Bangla Script: A Structural Study

Niladri Sekhar Dash and **Bidyut Baran Chaudhuri**
Computer Vision and Pattern Recognition Unit
Indian Statistical Institute, Kolkata, India
Email: ns_dash@yahoo.com

Abstract

In this paper we have tried to analyse the shape of the graphemes used in the Bangla script (as noted in printed documents). The study has focused on the formation of graphemes, their structural changes in case of compound grapheme formation, contextual use of graphemes and their allographs, statistical analyses of their occurrences in corpus and their positional and functional roles in case of semantic changes. The purpose of this study is to understand the role of the graphemes in the language; to show their behavioural peculiarities and to find out the reasons of such peculiarities. Information obtained from this study may be useful for optical character recognition, spelling checker designing, key-board designing, cryptography, language teaching, and natural language processing in Bangla.

1. Introduction

The script is the visual representation of a natural language. It is the collection of some unique symbols or characters known as graphemes, which are arranged in specific patterns with appropriate punctuation marks in texts. The total set of symbols or graphemes is known as *alphabet* of the language. Most of the linguistic features of a language are retained in the script so that it can be easily read and understood. The study of script is therefore very important for the study of any language. Moreover, this study is useful and necessary for computer key-board design, optical character recognizer development, language learning and teaching (both primary and secondary), speech analysis and synthesis besides other applied and interdisciplinary studies.

Historians have identified two types of Proto-Indian scripts in India. One is the *Khorosthi* script (also known as the *Cuneiform script* for it conical shape) which is used in the North-Western Frontier provinces of India. The other is the *Brahmi* script which is found in other provinces of the Northern India. The *Brahmi* script, deciphered by James Princep in 1838, is claimed (Ganguli 1994) to be the origin of all modern Indian scripts other than the Persian-Arabic and the Roman based ones. Historians believe that this script was formed nearly 1000 years B.C. in India. The historical evolution of this script has taken place differently in the Northern and the Southern India (Majumdar 1995).

Linguistics Today. Vol. 1. No. 2. Pp. 1-28. 1998.

Presently, in India there are four types of script in use as observed by experts (Majumder 1995):

(a) The scripts of Bangla, Devanagari, Gurmukhi, Assamese, etc., which are originated from the old *Brahmi script*,
(b) The scripts of the Dravidian languages like Tamil, Telugu, Kannada and Malayalam which are originated through old Battejhu.ttu and Pahllava script,
(c) The Persian-Arabic script like Urdu, Sindhi, Kashmiri etc, which are derived from the Semitic script, and
(d) The Roman based script which is used for some tribal languages like Santali, Mundari, etc. which have originally no script.

There exist some studies on the script of European languages, particularly for Roman (English) (Diringer 1968). The study on modern Indian scripts is limited to the modifications of script for ease of writing and printing. Thus Devnagari script has been simplified to write and print compound graphemes. Some proposals on the Bangla script modification are put forward during the past sixty years (Chatterji 1993). As a result, some graphemes used in Bangla script have been deleted from the alphabet, and modifications of some compound graphemes have been made. Nevertheless, the process of script modification and simplification in Bangla lags far behind Devnagari (for Hindi) and the compound grapheme shapes of Bangla, at present, are more complex than that of the Devnagari.

There is a need for statistical studies from the corpus of Indian language texts in this regard but very little work has been done so far. Years ago, the grapheme occurrence frequency in Hindi has been calculated on a small corpus (Tripathi 1971) but occurrence frequencies are not taken in a true graphemic way. In the Bangla script some statistical studies on grapheme occurrences are conducted on some Bangla texts a few decades ago (Bhattacharya 1965). We are not aware of any phonemic statistical analysis of Indian language speech corpus except the one done by a small group in ISI, Kolkata (Chaudhuri and Pal 1995).

Recently, however, a reasonably large corpus in Indian languages is available due to sponsorship of the Department of Electronics (DOE), Govt. of India. As a result, more complete, elaborate and robust statistical analyses are possible to carry out on this corpus. For the study of the Bangla script we have taken the TDIL corpus that contains more than 30 lakh words as well as a corpus consisting more than 5 lakh words which is developed at the Computer Vision and Pattern Recognition Unit of Indian Statistical Institute, Calcutta. Both the corpora are developed by collecting printed language data

from almost all disciplines of human knowledge published between the year 1981 and 1995. The corpora, under study, give a clear view about the characters used in the Bangla script with adequate information about their shape, size, occurrence, articulation and graphemic changes for various statistical analyses and observations.

This paper is organized in the following way: in Section 2, we highlight the basic features of the Bangla script, in this section also we present a comparison of the Bangla script with other Indian scripts. In Section 3, various statistical analyses on the Bangla script is reported. In Section 4, we analyse the shape of the graphemes in isolation as well as within words. In this section we also give tier-division of graphemes and allographs along with some description on compound graphemes. In Section 5, we evaluate the impact of characters on utterance in Bangla speech. The Section 6 is the conclusion where the importance of this type of character analysis is discussed.

2. Features of the Bangla Script

The Bangla script is evolved through hand written documents (Sen 1992) and as a result it has made some modifications on the shapes of the graphemes. But when it is first mechanically designed and stratified for the purpose of printing, the scope of structural modification is almost stopped. The first grapheme design for printing Bangla language was done by Charles Wilkinson and his script was used to print *A Grammar of Bengali Language* written by Nathaenial Brassi Halhead at the Hoogli district in 1778 (Banerjee 1981). Later, Panchanan Karmakar took this design from Wilkinson and modified the script to give the present shape.

Similar to other Indian scripts, the Bangla script has also no indigenous origin. It is highly influenced by its contemporary sister language scripts. Though it is directly evolved from *Brahmi,* the influence and interpolation of other scripts of India can not be ruled out. Structurally, the Oriya and all the Dravidian scripts are round, semi-round and twisted, whereas other Indian scripts are conic and triangular in shape and form. The Bangla script has both kinds of shape and structure though the similarity with the Aryan script is more than that of the Dravidian scripts. The followings are the main features of Bangla script as noted in our database.

(a) The Bangla graphemes are read and written from left to right direction both at word and sentence levels.
(b) There are nine (9) vowel graphemes, two (2) diphthong graphemes, twenty (20) vowel allographs, and thirty nine (39) consonant graphemes along with nearly 380 unique consonant grapheme clusters.

(c) Except the vowel grapheme অ (a) all other vowel graphemes have at least one allograph. For example, the vowel grapheme আ (A) has the allograph া, ই (i) has ি, ঈ (I) has ী, উ (u) has ু, ঊ (U) has ূ, ঋ (r) has ৃ, এ (e) has ে and ে, ঐ (ai) has ৈ, ও (o) has ো and ঔ (au) has ৌ.

(d) In printed form the vowel grapheme এ (e) has two allographs. One is without the *matra* (headline) and the other is with *matra* (headline). The contexts of their use are also different.

(e) Allographs can be used with consonant graphemes and consonant clusters but only one at a time. Allographs can never be used with a vowel grapheme or another vowel allograph. A single consonant grapheme or a cluster can use only one allograph with it at a time.

(f) Vowel graphemes and allographs are always articulated in words. It never happens that a vowel grapheme or an allograph is not articulated in spite of being present with a consonant or cluster in a word.

(g) On the other hand, a consonant grapheme may be silent in articulation in certain contexts despite being physically present within a word.

(h) Generally, the shape of an allograph is grapheme independent. However, there are some exceptions where the shape of an allograph is changed based on the shape of a consonant grapheme. For example, the original shape of the allograph ু (u) is changed when it is used with consonant graphemes like গ (g) and শ (sh) and takes shapes like গু (gu) and শু (shu), respectively.

(i) Consonant graphemes or clusters have no allograph. But the consonant grapheme র (r) has two modifiers: র্ (reph) and র্ (ra-phalA). Similarly, consonant grapheme য় (y) has ্য (yaphalA). These modifiers are normally used in cluster formation with other consonant graphemes.

