THE BANGLA-ASAMIYA SCRIPT AND ITS REPRESENTATION IN UNICODE

Probal Dasgupta
University of Hyderabad

Gautam Sengupta
Jadavpur University, Kolkata

0. Introductory Remarks

The script described here is one that the Eastern Indo-Aryan languages Bangla (or Bengali) and Asamiya (or Assamese), the established and traditional loci of this script, share with Santali (which is also written in other scripts), Manipuri and several other languages. As the linguistics and graphology of those languages come to be better understood, the behaviour of this shared script in those settings will play a crucial role, as all newly opened up domains do, in redrawing the overall picture of this script. Until that development takes place, we are confined to the study of the major languages that have traditionally used this and only this script. Furthermore, the background of the authors compels an emphasis on the way the script functions in the Bangla setting, treating the Asamiya material as a brief supplement to an overall Bangla-focused story. The present authors regret but cannot avoid this bias, which should be easy for Asamiya scholars to correct. For some current Asamiya-based remarks on this shared script, see Goswami and Tamuli (2003).

The present paper seeks to cover the synchronic realities at the levels of linguistic and graphological description, including the computational dimension that the study of scripts in our era is compelled to take on board for languages that have made the jump. Some rudimentary remarks about historical origins also appear, but will need to be supplemented by specialists of the history of script and language.

While eschewing phonological description for its own sake, we are compelled to provide a reasonably full picture of the segmental phonology of standard Bangla as the basic material that the graphology hooks on to. For convenience of reading, we use a basic phonemic transcription as our standard method of mentioning Bangla or Asamiya words, except where the specifics of a written form are under scrutiny. For the same reason, our discussion of the script is built around a romanization whose conventions are consistent with the phonemic transcription we employ.
We should ideally distinguish between the properties of the script proper, which can underlie several different orthographies for one or more languages, and those of an orthography that applies it to a particular language. But common parlance does not make this distinction. For immediate practical purposes, given the relatively marginal differences between written Bangla and written Asamiya at the structural level, we will therefore not make that distinction, and will speak of script somewhat loosely.

The distinctions we do need to make for our practical work are expressed in the transcription systems used. In our practice here, the graphological transcriptions are enclosed in « » to distinguish them from transcriptions at the graphic level enclosed in single angular brackets < >. We employ a i u for short vowel graphemes, ā ī ū for their long counterparts, ō for the functionally vocalic /ř/ originating historically from syllabic /r/, a’ a” for the diphthong graphemes, tilde for vowel nasalization (as in ā), n for the velar nasal, s for the palato-alveolar sibilant, n and s for the dentals, t d for retroflex plosives, r for the retroflex flap, ş for the retroflex sibilant, h for the palatal nasal, ṇ for the anusvāra and ḷ for the visarga to the extent that these may need separate representation, ȝ for the palatal grapheme with semivowel value (marked in the actual script by placing a dot under the independent allograph of the palatal grapheme), y for the palatal grapheme with non-semivowel value (this covers both the undotted independent allograph of the palatal grapheme and the postconsonantal dependent palatal marking known as ya-phalā), and backslash for the silencer that cancels the default vowel «a» on a consonant. The phonemic / // and phonetic [] transcription systems use æ ɔ for low vowels, c j for palato-alveolar affricates, and other symbols (including /a/ for the low back unrounded vowel) as in IPA and other standard transcriptions. It is possible that this stock needs to be eked out by drawing on IPA itself and other resources if one expands the scope of such an investigation. When Salomon (2003: 82), in a direct reference to Bangla, claims that ‘the written vowel e actually serves to represent three separate sounds’, he lists /ɛ/, in addition to our /ɛ/ and /æ/, which corresponds to nothing known in standard Bangla. The reference is perhaps to dialects of Bangla in which standard /æ/ does not appear, for we are not aware of any variant of Bangla in which lower-mid front and low front vowels appear to contrast with mid front, and Salomon does not clarify what he has in mind. We doubt that graphological studies are prepared to take on all the dialects at this stage of the enterprise.

The present exposition is a working draft. Section 1 provides a general overview of the script and some considerations involving
phonology and orthography. Section 2 takes up more concrete issues on the computational front requiring decisions.

1. Considerations of Script, Phonology, and Orthography

1.1. Diachrony and Description

We distinguish the content of descriptive linguistics, which needs to take on board whatever realities govern a particular state of a language, from the principles of linguistic synchrony, which interact with those of diachrony and other systems to yield these realities. The serious study of linguistic synchrony deploys scientific idealizations modelling the most transparent operations in, say, phonology or syntax. In contrast, the descriptive analysis of the phenomena of a language has to hug these phenomena closely and to suspend judgment about just how various factors interact to give rise to the appearances, if they are later scientifically judged to be as deceptive as all appearances.

