Corpus-based empirical analysis of form, function and frequency of characters used in Bangla

Niladri Sekhar Dash and Bidyut Baran Chaudhuri
Computer Vision and Pattern Recognition Unit
Indian Statistical Institute
203, Barrackpore Trunk Road
Calcutta - 700 035, India
Email: {niladri/bbc@isical.ac.in}

Abstract

In this paper an attempt is made to understand formal and functional aspects of Bangla characters used in the written texts compiled in a sample monitor corpus designed systematically from language data collected from various text documents published within 1980 and 1995. The purpose of this study is to understand the form and function of the characters, trace their behavioural peculiarities, and if possible, find out the reasons of such peculiarities. The study focuses on the formation of the characters, their structural change in case of compound and cluster formation, their contextual use, statistical analysis of their occurrence, and their position in words. The study also encompasses the use of different punctuation marks in the texts. Finally, some possible areas of application of such analysis are identified.

Key words: character, corpus, grapheme, allograph, vowel, consonant, cluster, frequency, statistics, punctuation.

1. Introduction

The Bangla[1] language corpus, used for the following empirical analysis, is represented by a set of unique orthographic symbols (letters) arranged in a specific pattern with appropriate punctuation marks for proper reading and comprehension of the language users. Most of the linguistic features of the language are captured by the character[2] set that leads the written text to be potentially representative of the spoken form of the language. The study of the characters constituting the corpus is important for accounting their pattern of use in different context of the texts as well for comprehension of the general characteristics of the language. Thus, multi-layered information of the characters can be important and necessary contribution to Natural Language Processing (NLP), Computational Linguistics (CL), Optical Character Recognition (OCR), key-board design, Word Sense Disambiguation (WSD), Parts-of-speech Tagging, cryptography, language teaching, Machine Translation (MT), besides other applied and interdisciplinary studies. Moreover, it can provide insight about how language is used by different users in different domains of knowledge representation.

For functional and structural analysis of characters used in Bangla, we have considered a written corpus of 4 million words comprising printed texts of nearly 85 disciplines, published within 1980 and 1995 (Dash and Chaudhuri 2000). To make our corpus non-skewed and sufficiently representative, we have tried to develop it after the standards proposed by Sinclair (1991), Atkins *at al.* (1992), Biber (1993) and others. The relative frequency of various published documents in the corpus is determined following the report of a survey conducted by the Central Institute of Indian Languages (CIIL), Mysore, as stated below:

[1] Among the languages used in the world today, Bangla is fifth (after English, Chinese, Spansh and Hindi) in strngth in ragard to the number of speakers. It is the national language of Bangladesh and one of the national languages of India, spoken by the people of West Bengal and Tripura, two states of India. Moreover, a sizable number of people living in Asia, Europe and America also use this language.

[2] The term *character* is used here in linguistic or paleographic sense to mean different orthographic symbols used for writing of a natural language. In most cases, it includes letters, punctuation marks and other similar symbols.

Texts from genre	%-age	Texts from genre	%-age
Mass media	30 %	Literature	15 %
Social science	15 %	Natural science	15 %
Commerce	10 %	Fine arts	5 %
Translation	5 %	Others	5 %

Table 1: Percentage of texts belonging to different genres included in the corpus

Moreover, an electronic Bangla dictionary of around 60,000 words is used for the current purpose. Thus, the corpus and the dictionary are used as representative linguistic data-base for description as well as for verification of hypotheses about the use of characters in the language. They provide information of use, shape, design, size, occurrence and change of the characters as well as insight into the role of punctuation marks employed in writing.

The paper is organised as follows: section 2 highlights some unique features of Bangla characters while section 3 gives an account of shape analysis of characters in isolation as well as within words. It also gives some idea of tier division, compounding and clustering of characters. In section 4, we provide a few frequency counts of characters in the corpus while in section 5, we try to provide a brief survey of the use of punctuation marks. The importance of such study is evaluated and some application areas are identified in section 6.

2. Bangla characters

In printed Bangla texts, five types of character are noted:

(i) vowel grapheme
(iii) consonant grapheme
(ii) vowel allograph[3]
(iii) consonant graphic variant
(v) graphic compound, and
(vi) consonant cluster.

Scholars like Banerji (1919), Chakravarty (1938), Diringer (1953), Sen (1993) and others more or less agreed that Bangla characters are developed from the proto-Bangla type. The present character set is claimed to be evolved from the hand-written letters inscribed in various Old Bangla documents and manuscripts (Sen 1993: 24). But, through the years, notable modification on the shape of the characters is introduced at different points of time. By the 15th and 16th century, the characters appeared to be fully developed. Indeed, during the 17th and 18th century, no change is registered on their design and structure. In the 19th century, the shape of the characters has been stereotyped by the introduction of printing press (Diringer 1953: 365). When the character set was mechanically designed and stratified for printing, the scope of structural modification was drastically reduced. However, that the first character set designed by Wilkinson in the 18th century, has been modified to give more recent shape (Bandyopadhay 1981: 47). Vidyasagar made an attempt to rearrange characters more scientifically and proposed 52 characters (12 vowels and 40 consonants) in place of earlier 50 characters (16 vowels and 34 consonants) (Bandyopadhaya 1981: 100). As a result, the total number of character has been reduced and some consonant clusters are simplified in form (Mukhopadhaya 1981: 102). It is noted that all Bangla vowels, except *a*, have allographs. Probably, allographs were meant to speed up handwriting, because basic vowels, compared to their respective allographs, take more time, space and energy. It is statistically observed that most recurrently used allograph is most simplified in shape and most suitably positioned in writing system (Dash and Chaudhuri 2000).