(j) A single grapheme most often represents a single abstract sound. But the reverse one is not true. That means two or three graphemes may represent a single sound. For example, both long ঈ (I) and short ই (i) represent /i/. Similarly, both long ঊ (U) and short উ (u) represent /u/. Among consonant graphemes, palatal শ (sh) and retroflex ষ (S) represent /ʃ/ while retroflex ণ (N) and dental ন (n) represent /n/.

(k) Only consonant graphemes can form *yuktabyAñjan varNa* (consonant cluster). Here consonant graphemes are physically joined in the operation. Clusters of three or four consonant graphemes are also possible in Bangla. There are nearly 380 unique consonant grapheme clusters of which the cluster of two consonant graphemes counts nearly 290, of three consonants count nearly 80 and that of four consonants count around 10.

(l) The sentence terminal marker in Bangla is ।(pUrNacched) "full stop". It is perhaps identical for all the Aryan scripts. All Dravidian scripts, however, use a dot (.) like the Roman script for the same function.

(m) Other punctuation marks used in Bangla script as well as in other Indian scripts are directly borrowed from the Roman script through English.

The position of a vowel allograph with respect to a consonant grapheme or a consonant cluster is not always uniform. The allographs like ে ,ে and ৈ are put on the left hand side of a consonant grapheme or a cluster; the allograph ি is used on left hand side and above of a consonant grapheme or a cluster; the allograph া is put on the right-hand side of a consonant grapheme or a cluster, the allograph ী is used on the right hand side and above of a consonant grapheme or a cluster; the allographs ু, ূ and ৃ are put at the below of a consonant grapheme or a cluster. Even some allographs are put around the consonant grapheme or a cluster, as in, ো and ৌ. The list below (Table 1) shows the forms and positions of allographs with respect to consonant graphemes.

Vowel Grapheme	Vowel Allograph	Use with Consonant Grapheme
অ (a)	Null	ক (k)
আ (A)	া	কা (kA)
ই (i)	ি	কি (ki)
ঈ (I)	ী	কী (kI)
উ (u)	ু	কু (ku)
ঊ (U)	ূ	কূ (kU)
ঋ (r)	ৃ	কৃ (kr)
এ (e)	ে	কে (ke)
ঐ (ai)	ৈ	কৈ (kai)
ও (o)	ো	কো (ko)
ঔ (au)	ৌ	কৌ (kau)

Table 1: Form and positions of vowel allographs with respect to consonants

Structurally, the Oriya and the Dravidian graphemes are round, semi-round and twisted whereas other Indian graphemes are conic and triangular in shape and form. The list of Bangla graphemes includes both kinds of shapes and structure though the similarity with the Aryan graphemes is more than that of the Dravidian. Among the vowel graphemes structurally, Bangla অ (a) and আ (A) are similar to those of Assamese, Devnagari, Oriya, Tamil and Telugu; ই (i) is similar to those of Assamese, Devnagari, Telugu, and

Kannada, ঈ (I) is same to those of Assamese and Devnagari; উ (u) and ঊ (U) are identical to those of Assamese, Devnagari, Oriya, Gujarati and Gurmukhi; এ (e) and ঐ (ai) are similar to those of Assamese, Oriya and Tamil; and finally, ও (o) and ঔ (au) are same to those of Assamese, Oriya and Malayalam.

In case of vowel allographs it is interesting to note that none of the Indian scripts, both Aryan and Dravidian, has allographic form for the vowel grapheme অ (a). Other vowel allographs in Bangla are similar to other Indian scripts in the following ways:

- া is same with Assamese, Devnagari, Oriya, Gujarati, Gurmukhi and Tamil;
- ি is same with Assamese, Devnagari, Oriya, Gujarati and Gurmukhi;
- ী is similar to Assamese, Devnagari, Oriya, Gujarati, Malayalam and Gurmukhi,
- ু is same with Assamese, Oriya, Tamil, Kannada and Malayalam,
- ূ is identical to Assamese, Oriya, and Malayalam; ৃ is same with Assamese, Devnagari and Gujarati;
- ে is similar to Assamese, Oriya, Tamil and Malayalam;
- ৈ is same with Assamese and Oriya;
- ো is same with Assamese, Oriya, Tamil and Malayalam, and finally,
- ৌ is identical to Assamese, Devnagari, Oriya, Gujarati and Tamil.

The Bangla consonant graphemes are closely similar to that of Assamese with slightest variation in case of র (r) and ব (b). In case of র (r) in Bangla, the grapheme has a dot (.) just below the lower arm whereas in Assamese the grapheme is crossed within with a slanted line as in ৰ. In case of ব (b) in Bangla, the grapheme has no mark at its below, whereas in Assamese there is a short slanted line parallel to the lower arm of the grapheme as in ৱ. The structural similarities of consonant graphemes with other Indian scripts are as follows:

- Assamese has similarity with all Bangla graphemes except র (r) and ব (b).
- Oriya has similarity with: ঙ (ng), ড (D), ড় (R), ঢ (Dh), ণ (N), ত (t), ন (n), থ (th),দ (d), ভ (bh),ল (l), ঁ (~) and ঃ (H).
- Devnagari has similarity with: ক (k), গ (g), ঘ (gh), ঙ (ng), ট (T), ড (D), ড় (R), ঢ (Dh), ঢ় (Rh), ণ (N), থ (th), দ (d), ধ (dh), ন (n), প (p), ব (b), ম (m), য (y), ল (l), ব (v), ষ (S), স (s), হ (h) and ঃ (H).
- Gujarati has similarity with: গ (g), ঘ (gh), ঙ (ng), ঞ (ñ), ট (T), ঠ (Th), ড (D),□থ (th), ন (n), ম (m), য (y), ষ (S), স (s), ঁ (~) and ঃ (H).
- Gurmukhi has similarity with: ট (T), ত (t), য (y), র (r) and স (s).

- Tamil has similarity with: ঞ (ñ), ণ (N), ল (l) and ঃ (H).
- Telugu has similarity with: ত (t), ং (M) and ঃ (H).
- Kannada has with similarity with: ত (t), ং (M) and ঃ (H).
- Malayalam has similarity with: ক (k), ণ (N), ন (n), ং (M) and ঃ (H).

One of the primary differences of the Bangla script from that of the Roman is that while the latter has both Upper Case (i.e., capital letters) and Lower Case (i.e., small letters), Bangla script has no such variation. This is also true to other Indian scripts. On the other hand, Bangla and other Indian scripts have some vowel allographs as well as consonant grapheme clusters which are not available in the Roman script. However, the Roman script had some vowel grapheme clusters (e.g., Œ, Æ, æ, œ, etc.) which were designed for specific purposes. These are no more in use now-a-days.

3. Statistical Analysis of Bangla Graphemes

The statistical analysis of the Bangla script is meticulously done over a corpus of two hundred thousand words indigenously developed in our laboratory along with a list of twenty five hundred thousand words collected from the TDIL corpus of *DOE, Govt. of India*. The DOE text is compiled from the printed materials of different disciplines like literature, social science, natural science, commerce, and mass media published in between 1980-1990. It should be mentioned here that not all kinds of statistical analysis are informed in this paper. For the convenience of understanding we have presented only a few statistics along with some discussions in this section.

Char	%	Char	%	Char	%	Char	%	Char	%
অ	3.56	ঔ	0.03	ঞ	0.00	ন	4.64	শ	2.19
আ	4.85	ক	9.81	ট	0.63	প	8.68	ষ	0.05
ই	0.76	খ	0.99	ঠ	0.23	ফ	0.89	স	8.24
ঈ	0.05	গ	2.29	ড	0.43	ব	8.58	হ	4.45
উ	1.70	ঘ	0.66	ঢ	0.12	ভ	1.91	ড়	0.00
ঊ	0.02	ঙ	0.00	ণ	0.01	ম	4.63	ঢ়	0.00
ঋ	0.04	চ	1.99	ত	4.50	য	3.11	য়	0.13
এ	5.43	ছ	1.99	থ	1.21	র	2.13	ৎ	0.00
ঐ	0.18	জ	2.44	দ	4.47	ল	1.23		
ও	1.52	ঝ	0.13	ধ	0.72	ব	0.01		

Table 2: Words with a particular grapheme in first position

The Table 2 given above shows the number of unique words starting with a particular grapheme in the first position as found in the Bangla corpus. The table highlights that Bangla speakers feel comfortable to articulate words starting with velar, labial, sibilant or vowel sounds. That is why words starting with ক (k) is highest in number followed by that of প (p), ব (b), স (s), এ (e) and আ (A) in a sequential order. All these sounds, by nature, are easy to articulate because they take less energy or puff of air in articulation than other sounds. It is the first premise, designed statistically, to show that Bangla is easy to articulate and sweet to listen.