The description of morphology, for instance, has to take into account both synchronic and diachronic principles. In the other direction, a study of diachrony must divide its attention between the tracking of phenomena through time, a historical matter, and the effects of diachronic phenomena as they show up within a particular state of a given language, where the archaic and the innovative sectors of the lexicon are layered differently.

In the study of language in recent decades, there has been some neglect both of diachronic issues and of the interface between spoken and written language. We face a certain problem of equipment and of preparedness of the discursive surroundings, therefore, when we approach the issue of the diachrony of the Bangla-Asamiya writing system.

For our purposes, let the following points suffice, inviting supplementation by specialists. To begin with a familiar first approximation (which we radically revise in section 2), the Bangla-Asamiya script is alpha-syllabic, or an akṣara script in the sense of Salomon (2003:71). Its remote ancestor was the Brāhmī script. Its classic shape was finalized over several decades when the sociology of print encountered the handwriting of Ishwarchandra Vidyasagar (1820-1891, the initiator of Bangla print literacy and the author of primers that have stood the test of time) and his associates in nineteenth century Bengal. There was one major revision of these shapes beginning in the nineteen fifties, with the rise of linotype printing and the need to reduce the number of distinct allographs. The current rise of computer technology has been occasioning much activity behind the scenes as far as the replication of older printing
standards is concerned, but little by way of script reform in the sense of the earlier impact of print technology on this script. At the level of orthography, Bangla-Asamiya writing is considerably more conservative than the phonologies of these languages. Many innovations in their phonology are not matched in the writing of words taken directly from Sanskrit and preserving the orthography of Sanskrit, standardly called tātsama words. As a result, there is a deep cleavage between speech and writing in these languages.

1.2. Generalities and Vowels

The basic graphological molecule of an akṣara script such as the Bangla-Asamiya script is an akṣara. It consists of a series of graphic atoms. The series must begin with zero or more consonants and end with either a marked vowel or the unmarked default vowel «a» (Bangla and Asamiya pronounce this as /ɔ/) or the silencer element transcribed by us as the backslash \ and called «hasanta» /hoʃɔnt/) in Bangla (the Nāgari term is «halanta>, Hindi /hɔlənt/). Thus, in «ud\bigna» /udbigno/ ‘anxious’, there are four akṣaras. The vocalic akṣara «u» consists of zero consonant plus graphological «u». The zero consonant triggers the graphic level choice of the independent vowel allograph of «u». The next akṣara «d» consists of graphic «d» plus the silencer, without which graphic «d» would be understood to carry a default «a» (modern Bangla strongly tends to omit all silencers at the graphic level, leaving it an open question whether they are even graphologically absent). A consonant «b» initiates the following akṣara in the word under consideration and triggers, for graphological «i», the graph level choice of the dependent allograph «i», which appears to the left of the graph «b» and is called the «hrsba-i-kāra» /hroʃɔikar/, the ‘short «i» supplement’. All vowels distinguish the dependent (or supplementary) from the independent allograph. The last akṣara of «ud\bigna» begins with the conjunct consonant «gn» and leaves the default vowel «a» understood.

Some conjunct consonants like «gn» combine the two graphs, in this case vertically (while «mph» combines m with ph horizontally). One consonant, «t>, in a conjunct-initial position, chooses a prefixal allograph called «khandYa ta» /kʰɔntɔtɔ/ ‘broken t’, whose use preempts that of the unattested «t» combination throughout the language. In the other direction, several consonants in conjunct-final position choose a dependent allograph known as a «phalā» /phɔla/. One such consonant, «r>, interacts with the dependent allograph choice of vowels. Thus, in «dru>, the «raphalā» /rɔphɔla/ (this word’s reduction of /ɔ/ to /ʊ/ reflects factors discussed in section 1.3) on «d» triggers quirky shapes for dependent «u, ū» attached to the right of the
akṣara, distinct from the usual dependent «u, ū» allographs that occupy the position below the akṣara body. Such quirky shapes persist in calligraphy to this day and were standard in pre-linotype print Bangla but have been standardized away in linotype Bangla during the shift mentioned in section 1.0. There also exist a few unique conjunct shapes «ṅ, ṇ, ṭh» etc.

We now take up the atomic segments, system by system, in the context of phonological and graphological exposition.