Structurally, Bangla characters are cursive and twisted as Diringer (1953: 363) observes:

[3] The term *allograph* is a cover term that includes all graphic variants found in the script. The variants are generally called *vowel signs*. Recently, Babulanam and Beena (1999) have mentioned them as *shorthand signs* which, however, we ignore for possibility of confusion with real shorthand signs.

> "The Bengali was a peculiar cursive script with circular or semi-circular signs, hooks or hollow triangles attached to the left of the tops of the vertical strokes. The triangle itself is a modification of the top stroke with a semi-circle below, and this form is connected with the common form of thick top-strokes, rounded off at both ends".

Ploughing through the corpus, the following features of Bangla character set are noted:

(i) It has 12 vowels, 20 allographs and 39 consonants along with nearly 280 consonant clusters which are read and written from left to right sequence in word formation.

(ii) In printed texts, the vowel ***e*** has two allographs: one is with and the other is without the ’head line’ (a short horizontal line put above most basic characters). The contexts of their use are also different.

(iii) A consonant or cluster can use only one allograph at a time.

(iv) Unlike vowel, in certain contexts, a consonant may be silent in utterance despite being physically present in the text.

(v) The shape of an allograph is not grapheme dependent. However, in some cases it is changed depending on the shape of a consonant. For instance, the shape of allograph of the vowel *u* is changed when used with consonant *g* and *sh*, and cluster *nt* etc.

(vi) The linear position of allographs with respect to graphemes is not uniform. They can occur before or after, above or below the consonants or clusters.

(vii) Consonants also register some graphic variants (called *reph* and *raphalaa*, *yaphalaa*, *vaphalaa* etc.) which are generally used in cluster formation.

(viii) A single grapheme most often represents a single phoneme with a slight phonetic variation (e.g. *ii* (long) and *i* (short) denote [i], *u* (short) and *uu* (long) denote [u], *j* and *y* denote [ɟ], *sh* (palatal), *s.* (retroflex) and *s* (dental) denote [ʃ], *n* (alveolar) and *n* (dental) represent [n] etc.)

(ix) Most of the consonants can join physically to form a cluster. Generally, clusters are formed by joining two consonants. However, clusters of three or four consonants are also possible. There are nearly 280 clusters of which those made by two consonants are around 240. Cluster of three and four consonants are nearly 35 and 5, respectively.

(x) The sentence terminal markers consist of *purnacched* (full stop) [< Skt. *purna* "full" + *cched* "pause"], interrogation mark and exclamation mark. Most punctuation marks used in Bangla are borrowed from English punctuation system.

3. Shape analysis of characters

Generally, the vowels and consonants are considered ’basic characters’ because of their independent existence and their role in governing the use of other characters in the texts. The entity and role of vowel allographs and consonant graphic variants are measured by their use within words or morphemes. The formation of graphic compounds is caused by the change of vowel allographs used with consonants or clusters. The consonant clusters are formed by joining two or more consonant graphemes the role of which is context-bound (because their role in the script can be properly studied when put in the context of words).

3.1 Characters in isolation

The shape of the basic characters is a mixture of straight line, circular and semi-circular curves, thick dot and conic shapes. All these shapes are not of equal length although the height of most of the basic characters are identical. Moreover, the lines and curves are not always used in their full length in every occasion of character formation. Sometimes, the full length of a line or curve, sometimes the half of them, or sometimes just a portion of a line or curve is used for designing basic characters. However, the shape and design of some basic characters (*ii, kh, g, gh, ch, j, th, ph, bh, s* etc.) are more complex than that of other characters. The reason of their complexity probably lies in their process of formation where all the properties (i.e. dots, curves, straight lines and conic sections) are used.

The head line (*shirorekhaa*) is considered as an important feature of the characters because it acts as a line of demarcation at the time of tier division (discussed in section 3.3) of the characters. It is a property by which the basic characters can be grouped into two broad classes:

- characters with head line (32 in number)
- characters without head line (14 in number).

Moreover, according to the arrangement of different structures and properties, the basic characters can be grouped into three major classes:

- characters formed with linear structures arranged in different angles (15 in number)
- characters formed by dot and curve shapes (11 in number), and
- characters having both kind of shapes (26 in number).