The Table 3 provides an interesting insight into the nature of the Bangla language. Out of 20 graphemes the number of vowel is 6, semivowel 1, liquid 2, nasal 2, stop 8, and fricative 1. The percentage of vowel includes both the percentage of their original and allographic forms. Almost all the graphemes are soft, mellow to listen and easy to articulate. This statistics easily establishes the common belief that the Bangla is virtually a soft and mellowed language, easy to utter and sweet to listen, and may be easier to learn than other languages.

Rank	Grapheme	%-age	Rank	Grapheme	%-age
1	আ	12.63	11	ম	02.62
2	এ	09.51	12	স	02.59
3	র	07.68	13	ও	02.68
4	ই	07.27	14	প	02.24
5	ন	04.46	15	দ	02.02
6	ক	04.12	16	য়	01.96
7	ল	03.52	17	ট	01.70
8	ব	03.48	18	ঈ	01.53
9	ত	03.20	19	জ	01.31
10	উ	02.67	20	গ	01.28

Table 3: Global percentage of grapheme occurrence in Bangla corpus

The Table 4 shows the occurrence of vowel is 41.69%, that of consonant is 51.70%, and that of consonant cluster is 06.61% in the said Bangla corpus. It shows that both vowel and consonant consist nearly 93.39% of the total graphemes used in the corpus. Though we have nearly 380 grapheme clusters in the language, the use of clusters in the corpus is quite less. If a page of a printed book contains 30,000 graphemes (having 300 words, each word containing 10 graphemes in average) then nearly 28000 graphemes are either vowel or consonant and the rests are clusters. It is noted that the percentage of cluster is

higher in similar contexts if the text is written in *sadhu* form (chaste version), which is older than the *calit* form (colloquial version), which is now in regular use in Bangla. We assume that the language is gradually becoming simplified as the consonant clusters are being removed from the regular use of the text.

Grapheme	%-age of use
Vowel graphemes	41.69 %
Consonant graphemes	51.70 %
Consonant grapheme clusters	06.61 %

Table 4: Percentage of grapheme occurrence

The Table 5 shows the percentage of allographs used in the corpus. It is found that the use of া (A) is the highest among all the allographs available in the language. A simple query confirms that the sound /a/ is the maximally used sound among the vowel sounds in regular Bangla speech. This observation is equally true for Hindi spoken corpus also (Khan, Gupta and Rizvi 1991). Next comes the allographs ে (e), ি (i), ু (u) and ো (o), respectively. The first two among these alographs are low and mid-high respectively in cardinal vowel diagram. This proves that Bangla native speakers prefer low or mid-high vowel sounds in speech to high or mid-low vowel sounds or diphthongs, the frequencies of use of which come at the end in the table.

Allograph	%-age of use	Allograph	%-age of use
া (A)	34.00	ী(I)	03.64
ে (e)	29.81	ূ (U)	00.99
ি (i)	18.56	ৃ (ri)	00.95
ু (u)	06.70	ৈ (ai)	00.30
ো (o)	04.72	ৌ (au)	00.28

Table 5: Percentage of use of vowel allographs in the Bangla corpus

The Table 6 given below shows which group of consonants are most frequently used in the language. It shows that in Bangla the use of alveolar and dental consonant graphemes is quite high. Next comes the labial followed by velar and nasal consonants, respectively. The maximum use of soft and liquid consonants in the language proves the lucidity and softness of the language.

Consonant Group	%-age of Use	Consonant Group	%-age of Use
Alveolar Group	24.81	Nasal Group	10.78
Dental Group	18.78	Palatal Group	07.59
Labial Group	16.83	Retroflex Group	06.23
Velar Group	12.36	Glottal Group	02.62

Table 6: Percentage of use of consonant groups in the Bangla corpus

The Table 7 shows the most frequently used consonant clusters as found in the Bangla corpus. It shows that the cluster প্র (pr) is maximum in use followed by others. It is highest in use because both the consonant grapheme প (p) and the consonant modifier ্র (ra-phalA) are mostly used characters in the corpus text. In Bangla primary text books, however, the cluster ক্ষ (kS) is considered as a unique consonant grapheme. Perhaps, its unique combination and high frequency of use have motivated the script designers to consider it as a basic character. The same case does not happen for প্র (pr), because it is a cluster which made with a kind of modifier ্র (ra-phalA) which, unlike ষ (S), is used with almost all other consonant graphemes. Probably, for this reason it is not considered as a unique grapheme.

Cluster	%-age of use	Cluster	%-age of use
প্র (pr)	08.16	ন্য (ny)	01.99
ক্ষ (kS)	03.94	স্থ (sth)	01.98
ন্ত (nt)	03.51	গ্র (gr)	01.87
ঙ্গ (ngg)	02.50	ব্য (by)	01.73
ত্র (tr)	02.48	জ্ঞ (jñ)	01.58
ন্দ (nd)	02.46	চ্ছ (cch)	01.52
ক্ত (kt)	02.33	র্থ (rth)	01.48
স্ত(st)	02.27	ন্ন (nn)	01.48
স্ব (sv)	02.06	দ্ধ (ddh)	01.47
ষ্ট (ST)	02.01	শ্য (shy)	01.31

Table 7: Percentage of use of consonant clusters in Bangla corpus

A question may raised in this context: why some consonant modifiers like ্র (ra-phalA), ্য (ya-phalA), র্ (reph), ্ব (va-phalA) etc. are specially designed in the script for cluster formation? The counts taken from the corpus (as shown in Table 8) show that the use of *ra-phalA, ya-phalA, reph* and *va-phalA* are very high in the corpus. The total percentage of their use is 46.63% whereas the total percentage of use of other consonant clusters in the corpus is 53.37%. This count supports our assumption that because of their frequent

use of in texts their unique form is designed to make the act of writing easy and simple. There is also a possibility that these modifiers can be considered as unique consonant grapheme in the script in near future as it happens for the cluster of ক্ষ (kS).

Consonant Modifiers	%-age of use
্র (ra-phalA)	18.97
্য (ya-phalA)	13.33
র্ (reph)	11.35
্ব (va-phalA)	02.98
Others	53.37

Table 8: Percentage of use consonant modifiers in the Bangla corpus

4. Structural Analysis of Graphemes

In a running Bangla text one can trace four types of grapheme:

(a) Vowel and consonant graphemes (basic characters),
(b) Vowel allographs and consonant modifiers,
(c) Compound graphemes (when a vowel allograph changes shape of a consonant grapheme), and
(d) Consonant grapheme clusters.

In the following subsections the basic characters are dissected in isolation to find out the unique properties by which a grapheme is different from the other. Modifications and changes are also noted whenever these isolated graphemes are used in running texts.

4.1 Graphemes in Isolation

The structure of the basic Bangla graphemes is a mixture of straight lines, circular and semi-circular curves, thick dots, and conic shapes - normally known as glyphs. All these glyphs are not of equal size and length and each glyph is not used in its full length in every occasion. Sometimes the full length of a glyph, sometimes the half of it and even sometimes just a portion of the glyph is used for designing the basic graphemes. The arrangement of these glyphs is not complex like that of Dravidian scripts. However, the physical shape of some graphemes like ঈ (I), ঋ (r), খ (kh), গ (g), ঘ (gh), ঙ (ng), ছ (ch), জ (j), ঞ (ñ), থ (th), ফ (ph), ভ (bh), শ (sh) and স (s) is more complex in form than other graphemes. The reason of their complexity might be due to use of dots, curves, straight lines, and conic glyphs in their shape formation.