The Bangla vowel system comprises the front unrounded vowels /e, æ/, respectively high, mid and low, the low back unrounded /a/, and the respectively low back rounded, mid back rounded and high back rounded /o, u/. The script treats /a/ as a default vowel, left unmarked when it serves as nucleus for a consonantal onset. Thus, a graphic <kh> is to be pronounced as /kha/ unless ‘silenced’ by the silencer \, the way other scripts derived from Brāhmī treat the schwa. One standardly represents graphic <kh> as graphemic «kha», even in the word-final position where, as in other modern Indic languages, the default vowel is typically silent. The on the whole conservative script continues to mark diachronically lost contrasts between short and long high vowels. That the orthography of particular words uses long vowel symbols has to be rote-learnt, as in «dina» ‘day’ and «dīna» ‘poor’, both phonologically /din/. Other features of the way the script treats vowels become clear in the discussion of vowel harmony in the Synthesis subsection 1.3.


The consistent graphological marker of nasalization that features in the written forms of the examples above, the candrabindu, is distinguished from the anusvara and always maps on to phonological nasalization. In addition, certain words in the tatsu lexical stratum with what in Sanskrit was a consonant plus non-dental nasal sequence such as /jñ, sm, tm/ have nasalization (plus, if metrically appropriate, compensatory consonant gemination) in Bangla, whose conservative graphism therefore has <<jñāsā, smarāna, jībāśma, barta, bismita>> for /jiggāʃa, ʃrōn, jibāʃō, borto, bijʃto/ ‘query, inquiry; memory; fossil; road; surprised’. For more, see the algorithms in section 1.8.

Lacking a monographic symbol of its own, the vowel /æ/ is written variously as «e», «ya», or «ya»: «becă, byăkaraṇa, byastă» for
Vowel length is phonologically non-significant. Phonetic length is a function of position. Typical contexts for phonetically longer vowels include monosyllables like /gal/ ‘cheek’ (in contrast with the phonetically short /a/ in /galo/ ‘strain-2PN’), where length is mandatory, and final syllables like /gelo/ ‘went’ whose final /o/ can be lengthened for prosodic or emotional reasons. Stress and tone do not distinguish words in Bangla. Phonetic “stress” is a component of the prominence of word-initial syllables. The distinction between low /a/ and mid /o/ is associated with patterns of such prominence. For details, see the Synthesis section 1.3.

Nasalization as a phonetic spread feature affects pre- and post-nasal vowels. Therefore no phonologically nasalized vowel occurs right before or after a nasal consonant. Sister languages are different. Hindi-Urdu, which replicates the Bangla pattern only pre-nasally, distinguishes post-nasal /e/ from /ë/ in verbs like /pəhne/ ‘wear-Opt-3PSing’ vs /pəhñe/ ‘wear-Opt-3PPPlur’.

Unlike many nasalized vowel languages, Bangla distinguishes V*P from Vn*P, where P is an occlusive (a superordinate term subsuming plosives and affricates), and n* is a homorganic nasal, for both voiced and voiceless P, as in /pāja/ ‘pile (of bricks)’ vs /panja/ ‘pānjā’ ‘arm wrestling’, or /jātar/ ‘sātāra’ ‘swimming’ vs /jantar/ ‘sāntāra’ ‘Shanta’s’. A nasalized vowel in Bangla does not force a homorganic nasal transition into a voiced plosive in the manner of Hindi-Urdu /cən:di/ ‘silver’, which thus contrasts with Bangla /cādi/ ‘silver’. The only type of stop that a nasalized vowel in Bangla never precedes is /g, gh/, a gap reflecting the velar nasal’s special status. But even there, arguably, pairs like /føngə/ ‘with’ vs /føne/ ‘clown-Loc’, or /føŋghatik/ ‘terrible’ vs /føŋhæi/ ‘Shanghai’, really instantiate Vn*P vs V*P, if the phonotactics makes Vn(h) count as standing in for the phonologically equivalent (but never phonetically realized) Vg(h).