The use of vertical line is maximum in the formation of the basic characters. Nearly 33 basic characters have vertical line in its full horizontal span. It is observed that while some characters (*n, b, r, dh*) contain vertical line at their rightmost side, some characters (*c, ch, T, Dh, Rh, d*) have it on their left most side. Similarly, while some characters (*aa, jh*) use vertical line twice where the second line is placed just parallel to the first one, some other characters (*u, uu, ch, D, d, R*) use only a half-length vertical line in their shape design. Moreover, the width of a character is not always proportionate to its height. While some characters (*g, l, sh*) are wider than their height some others (*N, n*) are less wide than their height. For some characters (*k, b, r*) the width and height are nearly equal

For automatic recognition an analysis of shape similarity of characters is required since some basic characters (*a/aa, u/uu, o/tt, kh/th, k/ph, t/bh, l/n, sh/n, b/r, D/R, T/Dh/Rh, g/p, y/gh/s, ks/hm*) are nearly similar in shape. They are generally considered as confusing characters because one can easily be confused with the other either by man or in machine recognition problem.

3.2 Characters in string

Characters used within words sometimes differ from their features noted in their isolation, because context can add some more features. Generally, the following factors control their roles in context:

- restrictions in positional use,
- modifications in original form or shape, and
- limitations in functional role

These factors generally lead vowels to be converted into allographs at word intermediate and final positions despite their use in original shape at word-initial position. Therefore, the occurrence of vowels (*i, o*) in basic shape at word-middle or final position carries separate implication (e.g., a particle for emphasis). Similarly, some consonants (*n, R, Rh, y*) cannot occur at word-initial position because of the restriction in their positional use. Moreover, the variant of the consonant *t* (called *khandata* "half-t") is not entitled to accompany a vowel allograph. Therefore, whenever such situation arises (particularly at the time of using case markers) it changes into the basic character (*t*).

The consonant *r* generally allows its two variants (*reph* and *raphalaa*) to occur only in cluster formation. While *reph* occurs above a consonant, *raphalaa* finds place at the bottom of the character. Moreover, *reph* is too weak to cause any structural change of the character but *raphalaa* is strong enough to modify the basic shape of some consonants.

3.3 Tier division of characters

An analysis of tier division of the characters is important for understanding actual behaviour of the characters within the words. In Bangla words, the characters are arrayed in three tiers: upper, middle, and lower. While the upper tier contains signature of basic characters and allographs, some consonant graphic variants (*candrabindu, reph* etc.), the middle tier contains the bulk of character shape, and the lower tier contains some allographs (*u-kaar, ri-kaar* etc.) and some consonant graphic variants (*raphalaa, vaphalaa* etc.). The allographs, when used with consonants or clusters, are distributed in all three tiers (Pal and Chaudhuri 1995). This kind of analysis has helped us in OCR system development in Bangla.

3.4 Graphic compounds

The graphic compounds, generally formed by joining two or three graphemes with allographs or graphic variants, can either be a combination of consonant and allograph, or a combination of consonant, graphic variant and allograph. In this process some change takes place in the original form of the characters. Compared to basic characters these are complex in form and design. It is noted that:

(a) the allograph of *u*, when used with consonants, creates three different grapheme-dependent compound shapes: (i) with consonant *r* or a cluster with *raphalaa* (e.g., *dr, gr, shr, br* etc.) the allograph changes its shape and is attached at the right hand side of consonant or cluster. The notable point is, the change takes place only with consonant *r*, either in its original shape or in graphic variants, (ii) with consonant *sh, g* and cluster *nt*, it changes its shape and is attached just at the bottom of consonant or cluster, and (iii) with the consonant *h* it entirely merges with the grapheme making the shape of the grapheme little more twisted.

(b) the allograph of *uu* also goes through structural change when used with consonant grapheme *r* and clusters with *raphalaa* (e.g. *gr, shr*). The allograph changes thoroughly in shape and is attached on the right hand side of the character.

(c) the allograph of *ri* goes through a directional change (not structural change) when used with consonant *h*. Here, the allograph changes its direction from horizontal to vertical and is attached to the right hand side of grapheme.

3.5 Consonant clusters

The consonants in clusters (formed by joining with other consonants) undergo three types of structural change:

(i) the shape of the graphemes are entirely modified to generate a new shape (e.g., *ks, kt, kr, ng, nc, tt, tr, hm* etc.)

(ii) the shapes of the graphemes are partly modified. We have found that for nearly 45 clusters, the shape of the first grapheme is modified, while for nearly 60 clusters, the shape of the last grapheme is affected.

(iii) three (nearly 35) or four consonants (nearly 5) can also join to form a cluster, where the last grapheme (either in full or in part) is attached at the bottom or right hand side of the cluster.

Recently, some compounds and clusters are simplified in shape for transparency and easy access in typewriter and computer (Sarkar 1993: 42), which is yet to be accepted widely by printing organisations.

4. Some quantitative findings

The introduction on different sub-disciplines like quantitative linguistics, stylometrics, applied linguistics, forensic linguistics etc. has raised a demand for different statistical and quantitative analysis of the occurrence of characters in the texts of a language for making various observation and hypotheses, developing primers for language learners, designing tools for CL and NLP etc. The use of statistics on language study is rejuvenated once the computer accessible huge corpus is available to the investigators. A corpus with a huge collection of empirical data with innumerable variations of use of different characters can easily be subjected to quantitative analysis which allows us to discover which characters are more regularly used in the language and which occur rarely. It allows us to get a precise picture of the frequency and rarity of particular character, and thus helps us to determine their relative normality or abnormality. This has led Yule (1964: 15) to comment that linguists without adequate knowledge of statistical information about different properties of language can make mistakes in handling linguistic data as well as in observations.