The *shirorekhA* (i.e., headline) is considered as an important feature as it can act as a line of demarcation at the time of *tier division* of basic characters. It is the main feature by which the basic graphemes can be grouped into two broad classes:

(i) Graphemes with *shirorekhA* (32 in number), and
(ii) Graphemes without *shirorekhA* (14 in number).

According to the arrangement of different glyphs the basic graphemes can be grouped into three major classes:

(i) Graphemes made with linear structures arranged in different angles (15 in number)
(ii) Graphemes made with dot and curve shapes (11 in number), and
(iii) Graphemes made with both kinds of shape (26 in number).

The use of vertical line is maximum in formation of basic graphemes. Nearly 33 basic graphemes have vertical lines in full span with them. The graphemes such as আ (A), ঋ (r) and ঝ (jh) have used this vertical line twice - the second line is placed just parallel to the first line. For most of the graphemes this vertical line is at the right most side of the grapheme, as in, ন (n), ব (b), etc. However, there are some graphemes in the script like চ (c), ছ (ch), ট (T), ঢ (Dh), ঢ় (Rh), etc. for which this vertical line is situated on their left most side. Graphemes such as উ (u), ঊ (U), ঙ (ng), ছ (ch), ড (D), দ (d), ড় (R) etc. have half length vertical line in their shape design.

The width of a grapheme is not always proportionate to it height. For some graphemes the width is more than height, as in, গ (g), ল (l), শ (sh), etc. For some other graphemes width is less than height, as in, ণ (N), ন (n), etc. . Finally, for other graphemes the width is nearly the same to height, as in, ক (k), ব (b), র (r), etc.

For automatic character recognition, the information of shape analysis of each grapheme is essential. In its first step it is noted that some graphemes are extremely similar in shape to other graphemes, such as, অ (a) is similar to আ (A) except the vertical line with the later; উ (u) is similar to ঊ (U) except that extra curve with the later; ও (o) is identical with ত্ত (tt) except that *shirorekhA* (headline) on the later; খ (kh) is almost same with থ (th) except that initial extra fold; ক (k) is almost identical with ফ (ph) except the gap on the right-most vertical line; ত (t) is similar with ভ (bh) without the initial fold for the later; ল (l) is same with ন (n) except the an extra fold with the first one; শ (sh) is same with ণ (N) except that extra loop with the first one; এ (e) and ত্র (tr) are same except the

upper line with the later; ব (b), ড (D), ঢ (Dh) and য (y) are same with র (r), ড় (R), ঢ় (Rh) and য় (y), respectively, without the dot just below the later characters; গ (g) is almost same with প (p) except that the last one does not have the short slanted line connecting the front end with the upper end of the vertical line; ঘ (gh) is same with ষ (S) except that slanted line that runs through the middle of the later.

Among the consonant grapheme clusters the forms ক্ষ (kS) is nearly similar to হ্ম (hm) except the loop which hangs on the right hand side of the right-hand vertical line of the later character, etc. These graphemes are considered to be confusing graphemes as one grapheme can be confused with the other in shape easily by man and machine.

The allographs of vowel graphemes, when these are used with consonant graphemes or clusters are distributed in all three tiers. Some are distributed between upper and middle tier, some are used only in lower tier, while some are distributed only on the middle tier (see Sub-section 4.3). The reason for formation of these allographs may be to reduce the recurrent use of vowel graphemes after the consonant graphemes and clusters in words. The vowel graphemes, in comparison with their respective allographs, usually take more time, space, and energy in writing. So the script designers, thinking that the use of an allograph can be the best possible option for relieving a writer from the extra burden of repetition, may have designed the allographs. It is observationally justified that the most recurrently used allograph is most simple in shape and most suitably positioned in the Bangla writing system.

4.2 Graphemes within Words

Sometimes some graphemes when used within words differ from their features noted in isolation. On the contrary it can be said that their contexts can add some more features which are not noted in some graphemes in their isolation. Moreover, these graphemes can have some restrictions in their positional use; can have some modifications in their original shape and size; and also can have some limitations in their functional role in the strings, etc. For example, the vowel graphemes in their original forms are mostly used at word-initial position. They can, however, be used at word-final position but at that context they mostly function as emphatic markers, as in, কলমই (kalamai) "the pen itself", তুমিও (tumio) "you too", etc. Very rarely they are found to be used at the word-middle position, such as, চাঅলা (cAalA) "tea vendor", অতএব (ataeb) "therefore", মাঈজি (mAIji) "mother", মউতাত (mautAt) "relish", etc., and mostly in case of transliterated foreign words, such as, জানুআরি (jAnuAri) "January", ওআটার (oATAr) "water", আইন (Ain) "law", etc.

Among consonant graphemes, ঙ (ng), ঞ (ñ), ড় (R), ঢ় (Rh), য় (y) and ৎ (t) cannot occur at word-initial position because in a normal situation it is quite difficult for a Bengali speaker to articulate a word starting with any one of the consonant graphemes.

The consonant grapheme ত (t) has a *modifier,* namely, ৎ (t) (khaNData), which cannot use vowel allograph. Generally, it occurs at the word-middle and word-final positions. However, when there is a need to use a vowel allograph with this modifier, particularly when a case marker is added to it, it changes into the original grapheme ত (t) because the modifier cannot carry the load of the vowel allograph, as in,,মহৎ (mahat) "great" but মহতের (mahater) "of great", ভবিষ্যৎ (bhabiSyat) "future" but ভবিষ্যতের (bhabiSyater) "of future", etc.

The consonant grapheme র (r) has two distinct graphic modifiers which occur at the time of cluster formation. One is the ◌ঁ (reph) which is placed in the upper tier just above the consonant grapheme and which cannot cause any structural change of the consonant grapheme. The other one is ◌্র (ra-phalA) which is placed at the lower tier just below the consonant grapheme. In some occasions it can cause change in the original shape of the grapheme in the middle tier as has happened for the consonant clusters like ক্র (kr), ত্র (tr), ভ্র (bhr), etc.

To understand the actual behaviour of the graphemes used in the Bangla script we need to scrutinize their two important criteria within a word string:

(a) Tier division of characters, and
(b) Ability of compound grapheme formation.

Both the processes (i.e., tier division and compound grapheme formation) are necessary and useful information for proper identification and recognition of each grapheme and for identification of the methods used in compound grapheme formation.

4.3 Tier Division of Graphemes

In a running text the Bangla graphemes are arrayed in three tiers: upper tier, middle tier and lower tier. The upper tier generally contains the signatures of the basic graphemes and allographs along with some consonant modifiers like *candrabindu* and *reph.* The middle tier virtually contains the bulk of the graphemes' weight and the lower tier carries some allographs and consonant modifiers like *ra-phalA* and *va-phalA*. The graphic representation of tier division of Bengali graphemes are given below (Fig. 1).

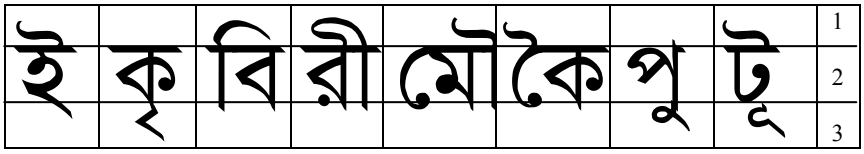

Fig.: 1: Tier-division of Bengali graphemes and allographs
(1: Upper Tier, 2: Middle Tier, 3: Lower Tier)

The above diagram clarifies the concept of tier division. Some times the in the middle tier the vowel allographs are also accommodated. This division is required for grapheme recognition and for structure analysis. Moreover, for automatic grapheme recognition by computer, this tier division helps to identify a single grapheme in a string of multiple different graphemes.