1.3. Semivowels

We may usefully describe as y-type and w-type semivowels the intervocalics in sequences like /dʌf(ı) (y) at/ ‘two-hundred eight’ (inter-word intrusives are consistently of the y type) and words like /di(ı)y/ ‘give-ImperFut-2PN’, /sthani(y)/o/ ‘local’, /khe(y)/ ‘eat-ImperFut-2PN’, /prode(y)/o/ ‘payable’, /bæ(y)e/ ‘expense-Loc’, /bəyam/ ‘exercise’, /bəyu/ ‘wind’, /bɔyɔn/ ‘weaving’, /moyur/ ‘peacock’, /juyo/


Vowel length is phonologically non-significant. Phonetic length is a function of position. Typical contexts for phonetically longer vowels include monosyllables like /gal/ ‘cheek’ (in contrast with the phonetically short /a/ in /galo/ ‘strain-Pres-2PN’) where length is mandatory and final syllables like /gelo/ ‘went’ whose final /o/ can be lengthened for prosodic or emotional reasons. Stress and tone do not distinguish words in Bangla. Phonetic “stress” is a component of the prominence of word-initial syllables. The distinction between low /a/ and mid /o/ is associated with patterns of such prominence. For details, see the Synthesis section 1.3.
As illustrated, y-glides contiguous to /i, e/ and w-glides adjacent to /u, o/ are consistently optional. The w-glide is variably (across speakers and across tokens) present in /de(w)a/ ‘wall’ and, for some speakers, even /de(w)a/ ‘to give’ and /ne(w)a/ ‘to take’; when the words omit this glide, they often substitute /y/.

The orthography omits inter-word intrusives: «dušo āta» ‘two-hundred eight’. Written «y» is optional only between stem-final «i, e» and desinential «o». Otherwise -- partly due to a convention of inserting a graphic <j> between any vowel and a following vowel other than «u, o» -- nearly every potential /y/ is written as a mandatory «y», as in «sthānīyā», «pradeyā», «byaye» among the forms listed above. In this context it is important to note the exceptionally optional written «y» in verbs like «di(y)o» and «khe(y)o», Future Imperative forms mentioned above. All verbs of these canonical shapes fall under that exception.

The orthography implements /w/ systematically as «oy» (as in «deoya» for /Idewal/ ‘to give’) between non-round vowels, treats the «0» in that «oj» as optional if preceded by «0» (thus «So(o)yā» for /So(w)a/ ‘to lie down’), and omits it in contiguity to «u», whence «hāluyā, bāula» for /halu(w)a, ba(w)ul/ ‘semolina dish, Baul’.

Any account of intervocalic glides must note the unavailability of semivowel onsets. Items like /yar/ ‘mate’, /yarki/ ‘jest’, /warif/ ‘heir’, /wat/ ‘pillow-case’ can be pronounced with initial semivowels. But most speakers and all writers substitute /i(y)a, o(w)a/, «iya, oya» in «iyara; iyarki, iyaraki; oyarisa; oyara». We conclude that intervocalic semivowels count as interludes. The language has no true onset semivowels.


— 220 —
Crucial contrasts that the orthography conceals distinguish diphthongs as in /deule/ «deule» ‘bankrupt’ or /paona/ «paoanaly» ‘debt due’ from bisyllabic sequences with an optional semivowel buffer as in /de(w)ule/ «deule» ‘temple-Loc’ and /fa(w)oner/ «fanoanaly» ‘Shaon’s’. Spellings betoken the reality of the contrast but mark its incidence only variably. Orthography records an obligatory «y» in /da(y)i/ «dayi» ‘responsible’ and no «y» in /da/ «day» ‘nurse’. But one writes /ga(y)e/ ‘body-Loc’ always as «gae» and /gae/ ‘sings’ only as «gaya». No devices exist to mark a contrast like /deule/ vs /de(w)ule/.

1.4. Synthesis

This subsection is placed where it is because word integration phenomena have to do with how vowels (and semivowels) are handled.

Bangla is a vowel harmony language. Compare /pɔcaʃi/ «pɔcaši» ‘eighty-five’ with /pɔcisi/ «pɔcisi» ‘twenty-five’, /bæca/ «beca» ‘to sell’ with /beci/ «beci» ‘I sell’, and /kena/ «kena» ‘to buy’ with /kini/ «kini» ‘(I) buy’ to see that a lone vowel ‘eligible’ nucleus – of low or mid default height and having (unlike /a/) the same value for rounding and backness (back rounded or front unrounded) – is ‘raised’ by a contiguous high trigger, compelled to climb up one notch to nondefault mid and high respectively, and that, for /o/ «a» climbing to /o/ «a» and for /æ/ «e» climbing to /e/ «e», the orthography conceals this morphological mutation affecting the phonology (which we will call a phonological mutation for short, without prejudice to various positions on the issue of which module of core linguistics should handle it). A two-vowel ‘eligible’ nucleus obeys the same pressures, compare /pɔtəlisi/ «pɔtəlisi» ‘forty-five’ with /poitiʃi/ «poitiʃi» ‘three-five’; /baper bo(y)eʃi/ «baper bayasa, bayesa» ‘father’s age’ with /baper boiʃi/ «baper bayasi» ‘(someone who is your) father’s age’; and /ɔpɔbbæeʃ/ «apabyaʃa» ‘wasteful expenditure’ with /ɔpɔbbe(y)i/ «apabyayi» ‘(one) given to wasteful expenditure’. Low vowels represented in writing by the default vowel do not indicate the vowel harmonic raising: the first syllables of /pɔciʃi/ «pɔciši» ‘twenty-five’, /poitiʃi/ «poitiši» ‘three-five’, /boiʃi/ «bayasi» ‘(who is...) age’, /be(y)i/ «byayi» ‘(one) given to (…) expenditure’ are written as if they contain unmutated low vowel nuclei – one source of the speech-writing gap in Bangla.