Probably, the Bangla language has been first quantitatively studied by Chatterji (1926/1993). He made a frequency count of lexical items in a Bangla dictionary and in some writings of Old Bangla. Bhattacharya (1965), on the other hand, has made different statistical analysis of phonemes, syllables, words and sentences on a collection of prose texts. Similarly, Das *et al.* (1984) have made some statistical studies on global character occurrence in Bangla, Assamese and Manipuri on some selected texts. The efforts of Mallik and Nara (1994, 1996) are mostly centred around the writings of the poet Tagore. To the best of our knowledge, multi-dimensional quantitative analysis on Bangla corpus can be credited to Chaudhuri and Dash (1998) and Chaudhuri and Ghosh (1998).

Before starting frequency count, some issues regarding character identification should be resolved. The decisions thus made are followed throughout the study on corpus at the time of programming which saved us from problems of wrong observation or deductions. The decisions are as follows:

- at character level frequency count, importance is given on each character’s position in the corpus.

- each grapheme, allograph, graphic variant, consonant cluster and graphic compound is considered as a single character.

- the percentage of vowel is obtained by adding the occurrence of their basic as well as allographs, while that of consonants is counted by adding the occurrence of their basic as well as their graphic variants (where applicable).

- for uniformity in processing and identification of the characters, the punctuation marks are uniformly separated from the characters in the texts.

The following section presents the frequency statistics of different characters used in corpus along with some discussions on the findings. Among statistical methods, frequency count is the most straight-forward and rudimentary approach to work with quantitative data. In this process the characters in corpus are classified according to a particular scheme and an arithmetical count is made on the number of items within corpus which belong to each classification of the scheme. The frequency count provides number of occurrences of each type used in corpus. The results may be used for estimating position of different characters in script, to design primers for language users, as well as to develop different tools for computer implementation. Moreover, various hypotheses presented by earlier scholars about the use and occurrence of characters are also verified by the findings. Keeping these factors in mind, we have counted a few simple statistics of the characters which are given below:

(i) Among total number of characters used in the corpus, the occurrence of *aa* is maximum followed by *e, r* and *i*. Among the vowels, *aa* (11.965%) including allographs comes first, followed by *e* (9.793%), *i* (7.745%), *u* (2.379%) and *o* (2.027%), while among the consonants, *r* (8.633%) is maximally used followed by *n* (5.033%), *k* (4.898%), *t* (4.312%), *b* (3.800%), *s* (2.942%), *l* (2.866%), *m* (2.826%), *p* (2.562%), *y* (2.143%) and *d* (2.127%). The occurrence of *r* is maximum in corpus probably because of the occurrence of its two variants. Similarly, the percentage of use of *t* is increased due to the presence of its variant. Among the first 10 most recurrently used characters in the corpus, 6 are consonants while the remaining 4 are vowels and all these characters are easier to articulate than others present in the language. In Hindi corpus also, the vowel *aa* occurs maximally in the script (Tripathi 1971: 26).

(ii) The occurrence of vowel (39.63%) and consonant (52.76%) consists nearly 92.39% of the total characters used in the corpus. Though there are nearly 280 types of cluster in the script, their use is quite less (07.61%). The percentage of cluster is higher in a similar context if the text is written in chaste (*saadhu*) version which is older than the colloquial (*calita*) version now in vogue. This supports our argument that Bangla is gradually simplified in utterance and pronunciation and the clusters are gradually replaced by single consonants.

(iii) It is found that words starting with the consonant *k* (9.81%) is highest in the language followed by *p* (8.68%), *b* (8.58%), *s* (8.24%), *e* (5.43%), *aa* (4.85%), *n* (4.64%), *m* (4.63%), *t* (4.50%), *d* (4.47%) and *h* (4.45%). Moreover, words starting with consonants (81.52%) is more in number than words starting with vowels (18.48%). It provides an interesting insight into the nature of the language. Out of top ranking 20 characters, the vowels are 5, while the remaining 15 are consonants which are easy to articulate. It supports the commonly used statement that Bangla is easier to speak and perhaps sweeter to listen than many other Indian languages.

(iv) Among allographs, the occurrence of *aa* (34.20%), similar to Hindi (Khan *et al.* 1991: 272), is highest in the script followed by *e* (29.44%), *i* (18.75%), *u* (06.64%) and *o* (04.50%), consecutively. The counting of *a* is not possible as it has no allograph, though our common belief is that the occurrence of *a* would have been highest in the corpus, because in most cases a consonant or cluster if not attached with any allograph carries an inherent [ɔ] with it. However, this hypothesis can only be authenticated once a speech corpus is analysed.

(v) The use of cluster in language is gradually reduced because in corpus the number of words without cluster is much more (81.57%) than words having one (14.25%), two (2.72%) or three (1.00%) clusters. Words having more than three clusters are very rare, mostly *tatsama* compounds. The statistics hints for almost complete loss of clusters from the language in future. Among the first 20 most frequently used consonant clusters, *pr, ks, tr, st, sv, ny, sth, gr, by, jn* and *shy* can occur at any position of words, while the remaining clusters can occur only at word-intermediate and/or final position. Among these, *pr* (8.16%) occurs the most because of its maximum occurrence at word-initial position, and the graphemes comprising the cluster, are quite frequent in the language.