4.4 Compound Graphemes

For the convenience of our discussion we call them compound graphemes which are formed by physically merging two or three graphemes together. They can be vowel and consonant graphemes as well as vowel allographs and consonant modifiers. In this process of formation some changes usually take place in the original structure of the participating graphemes. Moreover, the change takes place only in the middle tier mentioned above. Compared to the basic graphemes these graphemes are complex in structure. For example, the allograph of the vowel grapheme উ (u) when used with a basic grapheme can generate three different compound shapes which are grapheme dependent, as described below:

First, the allograph of the vowel grapheme উ (u) takes a shape like ু while it is attached in the right hand side of the grapheme র (r) giving a final shape like রু (ru). The notable point is that the change takes place only with this particular grapheme র (r), either in its original shape or in its is্র (ra-phalA) version with other consonant grapheme cluster like দ্রু (dru), গ্রু (gru), শ্রু (shru), ব্রু (bru), etc. This shape of the allograph is similar to that of allograph in Telugu and Kannada script. It has probably come into Bangla from Telugu or Kannada as an outcome of cultural fusion between Bengal and South India.

Second, the allograph of the vowel grapheme উ (u) takes a shape like ু while it is attached at bottom of the consonant graphemes শ (sh) and গ (g) and the cluster ন্ত (nt) to generate final shapes like শু (shu), গু (gu) and ন্তু (ntu), respectively. This form of the allograph is similar to that of Devnagari and Gujarati script.

Last, the consonant grapheme হ (h) takes the shape like হু (hu) forcing the allograph to merge with the grapheme. This variation is noted only with this particular consonant grapheme which has no parallel form in any of the Indian scripts.

The allograph of long vowel grapheme ঊ (U) also go through structural change when it is used with the consonant grapheme র (r) and in clusters with র (r) at the final position, such as, গ্র (gr), শ্র (shr), etc. The allograph changes thoroughly into a shape like ৃ and is attached on the right hand side of the grapheme or cluster as in রূ (rU), গ্রূ (grU), শ্রূ (shrU), respectively. It is noted that the deformed allograph of this form is also similar to that of Kannada and Malayalam script.

The allograph of the vowel grapheme ঋ (r) goes through directional change (not the structural change) when used with the consonant grapheme হ (h). Here the allograph, changing its direction from horizontal to vertical is attached to the right hand side of the grapheme as in হৃ (hr).

Some consonant graphemes, when join physically with other consonant graphemes can form consonant clusters. At the time of cluster formation the participating graphemes may undergo three types of structural changes. Moreover, as said before, these changes are noted only in the middle tier of the graphemes.

First, primary shape of the participating consonant graphemes are thoroughly changed thereby forming a new compound shape, such as, ক্ষ (kS), ক্ত (kt), ক্র (kr), ঙ্গ (ngg), ঞ্চ (ñc), ত্ত (tt), ত্র (tr), and হ্ম (hm) (8 in number). In such case, it becomes almost impossible to trace out the original shapes of the participating graphemes.

Second, original shapes of the participating graphemes are partly modified. This can be either on both or on one of the participating graphemes. It is counted that for nearly 65 clusters, the shape of the first grapheme is affected where as there are nearly 90 clusters where the shape of the last grapheme is affected. The reasons of such differences may be that in the first occasion the phonetic property of the second grapheme of the cluster holds the importance in articulation whereas in the second occasion the process becomes just the reverse.

Last, for clusters of three and four graphemes (around 30 and 10, respectively) there is virtually no change in final form for the first two graphemes. The last grapheme of the cluster is placed either below or on the right hand side of the immediately preceding grapheme, generally in the middle tier.

In this context it should be mentioned that there are a few phonetic clusters in the Bangla language, such as, /Tl/ /Dl/, /tl/, etc. for which there is graphemic representation in the script. These are not discussed here as we intend to deal with the graphemic clusters that are available in the printed script.

Some compound graphemes are modified for the purpose of transparency as well as for easy access in typewriter and computer implementation. However, these modifications are not universally accepted by all printing organizations using the Bangla script. So in the corpus both old and new shapes of compound graphemes are available almost in equal proportion. In the following Table 9 we have shown some opaque shapes and their respective transparent shapes for compound grapheme design.

Opaque Shape	Transparent Shape	Opaque Shape	Transparent Shape
গু (gu)	(gu)	শু (shu)	(shu)
সু (su)	(su)	হু (hu)	(hu)
হ্র (hr)	(hr)	হ্ন (hn)	(hn)
ন্তু (ntu)	(ntu)	রু (ru)	(ru)
রূ (rU)	(rU)	স্ত (st)	(st)

Table 9: Opaque and transparent shapes of some compound graphemes in Bangla

5. Variations in Utterance

To locate different utterance variations of the graphemes within a word some utterance rules are described by earlier scholars. Among them Rabindranath Tagore (1995), Jamil Chaudhury (1990), Pabitra Sarkar (1992), Subhas Bhattacharya (1992), Enamul Haque (1995), Mahbabul Haque (1995), Punya Sloka Ray (1997), Paresh Chandra Majumdar (1998) are notable. Besides, some efforts are made for utterance regularization of Bangla words by Calcutta University (1936/37), Paschimbanga Bangla Academy (1992), Ananda Bazaar Patrika (1994), Bangiya Sahitya Parisad (1986) and Bangla Academy of Bangladesh (1993). For our analysis and observation the utterance pattern around Calcutta is considered as the standard one. In case of confusion, utterance dictionary (Bhattacharya 1993) as well as some experts in the field is consulted.

The Bangla vowels, allographs, consonant, consonant modifiers and consonant clusters constitute more than 300 unique graphemic forms. It has eight vowel graphemes: অ (a), আ (A), ই (i), ঈ (I), উ (u), ঊ (U), এ (e) and ও (o) to represent seven vowel sounds: /O/, /a/, /i/, /æ/, /e/, /o/, and /u/. According to their articulations these graphemes as well as their respective allographs can be grouped in the following way:

(a) One vowel grapheme denotes two vowel sounds: অ (a) denotes /O/ and /o/; আ (A) (and allograph) denotes /a/ and /æ/; and এ (e) (and allograph) denote /e/ and /æ/.
(b) Two vowel graphemes denote one vowel sound: ই (i) and ঈ (I) (along with their allographs) denote /i/; উ (u) and ঊ (U) (along with their allographs) denote /u/.
(c) O vowel grapheme denotes one vowel sound: ও (o) (and its allograph) denotes /o/.

In major cases, articulation problem arises due to absence of an allograph for the vowel grapheme অ (a). Data shows that the vowel grapheme is used, if required, only at the initial position of a word. It is sometimes articulated as /o/ creating confusion in its primary utterance. Moreover, all non-allographed consonants and clusters in words can be either articulated with /O/ or /o/ or may be simply non-vocalic. As a result, an unaccustomed reader does not know which consonant or cluster should be articulated with /O/ or /o/ or should be non-vocalic in utterance. In a similar manner, at word-initial position the utterance of the vowel grapheme এ (e) is either /e/ or /æ/. Moreover, there are utterance variations of the consonant clusters. Thus, there are four major utterance variations in Bangla script, namely:

(a) if a consonant grapheme is vocalic or not;
(b) if vocalic then whether it is uttered with /O/ or /o/;
(c) if the vowel grapheme এ (e) (and its allograph) is uttered as /e/ or /æ/, and
(d) if the consonant grapheme or cluster is uttered with /O/ or /o/ sound or it is simply non-vocalic in utterance.

In the Bangla corpus, in a rough count, there are nearly 5,000 words which are formed without the use of any vowel allograph. These words are formed either by combining vowels and consonants or by combining consonant graphemes (and clusters) only. It is, therefore, difficult to determine the utterance of such words as it is difficult to determine which character is vocalic and which character is not, and whether the character is vocalic with /O/ or /o/ sound. Positional occurrence of the characters can determine their actual utterance. The following observations may be made for non-allographed words:

(i) A non-allographed consonant or cluster at word-initial position is usually vocalic with /O/ or /o/ sound.
(ii) A non-allographed consonant is mostly non-vocalic at word-final position except those word that end in ত (t).
(iii) For non-allographed words the normal vocality pattern is: v~n~v~n (where /v/ denotes vocalilty and /n/ denotes non-vocality).
(iv) The normal vocalization pattern is: /O~o~O~o~O~o/ or vice versa.