As a fuller study of Bangla vowel harmony (not the issue here) would show, in this normal or dominant vowel harmony pattern, a
triggering high vowel on the right raises a left-hand eligible. This norm is absolute and strong in the conjugation, while, among nonverbs, the norm is absolute only if a low must be raised to mid, as in the reciprocity reduplicates /bolabolil/ ‘speaking with each other’, /ghâjaghêjì/ ‘ghësâghësi’ ‘rubbing shoulders with each other’, where the second copy must have a mid nucleus.

The Bangla vowel harmony system also exhibits a graphologically pertinent counter-normal pattern, where a preceding trigger affects a right-hand eligible /a/, with verb and non-verb subpatterns. The latter turns the target /a/, which we see intact in the adjectives /phaṭa/ ‘burst’, /bhaṉa/ ‘broken’, /paka/ ‘ripe’, /poca/ ‘rotten’, into a mid vowel copying the backness of the trigger: /bhije/ «bhije» ‘wet’, /bhulo/ «bhulo» ‘forgetful’, /dëbo(jahaj pahař)/ «dëbo(jahåja, pähåra)’ ‘underwater (ship, mountain)’. The subpattern for verbs turns /a/ uniformly, if unexpectedly, into a phonological /o/ which, unlike the /o/ «o» of «bhulo» and «dëbo», is rendered in the orthography by some authors with a consistent «0» but by many other authors with «o» or «a» depending on the lexeme. The stem formative that appears as /a/ «ā» in triggerless /pala/ «pâlā» ‘run away’ or /cëca/ «cëcā» ‘shout’ comes out as /o/ after a raising trigger: /phuro/ «phura, phuro» ‘be finished, get used up’, /cibo/ «cibo, never ciba» ‘chew’, /douro/ «da"ra», in some tenses da"ra» ‘run’ (cf «da"rala» ‘ran’ but rarely «da"ra» ‘will run’ and never «da"ra» for the imperative).

The verb subpattern equally unexpectedly treats mid /e/ as a mutation trigger: /bero/ ‘go out’, /ego/ ‘go ahead’, /pero/ ‘cross’, where the graphological patterns are similar. Phonologically speaking, the verb subpattern is unexpectedly also active outside the conjugation, in reciprocal reduplicates like /jhuoljholi/ ‘insistence’, /mukhomukhi/ ‘face to face’, /culoculi/ ‘hair-pulling combat’, /piṭhopiṭhi/ ‘back to back, one after the other’, a template characterized, as cases given earlier show, by an inter-copy default /a/. But the script always implements these non-verb instances of /o/ with a consistent «o».

The counter-normal pattern is, like the normal pattern, omnipotent in the conjugation and a lexical choice ridden set of tendencies outside it. Both the patterns of vowel harmony refrain from touching English loans: /ofis/, /bætjun/ and the like violate the norms with impunity.

In addition to vowel harmony, Bangla exhibits prosodic patterns associating low vowels with positional prominence. Thus, only a root-initial syllable can host æ /: /æk/ «eka» ‘one’ contrasts with its formal negation /onek/ «aneka» ‘many’, where the low vowel is raised to mid /e/ and the graphology leaves the mutation unregistered. As for /o/, it
is allowed only in root-initial syllables and at strong niches in the rhythm, giving way to its weak congener /o/ elsewhere, whence the alternation between the initial syllable’s /a/ in /porajito/ ‘defeated’ and /o/ in its morphological negation /oporajito/ ‘undefeated’. As the spellings <parājita, aparājita> show, the graphology ignores this.