(vi) Among consonant graphic variants *raphalaa* (37.94%), *yaphalaa* (26.66%) and *reph* (22.70%) are highest in occurrence. They are used in the script for the purpose of cluster formation. Probably, because of their recurrent use in the language they are made simple in shape so that they can be used easily in the script. Here we argue with statistical support that because of their recurrent use in script they are most simple in shape and designed to make writing easier and faster.

(vii) Following Miller *et al.* (1958) the statistics of relative frequency of use of punctuation marks in Bangla corpus is taken. It is noted that the use of comma, like that of English (Bayraktar *et al.* 1998), is highest (22.32%) followed by *purnacched* (17.26%), semicolon (15.27%), hyphen (8.89%), note of interrogation (7.38%), colon (6.16%) and note of exclamation (04.36%), respectively.

5. Use of punctuation marks

Generally, some syntactic and semantic properties of a sentence are controlled by punctuation marks as they are used in texts to mark out strings of words into manageable groups. The primary role of punctuation marks is to show the pause of breath in a sentence besides clarifying meanings or, in some cases, preventing wrong meaning being deduced form a sentence. Traditionally, two functions of punctuation marks are considered: (i) grammatical function where they help the construction or structure of a sentence, and (ii) rhetorical function where they help to deduce the hidden implication of a statement. However, Crystal (1995) has defined four functions: grammatical, prosodic, rhetoric and semantic function.

The role of punctuation marks in Bangla is not yet scientifically estimated though there have been some attempts by Saraswati Press (1956), Roy (1989), Chakravarti (1994), Chatterji (1973), Bhattacharya (1999) and others. We here try to present an estimate by citing examples from the corpus.

The punctuation marks most commonly used in Bangla to divide a piece of prose writing are: *purnacched*, semicolon, comma, colon, note of interrogation, note of exclamation, apostrophe, quotation mark, brackets (round, braces and square), dash, ellipsis, hyphen and space, besides arrow mark, percentage mark, equal mark, therefore mark, underline, implication mark etc. Moreover, various mathematical and geometric symbols are used in scientific books and articles. Full stop, therefore, marks the main division into sentences, semicolon joins sentences, and comma which is most flexible in use, separates smaller elements with the least loss of continuity. Brackets and dashes also serve as separators - often more strikingly than commas. In the following sections only a brief outline is given:

- In Bangla a *purnacched* is regularly used as a terminal marker at the end of a sentence which is a statement (not question or exclamation) either in descriptive, declarative, narrative or imperative sense.

- The use of comma has a lot of variation in practice. Its primary role is to give detailed description to the structure of sentences, especially longer ones, and to make their meaning clear. Too many commas can be distracting; too few can make a piece of writing difficult to read or, worse, difficult to understand. It is widely used to separate main clauses of a compound sentence when they are not sufficiently close in meaning or content to form a continuous unpunctuated sentence, and are not distinct enough to warrant a semicolon. Moreover, it is used for a short time break during utterance that indicates a pause between parts of a sentence, or dividing items in a list, string of figures etc. Besides, it is used in pairs to separate elements in a sentence that are not part of the main statement, to stop for a short while in various sections within a sentence, to provide a slight pause between words in a sentence and between dates of days, month and year. Generally, this technique is not practised in Bangla though our corpus contains some such instances. It is used in numeral of four or five figures, to separate each group of figures starting from the right; before the reporting of speech; and sometimes used in place of parenthesis where it is used just before and after the parts of text.

- Colon is generally used to introduce a quotation or a list of items; or to separate clauses when the second clause expands or illustrates the first. It acts to separate main clauses when there is a step forward from the first to the second, especially from introduction to main point, from general statement to example, from cause to effect, or from premises to conclusion; to deliver the terms (name of books, persons, countries or a statement of people) that have been mentioned in the preceding words; to introduce a list of items; to indicate time in hours, minutes and seconds in writing; and between numbers in a statement of proportion or ratio.

- Semicolon is of intermediate value between a comma and a full stop as it denotes a time-break which is longer than comma but shorter than full stop. The main function of semicolon is to unite sentences which are closely associated or which are complement or parallel to each other in some way; to join two sentences different in meaning or content;

to divide some parts of a sentence meaningfully where each part has commas for its own use etc. Moreover, it is used as a stronger division in a sentence that already includes division by means of commas and when the name and designation of some persons are to be shown in a sentence.

- The question mark is primarily used to indicate that the sentence is an interrogative one. Sometimes it is used within brackets in the middle of a sentence to express uncertainty or doubt about a fact, word or phrase immediately following or preceding it.

- The exclamation mark is generally used at the end of a sentence to denote a sense of exclamation. Moreover, it is also used after words within a sentence to express absurdity, command, warning, contempt, disgust, emotion, pain, encouragement, wish, regret, wonder, admiration, surprise, grief etc.