It is already noted that the Bangla vowel grapheme অ (a) has no allograph. Initially, the vowel sound denoted by this grapheme was inherent with all consonant graphemes and clusters (both in isolation and within words). Therefore, at the time og articulation, all consonant graphemes and clusters are vocalic if not specified in other way. So it is quite rational not to design an allograph for the vowel grapheme অ (a) when other vowel graphemes have one each. Whenever a consonant or a cluster without a vowel allograph appears in texts, by default it is considered to be vocalic in sound. However, in present day utterance we have many consonant graphemes and clusters within words which are sometimes vocalic and sometimes not. Such differences create problem in determination of vocality of consonant graphemes and clusters in words. Keeping these factors in mind we have tried to define articulation patterns of some of the vowel graphemes, consonant graphemes and clusters that may posit difficulties in utterance.

The utterance of আ (A) and its allograph (া) is mostly /a/ in all positions of a word. However, at certain contexts, the allograph is uttered as /æ/ if used immediately after the cluster জ্ঞ (jñ) in a word. We have found some words where the grapheme আ (A) is replaced by its allograph preceded by the semi-vowel grapheme য় (y), as in, বে-আড়া (be-ARA) > বেয়াড়া (beyARA) "obstinate", দো-আব (do-Ab) > দোয়াব (doyAb) "basin", বে-আদপ (be-Adap) > বেয়াদপ (beyAdap) "obstinate", etc. Such replacement takes place because the grapheme আ (A) is not generally used at word-middle position in Bangla. On the other hand, the use of য়া (yA) at this position is very common in Bangla. Moreover, both are similar in pronunciation. So there is no hesitation in replacing আ (A) by য়া (yA). However, as exceptions, we have found a few words in the corpus where the grapheme আ (A) is used at word-medial position, as in, জানুআরি (jAnuAri) "January", ওআটার (oATAr) "water", বেআইনী (beAinI) "illegal", বেআক্কেলে (beAkkele) "foolish", etc. These are mostly transliterated foreign words.

The grapheme এ (e) and its allograph has two utterance variations: /e/ and /æ/. For the Bangla tongue it is always easier to glide from /e/ to /æ/ than from /e/ to /a/. Thus, to relieve our tongue we replace /e/ before /a/ by /æ/ (Tagore 1995: 25). This is, however, noted only at word-initial position. At other position, irrespective to any context, the vowel grapheme or its allograph is always uttered as /e/. Other vowel graphemes and their allographs show no variation in utterance, although, a negligible variation of length (short and long) can be noted for /i/ and /u/.

The number of consonant graphemes (35) used in Bangla script is slightly more than the number of consonant sounds (30) found in the language. That means some consonant

graphemes are identical in articulation. For instance, consonants like জ (j) and য (y) are almost similar in articulation; so do শ (sh), ষ (S) and স (s); ণ (N) and ন (n); ত (t) and ৎ (t) etc.

Consonant graphemes are usually vocalic in isolation but when an allograph is attached with them they usually drop their inherent vocalic properties to take up that particular vowel sound the allograph denotes. However, the consonant graphemes হ (h) and ঢ় (Rh) (except আষাঢ় (ASARh) "rainy season") are always vocalic, while consonant grapheme ৎ (t) (khaNData) is always non-vocalic. Among consonant modifiers ঁ (candrabindu) and ঃ (bisarga) are always vocalic while ং (anusvAr) is non-vocalic.

The existence of yuktabyAñjanvarNa (consonant grapheme clusters) in Bangla script can be traced back in the Brahmi script, which is considered as the mother of most of the modern Indian scripts (Chatterji 1995: 56). The use of yuktabyAñjanvarNa is still prevalent among most of the modern Indian language scripts including Bangla, Oriya, Assamese, Devnagari etc. The Bangla script includes a large set of consonant clusters which are formed by joining two or more consonant graphemes. Such physical merging does not happen for vowel graphemes. The reason behind formation of such graphemic clusters (eye cluster) can be traced in phonetic clusters (ear cluster), which occur quite frequently in normal Bangla speech. In a random speech sequence, it is observed, two or more consonant sounds (mostly at *coda* and *onset* of syllables) are combined together to produce phonetic clusters, many of which are eventually realised as graphemic clusters. Such graphemic clusters have now become an integrated part of the Bangla script and writing system, and due to this reason they carry great importance in language learning, text to speech conversion, OCR development and spelling checker system designing in Bangla. A study on their formation and function can throw some light into the linguistic behaviour of the native Bengali speakers.

We have found some discussions on the articulation patterns of consonant clusters in Bangla (Bhattacharya 1993: 11-13, Sarkar 1994: 31-46, Ray 1997: 14-16) which have given us some insights about the overall scenario of consonant cluster articulation. In the following sections we have tried to give a more systematic study of the issues and problems of this area. Out of total set of consonant grapheme clusters available in the language, majority of the clusters are phonemic in nature because they follow the general norms of utterance of the graphemes. But, there are some clusters which deviate from the standard norms and create problems at the time of utterance. Only these clusters are discussed here.

For consonant clusters the general observation is that the *cluster-final* consonant is always vocalic. A *cluster-final* consonant is one which occurs as the last member of a cluster. At the time of utterance the characters generally follow the sequence of their occurrence. However, for some clusters the sequence is slightly changed. Besides, some modifications (e.g., deletion, addition and displacement of sound) in articulation also occur due to contextual use of characters as discussed below:

At word-final position a non-allographed consonant cluster is always vocalic with /o/, e.g., কলঙ্ক (kalanka) "defame", গল্প (galpa) "story", অব্দ (abda) "year", আরম্ভ (Arambha) "beginning", বরঞ্চ (barañca) "rather", প্রকাণ্ড (prakANDa) "huge", বরাদ্দ (barAdda) "allotted", বিদগ্ধ (bidagdha) "wise" etc. However, for some borrowed foreign words the word-final non-allographed cluster is non-vocalic in utterance, as in, আর্ট (ArT) "art", আগস্ট (AgaST) "August", এলার্ম (elArm) "alarm", গার্ড (gArD) "guard", বাল্ব (bAlb) "bulb", সন্স (sans) "sons", ফিল্ম (philm) "film", হিন্দ (hind) "Hind", গাল্ফ (gAlph) "gulf", হর্ন (harn) "horn", নার্স (nArs) "nurse", etc.

In case of the cluster স্ক (sk) the sequence of the characters, due to anaptyxis, is trans-positioned in utterance. Thus, works like বাক্স (bAksa) "box" and রিক্সা (riksA) "ricksaw" are sometimes uttered as /basko/ and /riska/, respectively.

In case of the cluster ক্ষ (kS) both the characters lose their respective individual utterance to produce two utterance variations:

(a) At word-initial position it is uttered as /kh/, as in, ক্ষত (kSata) "wound", ক্ষার (kSAra) "alkaline", ক্ষুর (kSur) "hoof", ক্ষেত (kSet) "field", ক্ষোভ (kSobh) "anger", ক্ষতি (kSati) "loss", etc.
(b) At other positions it is uttered as /^{k}kh/, as in, অক্ষ (akSa) "orbit", কক্ষ (kakSa) "chamber", চক্ষু (cakSu) "eye", পক্ষ (pakSa) "wing", পক্ষি (pakSi) "bird", সক্ষম (sakSam) "able", অক্ষম (akSam) "unable", etc.

The cluster জ্ঞ (jñ) has also two utterance variations within words. In both cases the first consonant grapheme জ (j) loses its own utterance to be pronounced as /g/:

(a) At word-initial position the consonant grapheme জ (j) is uttered as /g/, as in, জ্ঞান (jñAn) "knowledge", জ্ঞানত (jñAnata) "in sense" etc.

(b) At other positions the consonant grapheme জ (j) is uttered as /^{g}g/, as in, অজ্ঞ (ajña) "idiot", বিজ্ঞ (bijña) "wise", অভিজ্ঞ (abhijña) "experienced" etc.

