "*Bigram Counts and Association Statistics*",

http://www.umiacs.umd.edu/~resnik/nlstat_tutorial_summer1998/Lab_ngrams.html

[57] Qiuping A Wang, "*Probability distribution and entropy as a measure of uncertainty*", Journal of Physics A: Mathematical and. Theoretical 41, 2008, 8pp.