- Apostrophe is used to indicate omission of letters or numbers; to denote loss or contraction of some letters in words; and to denote loss of digits in years.

- Generally, two types of quotation mark are used in Bangla to indicate direct speech and quotation: single and double quotation. While single quotation is used to quote or mention title of books or other things; to denote special symbols or words used in texts; to mean that a word or phrase is cited in the sense of pun etc.; the double quotation is used to denote speech or dialogue in a story or novel; and to quote other's speech correctly in a piece of writing. It is noted that the closing quotation mark should come after any punctuation mark which is a part of the quoted matter, but before any mark which is not.

- In corpus the use of three types of brackets are noted: parentheses, brace and square bracket. Among these, use of braces is very rare, mostly for denoting some mathematical symbols or notations while use of parentheses and square bracket is almost regular. Parentheses are used to enclose explanation and extra information or comment; to give reference or citation to any date or event or work or person etc.; and to enclose reference letter and number. Square brackets are used to enclose extra information which is attributive to something, someone or some place; and to convey special kinds of information, especially when parentheses are used for other purposes. For instance, in standard Bangla dictionaries they are used to give the etymologies at the end of the entries.

- The purpose of using dash in writing or printing is to mark a break in words; or to represent omitted letters or words. It is also used in place of parenthesis. Sometimes a single dash is used to indicate a pause or hesitation in speech; to introduce an explanation or expansion of what comes before it; and to indicate omitted words especially slang or coarse words in reported speech.

- The use of dot is very rare in Bangla. Generally, it is used to denote abbreviation or shortening of some portion of words. The use of ellipsis is also very less in Bangla. Generally, it is used to indicate fumbling of speakers or incompleteness of sentences which are already started. Moreover, if one does not want to quote the full texts of some writings one can use this sign in those places of the text where it is left.

- Asterisk is used as a reference mark in the footnote to explain some ideas or to denote source of some texts. It is used just immediately before or after word or sentence form which one can directly refer to the footnote. It is also used to denote some words or characters silent in texts; to emphasise on or to draw attention of readers to a particular item in the text. For instance, in a list of books, some can be marked with asterisk to denote that these books are either very important or rare etc. Sometimes, more than one asterisk marks are used at a time to indicate a break or lapse of a section of an article or text.

- The alternative sign (generally called as *stroke*) is used to denote break of lines of a poetry, to describe two or more related words or items etc. In linguistic analysis it is used to denote phoneme or phonetic segment or a syllable.

- The space between words is never considered as punctuation mark but its role in the text is of equal importance like other punctuation marks. What is a measured pause in speech is probably a calculated space in writing. It provides gap between two consecutive words in a writing to identify words in a sentence or texts.

The hyphen is probably the most complex punctuation mark used in Bangla. It's role is not yet fully defined as there is no regularity in its use in the language. In standard Bangla dictionaries it is considered as a sign which joins two syllables or words. This definition is partly true as it captures only a fragment of the multiple roles of hyphen as noted in corpus. Its use in compounds is arbitrary, especially when elements of compounds are of one syllable. Except for some unavoidable situations, it is randomly used. In some occasions it is not used though needed, on the contrary, it is used in those occasions where it is not required. In the corpus more than twenty types of use are noted as given below. A hyphen is generally used:

- more often in routine and occasional couplings, especially when reference to the sense of separate elements is considered important and unavoidable. It is used between compounds of two nouns (*cor-Daakaat* "thief and robber"), two adjectives (*rogaa-moTaa* "thin and thick"), noun and adjective (*man-gaRaa* "fabricated"), pronoun and noun (*sei-din* "that day"), and cardinal adjective and noun (*tin-purus* "three generation") etc.

- between two proper names (*seli-kiTs* "Shelley-Keats"), words of similar meaning (*Taakaa-paysaa* "rupees and pennies"), department and post (*krisi-mantri* "agriculture minister"), institution and position (*skul-maasTaar* "school teacher"), place and occasion, (*baarlin-alimpik* "Berlin Olympic"), for direction (*uttar-pashcim* "North-West") etc.

- to connect words having syntactic link (*kathaay-kathaay-raag-karaa mejaaj* "getting-angry-in-every-word temper"), to link compounds and phrases used attributively (*haajaar-haat-kaali* "goddess Kali with a garland made of thousand cut-off hands"), to denote a sense of continuation (*co-o-o-o-r* "thief"), to indicate sounds of music or musical instruments (*taa-dhin-dhin-taa*), to write some names of non-Indian origin (*maao-se-tung* "Mao-Tse-Tung") etc.

- for all similar words to stop repetition of the second constituent when the second constituent of compounds of a sentence are common. The second constituent is used only in the last word in the sentence (*raajya juRe shramik-, bekaar-, bidyut-, khaadya-, jal-samasyaa dekhaa diyeche* "The problem of labour, unemployment, electricity, food, water are noted in the state"). Hyphen here performs the role of concurrence.

- between prefixes and nouns, (*ku-najar* "ugly look"), between monosyllabic noun and suffix (*paa-Ti* "the leg"), between monosyllabic pronoun and suffix (*ka-Taa* "how many"), in compounds where the first part is a single letter word (*bhu-prakriti* "geographical nature") etc.