In case of cluster ঞ্চ (ñc), ঞ্ছ (ñch) and ঞ্জ (ñj), the first consonant ঞ (ñ) is uttered like dental ন (n), as in, অঞ্জন (añjan) "eye-salve", কাঞ্চন (kAñcan) "gold", বাঞ্ছা (bAñchA) "wish", etc. However, in case of clusters made with reverse arrangement of consonants, as in, ঞ্চ (cñ) and জ্ঞ (jñ), it loses its own utterance to nasalize its immediately preceding character, as in, যাচ্ঞা (yAcñA) "want", অজ্ঞান (ajñAn) "senseless" etc.

In clusters of শ্ম (shm) and স্ম (sm) at word-initial position the utterance of ম (m) is lost to nasalize the preceding character, as in, শ্মশান (shmashAn) "burning ghat", শ্মশ্রু (shmashru) "beard", স্মর (smar) "cupid", স্মরণ (smaraN) "remember", স্মারক (smArak) "memento", স্মের (smer) "smiling" etc. A few exceptions are noted where ম (m) has retained its own utterance, as in, স্মিতা (smitA) "smiling" etc.

In case of clusters of ত্ম (tm), দ্ম (dm), স্ম (sm) and শ্ম (shm) at word-medial and word-final positions, the utterance of ম (m) is lost to nasalize and double the utterance of the preceding character, as in, আত্মা (AtmA) "soul", পদ্ম (padma) "lotus", রশ্মি (rashmi) "rays", গ্রীষ্ম (griSma) "summer", বিস্ময় (bismay) "surprise" etc. There are, however, some exceptions where the consonant ম (m) is distinctly uttered with its preceding character, as in, অস্মিতা (asmitA) "selflessness", কাশ্মির (kAshmir) "Kashmir", কুষ্মাণ্ড (kuSmANDa) "pumpkin" etc.

In cluster of ক্ষ্ম (kSm) the utterance of ম (m) is totally lost, as in, লক্ষ্মী (lakSmI) "Laksmi", পক্ষ্ম (pakSma) "eye lash", সূক্ষ্ম (sUkSma) "fine" etc. On the other hand, in case of clusters of গ্ম (gm), ক্ম (km), ল্ম (lm), ন্ম (nm), the utterance of ম (m) is retained mostly unaffected, as in, যুগ্ম (yugma) "two", রুক্মিনী (rukminI) "a name", গুল্ম (gulma) "shrub", জন্ম (janma) "birth" etc.

For the clusters of হ্ন (hN) হ্ন (hn), হ্ম (hm) and হ্ল (hl), the actual orthographic sequence of occurrence of the characters in words is just reversed in utterance. That means while their orthographic pattern is C_1C_2, their articulatory pattern is C_2C_1, as in, অপরাহ্ন (aparAhNa) "afternoon", চিহ্ন (cihna) "sign", ব্রহ্ম (brAhma), "Brahmin", জহ্লাদ (jahlAd) "executioner", etc.

The labio-velar ৎ (va-phalA) as a cluster-final member modifies the utterance of clusters in three ways, as shown below:

(a) At word-initial position its utterance is entirely lost, as in, জ্বর (jvar) "fever", ত্বক (tvak) "skin", দ্বীপ (dvIp) "island", শ্বাপদ (shvApad) "beast", (svapna) "dream" etc.
(b) At word-middle and word-final position, its own utterance is lost to double the utterance of the preceding character, as in, পক্ব (pakva) "ripe", সত্ব (satva) "right", বিল্ব (bilva) "a kind of wood apple", বিশ্বাস (bishvAs) "faith", বিদ্বান (bidvAn) "learned", etc.
(c) In case of cluster of হ্ব (hv) at word-middle and word-final position, it generates /bh/ sound, as in, আহ্বান (AbhAn) "invitation", বিহ্বল (bihval) "exulted", গহ্বর (gahvar) "hole", etc. (Sarkar 1994: 43).

Due to its orthographic similarity with the labial consonant grapheme ব (b), it is used at the same places within words where the bilabial ব (b) is normally used as a cluster-final member, as in, বাল্ব (bAlb) "bulb" and বিল্ব (bilva) "a kind of wood apple", উদ্বেগ (udbeg) "anxiety" and বিদ্বান (bidvAn) "wise", etc. In these cases, it is difficult to determine if the character is to be articulated or not because while bilabial ব (b) is always articulated the labio-velar ৎ (va-phalA) is always silent in utterance. We need both etymological and semantic information along with native language intuition to determine the utterance of the character. This information is handy for developing systems for text-to-speech conversion and language teaching.

The modifiers *reph* and *ra-phalA* of the consonant র (r) are always used with a character within words. However, while *reph* occurs only at word-medial and word-final position, *ra-phalA* can occur at all three positions of words. These modifiers can cause three types of utterance variation, as noted below:

(a) A non-allographed consonant grapheme with *reph* at word-middle and word-final position is a-ending or o-ending in utterance, as in, অর্ক (arka) "sun", কর্ম (karma) "worker", গর্ব (garba) "proud", বর্জন (barjan) "discard", সর্প (sarpa) "snake" etc. Exception is noted in some borrowed words where the utterance of the consonant is non-vocalic, as in, আর্ট (ArT) "art", গার্ড (gArD) "guard", হর্ন (harn) "horn", নার্স (nArs) "nurse", এলার্ম (elArm) "alarm" etc.,
(b) A non-allographed consonant with *ra-phalA* at word-initial position is normally uttered with /o/ sound, as in, ক্রমশ (kramasha) "gradual", গ্রহ (graha) "planet", প্রণয়

(praNay) "love", ব্রত (brata) "rites", ভ্রম (bhram) "illusion", শ্রম (shram) "labour", স্রবণ (srabaN) "secretion" etc. However, some exceptions are found where a non-allographed consonant with *ra-phalA* at word-initial position is uttered with /O/ sound, as in, হ্রদ (hrad) "lake", ক্রয় (kray) "buy" etc.

(c) A consonant with *ra-phalA* at word-medial and word-final position is doubled in utterance, as in, অঘ্রাণ (aghrAN) "a Bangla month", আব্রু (Abru) "cover", বক্র (bakra) "crooked", পত্র (pAtra) "bride", শত্রু (shatru) "enemy", সক্রিয় (sakriya) "active", বিগ্রহ (bigraha) "idol" etc.

The modifier *ya-phalA* occurs with a character at all positions within words. However, depending on its position in a word it varies in utterance. At word-initial position it has three utterance variations:

(a) With a consonant at word-initial position it has no utterance, as in, চ্যবন (cyaban) "name", দ্ব্যর্থক (dvyarthak) "ambiguous", দ্যুতি (dyuti) "glow", চ্যুত (cyuta) "expelled", ব্যোম (byom) "ether" etc. There is an exception, like ব্যক্ত (byakta) "expressed" etc.

(b) With a consonant tagged with the allograph of আ (A) it is uttered as /æ/, as in, ক্যাবলা (kyAblA) "un-smart", চ্যালা (cyAlA) "follower", ব্যামো (byAmo) "illness", ঠ্যালা (ThyAlA) "push", ল্যাজা (lyAjA) "tail" etc.,

(c) With a non-allographed consonant it is uttered as /e/ if its following character is tagged with /i/, as in, ব্যক্তি (bykti) "person", ব্যতিক্রম (bytikram) "exception", ব্যতিরেক (bytirek) "difference", ব্যাতীত (byAtIta) "except", ব্যাথিত (byAthita) "hurt", etc. However, there are some exceptions, where with some non-allographed consonants it is uttered as /æ/, as in, ব্যভিচার (bybhicAr) "lechery", ব্যয়ী (byayI) "expensive" etc.