Another indicator of the prominence of the initial syllable is its tendency to attract nasalization, which, in general, is allowed in non-word-initial syllables only rarely, and in non-root-initial syllables never. (French loans are exempt.) This has a tangible effect on the exceptional words /âtta; âtto; âttiyo/ ‘soul; self; relative’. Because of their Sanskrit-based orthographies with a consonant plus non-dental nasal sequence «âtma; âtma; âtmīya», one would have expected to pronounce these words as /âtâ; âttō; âttiyo/ (and indeed some stage accents which follow the orthography pedantically do use these spelling pronunciations) in accordance with the treatment of such Sanskrit sequences explained above, at the start of our Vowels section. That the nasalization, breaking this expectation, jumps to the initial syllable, must reflect a powerful tendency. Its force affects even the French loan /restorâ/, always written «restor~» but often pronounced as /restora/, moving the nasalization leftwards, if not far enough for total naturalization.

1.5. Consonant Segments: Sonorants:

The basically alveolar /n/ and /l/ have palatalized, retroflex, and dental allophones that precede corresponding consonants. The phenomenon requires graphological comment only because /ṅ ṇ ṇh ng/ and /nc nch nj njh/ are written with conjunct letters that the script pedagogy portrays as «ṅT ṇTh ṇD ṇDh» with a retroflex «ṅ» and «ṅc ṇch nj njh» with a palatal «ṅ».

The bilabial /m/, as in /mom/ «moma» ‘wax’, is of graphological interest to the extent that «Cm», for various values of C(onsonant), corresponds to various phonological forms. For instance, «tMV» and «dmV» surface as /ttV/ and /ddV/, respectively, as in «bartma» /bOrto/ (degemination after /t/ is automatic) ‘road’ and «padma» /poddo/ ‘lotus’. For a full listing, see the algorithms in section 1.8.


The dependent nasal grapheme in the alphabet (traditionally called
the «anusvara» /onuʃʃor/ in Bengal, rather than the «anuvāra», Hindi /anusvār/) can be transcribed in some special fashion such as ṅ for technical purposes. But Bangla orthography treats «ṅ» as if it were simply «n\», a silenced velar nasal (rather than default-vowel-endowed <ṅ> «ṅa»), which means that forms with a phonologically unambivalent velar nasal like /ʃonʃar; ʃonʃoe; ʃonʃbad; ʃonʃoron; ʃonʃhar; ʃonʃhita/ ‘domesticity; doubt; news; restraint; destruction; code-book’ in Bangla, written «saṃsāra; saṃsāya; saṃbāda; saṃbaraṇa; saṃhāra; saṃhitā» if we resort to this «ṅ» device, correspond to a wide range of contemporary Indic renderings of the older, slightly mysterious dependent nasal. We will, unless «ṅ» becomes necessary for the sake of focusing on this issue, use simply «ṅ\», as Bangla never uses graphic <ṅ\> in its orthographic practices.

The liquid /ɻ/, an advanced alveolar approximant in normal standard speech with a postdental tap allophone (but consistently a tap or trill in many nonstandard varieties and in much stage enunciation), remains distinct from the (always non-initial, and otherwise limited in distribution) retroflex flap /ɶ/ in the western standard. But in all eastern varieties, /ɶ/ becomes /ɻ/ unconditionally. Even in the west, /ɶ/ tends to approach /ɻ/ in words like /əcrano/ ‘to comb’, /hatəno/ ‘to grope’. Speakers of Bangla have no doubt that /ɶ/ contrasts with the plosive /dʃ/, and the availability of direct evidence bears them out, unlike many other South Asian languages. Though somewhat far-fetched, the minimal pair /kaʃəd/ ‘Canada’ vs /kaʃəɾa/ ‘Kannada’ is surely decisive. If common nouns are preferred, there is /reʃjo/ ‘even castor’ vs /reʃʃo/ ‘radio’. All this is graphologically straightforward; «r», «ɾ» and «ɖ» are distinct from each other. What needs comment is the phonological sequence /ɾi/, which sometimes corresponds to the graphological CV sequences «ri» and «ɾi» and sometimes to the graphological semiconsonant «r» as in «krti» /kritis/ ‘successful’. See section 1.8 for some details concerning this.

The ‘aspiration’ element /h/ has occasioned much discussion. We may take the convenient stand that /h/ «h» is murmured intervocally and voiceless elsewhere, leaving these characterizations for careful clarification at the phonetics-phonology interface. The visarga «h» in the alphabet serves as a final /h/ in exclamations like /bah/ «bāh» ‘bravo’, /uh/ «uh» ‘ouch’. But «h\» in «saḥ» ‘Shah’ or in «aḷḷah\» ‘Allah’ (with the usual standard option of omitting the silencer in practice) is silent, /ja, alla/.