- between reduplicated words (*maajhe-maajhe* "sometimes"), onomatopoeic words (*jhan-jhan* "tinkling"), echo words (*bis-Tis* "poison etc."), for exclamatory expression (*ho-yaaT* "what"), for emphasis (*maa-i* "mother herself"), for abbreviated words (*bi-bi-si* "BBC"), for avoiding awkward collision of homophonous characters (*taap-prabar* "heat-strong") etc.

- when an inflected word form is used as noun and is added with further suffixes for linguistic analysis (*moder-er byabahaar kameche* "Use of 'moder' is declined") etc.

- between proper name and case ending or suffix which is added with proper name of person (*MilTan-er* "of Milton"), institution, (*aai.es.aai-te* "in ISI"), book (*myaakbeth-er* "of Macbeth"), newspaper (*sTeTsmyaan-e* "In the Statesman"), day (*sombaar-e* "on Monday"), year (*1947-e* "in 1947") etc. However, such use is not a regular feature of the language. Generally, case endings and suffixes are attached with proper names without inserting a hyphen in between.

- to retain the original structure of some words unchanged when they use suffixes (*pad-er* "of words", *desh-er* "of Desh") though without this hyphenation the grammatical value of the words would not have been changed

- whenever a case marker is added with words ending with half-t (*bhabisyat-e* "in future").

- after Bangla characters when they are subjected to grammatical or linguistic analysis (*a-Taa baanglaar pratham varna* "a is the first letter in Bangla").

- in vowel allographs (*u-kaar* "u-allograph") for grammatical reasons, and between some consonants and their place of articulation (*dantya-sa* "dental-s").

- with affixes and case markers when used for linguistic analysis. Usually, prefixes are written with a hyphen immediately after them (*pr-, bi-*), while suffixes and case markers are written with a hyphen immediately before them (*-der, -ke, -te*). Hyphen works here as mark of their identity

- between numbers (*2-3*), between years (*1986-1990*), and between numbers and words (*6-phuT* "6-feet").

- when native post-positions or case markers are used with non-nativised foreign words (*mesin-er* "of machine"). Such kind of use is very rare in corpus. Generally, nativised foreign words use native post-positions or case markers without hyphen (*skuler* "of school").

- when some English group verbs (*pick up, by pass* etc.) are transliterated in Bangla (*pik-aap, baai-paash*).

- when a single-letter name is used to hide somebody's identity (*ka-baabu* "Mr. X").

- after an abbreviated word with a colon to indicate that the full form of the word is deliberately omitted (*pr:-* "question").

- between words which are not compounds but which by their peculiar combination either denote a sense of hesitation (*dicchi-debo* "dilly-dally"), or a sense of request (*baabaa-baachaa* "appeasing"), or a sense of adverb (*saat-taaRaataaRi* "in a haste"), or a sense of pun (*jyoti-hin pashchimbanga* "West Bengal without light").

As noted above the use of hyphen in Bangla is full of varieties. It serves as a means for dissolving lexical ambiguity embedded within the surface forms. Moreover, it helps us to find actual meaning of words or phrases as well as stops us from deducting wrong meaning. The corpus has cited some interesting instances where deletion or addition of hyphen changes the meaning of words. Table (2) shows some examples.

Without hyphen	Meaning	With hyphen	Meaning
Asukh	illness	*a-sukh*	not happy
KaTaa	brownish yellow	*ka-Taa*	some
PaaTaa	plank	*paa-Taa*	the leg
Amrita	nectar	*a-mrita*	not dead

Aakaar	shape	*aa-kaar*	a-allograph
CaaTaa	lick	*caa-Taa*	the tea
Ekaar	alone	*e-kaar*	e-allograph
Maar	kill	*maa-r*	of mother
Kushaasan	mat of Kush grass	*ku-shaasan*	bad ruling

Table 2: Change of meaning of the words with/without hyphen

Similarly, displacement of hyphen changes the meaning of words. For instance, the word *surat-ranga* means "game of coition" while *sur-taranga* means "waves of music", or *akhyaata-naamaa* means "notorious" while *a-khyaatanaamaa* means "non-famous" etc.

To divide a word with a hyphen at the end of a line is altogether a different matter because it is not a regular feature of spelling. It is more common in print, where the text has to be accurately spaced and the margin has to be justified. With little care it can be avoided totally in hand-written, typed or word-processed text materials. In printing the words need to be divided carefully and consistently taking account of the appearance and structure of words.

6. Conclusion

Researchers who are studying the evolution of thought process in human societies, believe that development of language and script may also influence human cognitive powers. Script is a form of knowledge representation in which the use of grapheme makes demand on humans to code and decode knowledge, convert auditory sounds into visual symbols, think deductively and order words to construct sentences. The study of characters used in a language is important and useful for developing various tools for NLP and CL such as OCR system, cryptography, key-board design for typewriter and printing, telegraphic codes design, spell-checker design, dictionary preparation, machine translation, information-theoretic analysis of language, language teaching etc. Even from pure linguistic point of view such study can help both primary and secondary language users to know how the characters are designed and used in the language.