At word-middle and word-final position a non-allographed consonant with *ya-phalA* is almost doubled in utterance, as in, কল্য (kalya) "tomorrow", শস্য (shasya) "food grains", পদ্য (padya) "poetry", সভ্যতা (sabhyatA) "civilization", যোগ্যতা (yogyatA) "ability", মধ্যম (madhyam) "middle", হত্যা (hatyA) "murder" etc. However, the doubling of sounds is of two types as shown below:

(a) The unvoiced aspirate consonant will generate unvoiced unaspirate consonant plus unvoiced aspirate consonant, such as, মুখ্য (mukhya) "main", পাঠ্য (pAThya)

"syllabus", পথ্য (pathya) "food for patient" etc. This process can be explained by the following rule:

Orthography		**Utterance**	
C + ya-phalA	→	C	C
[-voice]		[-voice]	[-voice]
[+aspirate]		[-aspirate]	[+aspirate]

(b) The voiced aspirate consonant will generate voiced unaspirated consonant plus voiced aspirate consonant, as in, ধনাঢ্য (dhanADDhya) "rich", বাধ্য (bAdhya) "forced", সভ্য (sabhya) "civilized" etc. This process of change can be explained by the following rule:

Orthography		**Utterance**	
C + ya-phalA	→	C	C
[+voice]		[+voice]	[+voice]
[+aspirate]		[-aspirate]	[+aspirate]

(c) The modifier *ya-phalA* with the consonant হ (h) at word-medial and word-final position will generate /jjh/ sound, as in, দাহ্য (dAhya) "inflammable", সহ্য (sahya) "tolerate", বাহ্যিক (bAhyik) "external", গ্রাহ্য (grAhya) "care", মুহ্যমান (muhyamAn) "morose" etc.

The functional roles and linguistic importance of *candrabindu*, *anusvAr* and *bisarga* are primarily contextual. When these are detached from contexts they lose their independent entity in the language.

6. Conclusion

Researchers studying the evolution of thought processes in human societies believe that development of language and script may also influence the cognitive powers of the members of a speech community. Since script is a form of knowledge representation, the use of alphabets makes demands on humans to code and decode knowledge, convert auditory sounds into visual symbols, think deductively and order words to construct sentences ["Language Instinct": *Know-how*: The Telegraph: *9th Feb.,1998*].

The script of a language is a form of knowledge representation. In the first few sections of the paper we have tried to understand form and structure of some Bangla characters.

The study is important and useful in problems related to spelling correction, speech recognition, computational linguistics, character recognition, text preparation, language teaching and cryptography, etc. The study of the Bangla script is not an exception. Even from pure applied point of view this study can help primary and secondary language learners and to know how Bangla script is designed and used in a running text.

In the last section of the paper we have tried to trace utterance peculiarities of Bangla graphemes with an intention to use these for NLP works. It is to be noted that some surface word forms, having identical grapheme arrangements, may be uttered differently either for difference of meaning, or for difference of lexical category or for imitation of utterance of foreign word forms. Moreover, a fixed sequence of grapheme arrangement in the words does not guarantee the similar sequence of utterance. So, intimate study of utterance of each grapheme and allograph is important for proper understanding the utterance of words. The study is important in automatic speech recognition, speech synthesis, etc.

Acknowledgment: We acknowledge help of Department of Electronics (DOE), Govt. of India for providing the Bangla corpus our study presented here. Discussions with Prof. Mrinal Kanti Nath, Dept. of Linguistics, University of Calcutta is acknowledged with thanks.

Reference

Banerjee, Chittaranjan (1981) (Ed) *Dui shataker Bangla Mudran o Prakashan* (Bangla Printing and Publication in Two centuries). Calcutta: Ananda Publishers.

Banerjee, Rakhaldas (1919) *The Origin of the Bengali Script.* Calcutta: Calcutta University Press. Reprinted by Nababharat Publisher, Kolkata in 1973.

Bhattacharya, Nikhilesh (1965) *Some Statistical Studies of the Bangla Language.* Unpublished Doctoral Dissertation. Calcutta: Indian Statistical Institute.

Bhattacharya, Subhas (1992) *Bangla Ucchaaran Abhidhan (Bengali Pronunciation Dictionary)*. Calcutta: Sahitya Sansad.

Bidyasagar, Iswar Chandra (1986) *Barna Parichay (Bangla Primer).* Calcutta: Sishu Sahitya Samsad.

Chattopadhyay, Sanat Kumar (Ed.) (1986) *Prasanga Bangla Bhasa (Issues on Bangla Language)*. Calcutta: Paschim Banga Bangla Akademi.

Chattopadhyay, Suniti Kumar (1962) *Bangala Bhasatatter Bhumika (Introduction to Bangla Linguistics)*. Calcutta: Calcutta University Press.

Chattopadhyay, Suniti Kumar (1988) *Bhasa Prakas Bangala Byakaran (Bangla Grammar).* Calcutta: Rupa Publications.

Chaudhuri, Bidyut Baran and Umapada Pal (1995) Relational Studies between Phoneme and Grapheme statistics in modern Bangla Language. *Journal of Acoustic Society of India.* Vol. 23., No. 1., Pp. 67-77.

Chaudhuri, Bidyut Baran and Umapada Pal (1996) OCR Error Detection and Correction of an Inflectional Indian Language Script. Presented in the *13th International Conference of Pattern Recognition*, Vienna, Austria.

Chaudhuri, Bidyut Baran, Umapada Pal and Pulak Kumar Kundu (1966) Non-Word Error Detection and Correction of an Inflectional Indian Language'. Presented in the *National Symposium on Machine Aids for Translation and Communication*. JNU, New Delhi, India.

Chaudhury, Jamil (1990) *Banan o Uccharan (Letter and Pronunciation).* Dhaka: Bangla Academy Press.

Coulmas, Florian (1989) *The Writing Systems of the World*. Oxford: Basil Blackwell.

Dash, Niladri Sekhar (1997) Applicability of NLP in Bangla: A Linguistic Perspective. Presented in the *International CSP Workshop on Approaches to Knowledge Representation*, Jadavpur University, Calcutta, 18-20th February, 1997. (MS).

Diringer, David (1968) *The Alphabet: A key to the History of Mankind.* Vol. I&II. London: Hatchinson.

Ganguli, Subrata (1995) *Lipir Padanka Rekhay (On the Footsteps of Script).* Calcutta: Samatat Prakashani.

Haque, Enamul (1995) *Bangla Bakdhvani : Svarup o Binyas (Bengali Speech Sounds: Nature and Distribution).* Dhaka: Ayantika.

Haque, Mahbabul (1995) *Bangla Bananer Niyam (Rules of Bengali Spelling).* Dhaka: Jatiya Sahitya Prakasani.

Khan, I.; S.K. Gupta and S.H. Rizvi (1991) Statistics of Printed Hindi Text Graphemes : Preliminary Results. *Journal of IETE.* Vol. 37. No. 3. Pp. 268-275.

Majumdar, Paresh Chandra (1998) *Bangla Banan Bidhi (Bengali Spelling Rules).* Kolkata: Dey's Publishing.

Majumder, Nepal (1992) (Ed) *Banan Bitarka (Issues on Spelling).* Calcutta: Paschim Banga Bangla Akademi.

Majumder, Paresh Chandra (1995) *Adhunik Bharatiya Bhasa Prasange (In the Context of Modern Indian Languages)*. Calcutta: Dey's Publishing.

Ray, Punya Sloka (1997) *Bengali Language Handbook*. Calcutta: Paschim Banga Bangla Akademi.

Sampson, Geoffrey (1985) *Writing System: A Linguistic Introduction.* London: Hatchinson.

Sarkar, Pabitra (1984) *Bhasa Desh Kal (Language in Space and Time).* Calcutta: G.A.E. Publishers.

Sarkar, Pabitra (1992) *Bangla Banan Sanskar: Samasya o Sambhabana (Bangla Spelling Reform: Problem and Possibility).* Calcutta: Chirayata Prakashan.

Sarkar, Pabitra (1993) Bangla Bhasar Yuktabyanjan. *Bhasa.* Vol. 1. No. 1. Pp. 23-45.

Sen, Dinanath (1993) *Mudrancarca (Printing Practices).* Calcutta: Paschim Banga Bangla Akademi.

Sen, Sukumar (1993) *Bhasar Itibritta (History of Language).* Calcutta: Ananda Publishers.

Tagore, Rabindranath (1995) *Bangla Shabdatatta (Bangla Philology).* Calcutta: Viswabharati Prakashani.

Tripathi, J.N. (1971) A statistical analysis of Devnagari (Hindi) text graphemes. *Journal of IETE.* Vo. 17. No. 1. Pp. 25-27.