The script appears to postulate greater unity for the aspirated flap phoneme /ɾh/ in /gurho/ ‘profound’, by using a single symbol for it, than for the /rh/ sequence which is shown as two graphemes in /gorhit/
'despicable'. We return to the matter after finishing with our consonants.

1.6. Consonant Segments: Plosives and Affricates:

Every stop series displays a voiceless and a voiced pair of simple vs 'aspirated' stops. Neither the traditional term 'aspirated stop' implying one segment nor conventional transcriptions like /kh gh/ implying two segments need be construed as encoding any specific analysis.

The velar series /k kh g gh/ is exemplified in /ka/ 'caw', /kha/ 'please eat', /ga/ 'body', /gha/ 'wound'. The series of alveolo-palatal affricates /c ch j jh/, as in /ca/ 'tea', /cha/ 'offspring', /ja/ 'husband's brother's wife', /jha/ 'a surname', shows considerable variation. Some of this variation is phonological: the series tends to become /ts dz s z/ before dentals and liquids, raising the issue of whether, say, the word for 'to choose' written as «bachate» and thought to be phonemically /bacht/, but standardly rendered as [baste], perhaps actually has the same phonology as «āste» /aste/ 'slowly'. Some of the relevant variation is subphonemic. It is a familiar shibboleth that many eastern dialects use alveolars as the principal allophones and have been making an impact on the western standard. Sociolinguistic work on this impact is long overdue.

The series of retroflex plosives /ṭ ṭh ṭʃ ṭθ/ is more apico-alveolar in Bangla than the true apico-palataIs of sister languages to the west. (This could be why the plosive /ḍ/ in Bangla is not phonologically close to the flap /ḍ/ the way it is in many other South Asian languages where the retroflex plosives are truly apico-palatal.) To the east of Bangla lies Asamiya, where these plosives and the dentals merge in an alveolar series. Examples of the retroflex series in Bangla: /ṭok/ 'sour', /ṭok/ 'a cheat', /ḍak/ 'mail', /ḍhak/ 'drum'. Skipping the alveolar position, unutilized in Bangla, we have a dental series /ṭh ṭḍ ṭdh/ as in /ṭan/ 'tune', /ṭhan/ 'white cloth', /ṭdan/ 'donation', /ṭdhan/ 'paddy'. In the labial series /p ph b bh/ as in /pal/ 'sail', /phal/ 'ploughshare', /bal/ 'child (learned form)', /bal/ 'forehead (learned form)'. The unstable aspirated /ph/ is often realized, under sociolinguistic and phonotactic conditions not yet examined, as [f]. Less often, /bh/ likewise gives way to a [v] which is in any case required in /aovan/ 'call', a point taken up presently.

1.7. Consonant Segments: Fricatives and Aspiration

The conjunct letter «hb» (encoding an older /hw/) in a few tatsama words like «gahbara» 'abyss', «āhbāna» 'call', «bihbala» 'upset', «jihbā» 'tongue' is rendered by careful speakers as a labio-dental /v/ (otherwise available only in foreign loans) preceded by a w-type coda:
IgJovJr, aovan, biuVJl, jiuv/. (Some users of /v/ strengthen it to [vh], a matter of subphonemic taste; many users don’t.) The less careful /goobhɔr, aobhan, biubhɔl, jiubhə/ are tolerated. Such variants as /abbhan, aohan/ have fallen into disuse. But /bh/ users appear to differ from /v/ users at the shibboleth level. Hence our decision to record a cross-speakers /v/ vs /bh/ contrast here rather than treat the matter as subphonemic. The language certainly has no general [v-bh] allophony under /bhl/ on a par with [f-ph] under /ph/. Based on these facts, some authors have introduced «hb» to represent foreign /v/ quite generally, in a «hbalphagār» ‘Wolfgang’ or a «hbebara» ‘Weber’. Others still use «bh» for foreign /v/, as speakers of Bangla unable to handle [f, v] routinely substitute [ph, bh] in foreign vocables.

The /z/ of foreign loans has the similar routine substitute /j/ in the case of Perso-Arabic origin. But the Islamic context, and the coincidentally correlated dominance of [z] over [j] in the eastern speech region, once led to some use in Eastern Bengal of the semivowel-derived special /j/ grapheme – our system’s «y» – to represent /z/ in words like «hayarata» /hɔzɾat/ ‘Hazarat’. This practice is now obsolete to the point of unrecognizability in Western Bengal; current proposals for [z] include «j» plus underdot or afterdot; no firm decisions have emerged. With the spread of English, some Bangla speakers now tend to preserve /z/ and even /ʒ/ in loan words. Even for them, the [z] of a [mazlo] ‘brushed (teeth)’ still counts, along with [madzlo], as a subphonemic allegro variant of lento phonology’s /mælo/, quite remote from the /z/ of a /mazl/ ‘muzzle’ which is [z] even in lento speech.