Acknowledgement: The Ministry of Information and Technology, Govt. of India is thanked for providing the corpus for studies presented here.

Bibliography

Atkins S, Clear J, Ostler N 1992 Corpus Design Criteria. *Literary and Linguistic Computing* 7(1): 1-16.
Babulanam S M, Beena K F 1999 The User-Oriented Bengali Easy Orthography. *Computers and the Humanities* 33: 241-245.
Bandyopadhyay C (ed.) 1981 *Dui Shataker Bangla Mudran O Prakashan* (The printing and publishing history of Bangla in last two centuries). Calcutta, Ananda Publishers.
Banerjee R D 1919 *The Origin of the Bengali Script.* Calcutta, Calcutta University Press.
Bhattacharya N 1965 *Some Statistical Studies of the Bangla Language.* Unpublished PhD thesis, Indian Statistical Institute, Calcutta.
Bhattacharya S 1999 *Tista Ksanakal: Biramcihna O Anyanya Prasanga* (Wait for a While: The Bangla Punctuations and other Issues). Calcutta, Ananda Publishers.
Bayraktar M, Say B, Akman V 1998 An Analysis of English Punctuation: The Special Case of Comma. *International Journal of Corpus Linguistics* 3(1): 33-58.
Biber D 1993 Representativeness in Corpus Design. *Literary and Linguistic Computing* 8(4): 243-257.
Chakravarti N (ed.) 1994 *Bangla: Ki Likhben, Kena Likhben* (Bangla: What to Write and Why to Write). Calcutta, Ananda Publishers.
Chakravarty S N 1938 Development of the Bengali Alphabet from the Fifth Century AD to the End of the Muhammadan Rule. *Journal of the Royal Asiatic Society of Bengal* 4: 351-391.
Chatterji S K 1973 *Bangla Bhasatattver Bhumika* (An Introduction of the Bangla Linguistics). Calcutta, Calcutta University Press.

Chatterji S K 1926/1993 *The Origin and Development of the Bengali Language*. Calcutta, Rupa Publications.

Chaudhuri B B, Dash N S 1998 Bangla Script: A structural Study. *Linguistics Today* (2)1:1-28.

Chaudhuri B B, Ghosh S 1998 A Statistical Study of Bangla Corpus. in the *Proceedings of International Conference of Computational Linguistics, Speech and Document Processing (ICCLSDP'98):* C32-37.

Crystal D 1995 *The Cambridge Encyclopaedia of the English Language*. Cambridge, Cambridge University Press.

Das G, Bhattacharya S, Mitra S 1984 Representing Assamese, Bengali and Manipuri Text in Line Printer and Daisy-Wheel Printer. *Journal of the Institution of Electronics and Telecommunication Engineers* 30: 251-256.

Dash N S, Chaudhuri B B 2000 The Process of Designing A Multidisciplinary Monolingual Sample Corpus. *International Journal of Corpus Linguistics*. (forthcoming)

Diringer D 1953 *The Alphabet: A Key to the History of Mankind*. London, Hutchinson's Scientific and Technical Publications.

Khan I, Gupta S K, Rizvi S H S 1991 Statistics of Printed Hindi Text Graphemes: Preliminary Results. *Journal of the Institute of Electronics and Telecom Engineering* 37(3): 268 -275.

Mukhopadhaya B 1981 Bangla Mudraner Car Yug (Four Era of Bangla Printing History) in Bandyopadhaya C (ed.) *DuiSataker Bangla Mudran O Prakashan* (The printing and publishing history of Bangla in last two centuries). Calcutta, Ananda Publishers.

Mallik B P, Nara T (eds) 1994 *Gitanjali: Linguistic Statistical Analysis*. ILCAA, Tokyo University Press.

Mallik B P, Nara T (eds) 1996 *Sabhyatar Sankat: Linguistic Statistical Analysis*. Calcutta: Rabindra Bharati University.

Miller G A, Newman E B, Friedman E A 1958 Length-Frequency Statistics for Written English. *Information and Control* 2 (1): 370-389.

Pal U, Chaudhuri B B 1995 Computer Recognition of Printed Bangla Script. *International Journal of Systems Science* 26(3): 2107-2123.

Roy A K 1989 *Unis Sataker Bangla Gadya Sahitya: Ingreji Probhab* (The Bengali Prose Literature of the 19th Century: The Impact of English). Calcutta, Jignasa Prakashani.

Saraswati Press 1956 *Rules for Compositors and Readers*. Calcutta, Saraswati Press.

Sarkar P 1993 Bangla Bhasar Yuktabyanjan (Consonant Clusters in Bangla Language). *Bhasa* 1(1): 23-45.

Sen S 1993 *Bhasar Itivrittva* (The History of Language). Calcutta, Ananda Publishers.

Sinclair J 1991 *Corpus, Concordance, Collocation*. Oxford, Oxford University Press.

Tripathi J N 1971 A statistical Analysis of Devnagari (Hindi) text graphemes. *Journal of Institute of Electronics and Telecom Engineers* 17(1): 25-27.

Yule G U 1964 *The Statistical Study of Literary Vocabulary*. Cambridge, Cambridge University Press.