Velar fricatives have no significant past in Bangla, even in loans, and no foreseeable future. That they occur in allegro speech, intervocically, has had no impact on the system.

Bangla has two sibilants, alveolar /s/ and palato-alveolar /ʃ/, contrasting in pairs like /aste/ ‘slowly’ vs /afte/ ‘to come’ in standard speech, a contrast strengthened by the growth of English literacy. In many Southern dialects, including working-class speech in the towns whose elite speech defines the standard, they merge into /s/. Contiguity with this fact puts some sociolinguistic pressure on the contrast. The contrast is phonotactically weak even in Standard Bangla, where the sibilants can contrast only between a vowel and a /t, n, r, l/. The rough generalization is that syllable-initially, only /sk, st, sr, sn, sl, sm/ clusters are possible; elsewhere, basically only /ʃ/ is possible, with some lexical exceptions: /bas/ ‘bus’ contrasts with /baf/ ‘residence’.

We turn now from fricatives to aspiration. The phonological unity and phonetic diversity of aspiration across the voiced-voiceless
boundary in the system are an area inviting research. Even if the phonetics of normal /gha/ ‘wound’ exhibits murmur and is quite distinct from the true aspiration of /kha/ ‘eat’, the two seem to fall together in whispered speech. The phonology seems to treat what have traditionally been called voiced and voiceless aspirated stops alike: both sets effectively lose their aspiration word-finally, and if no sonorant follows also syllable-finally.

Is an aspirated plosive (or affricate) best described as one segment, as the term Aspirated Stop suggests, or as a sequence of two, as the transcription suggests? Only time and phonological research will tell. One systemic fact that speaks against the one-segment view is the total non-existence in Bangla of any pair of morphemes /AX'B/ vs /AXhB/ where /X/ is a plosive or affricate, /X'/ is a true single segment aspirated congener of /X/, and the sequence /Xh/ is a clearly bisegmental sequence placing that /X/ in contiguity with an independent /h/ after it. If ‘aspirated stops’ are single segments, this gap in the system remains unexplained. Holders of the single segment view who provide an explanation of this gap are likely, in their explanation, to approach the essence of the segment sequence view. As in many other descriptive controversies, not much divides the two views.

1.8. A Script Chart

1.8.1. Vowels

<table>
<thead>
<tr>
<th>Independent forms</th>
<th>Transliteration</th>
</tr>
</thead>
<tbody>
<tr>
<td>আ ই ঈ উ ঊ এ ঐ ও ঔ a i i u u r e a' o a&quot;</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Combining forms</th>
</tr>
</thead>
<tbody>
<tr>
<td>ক কা কি ক্ষ ক্ত ক্ণ কে কেঁ কৌ কৌঁ</td>
</tr>
</tbody>
</table>

1.8.2. Consonants

<table>
<thead>
<tr>
<th>Characters</th>
<th>Transliteration</th>
</tr>
</thead>
<tbody>
<tr>
<td>ক খ গ ঘ ঙ</td>
<td>k kh g gh ń</td>
</tr>
<tr>
<td>ছ জ ঝ ঞ</td>
<td>c ch j jh ń</td>
</tr>
<tr>
<td>ট ঠ ড ঢ ত থ দ ধ র ঠ ড় ত়</td>
<td>t th d ķ dh rh ń</td>
</tr>
<tr>
<td>প ফ ব ভ ম</td>
<td>p ph b bh m</td>
</tr>
<tr>
<td>য র ল রি</td>
<td>y ķ r ṛ l w</td>
</tr>
<tr>
<td>শ ষ ষ্ট ষ্ট</td>
<td>s š s h</td>
</tr>
<tr>
<td>ছ</td>
<td>ṇ h</td>
</tr>
</tbody>
</table>

1.9. Algorithms from Graphology to Phonology

The script has no explicit space for the vowel phoneme /æ/. But tatsama words with an orthographic «Cyā» are pronounced /Cæ/; so the use of this graphism for /æ/ -- and of a special conjunct vowel,
with dependent «-yā» attached to the normal «a» -- has become standard. This amendment is not taught as part of the normative alphabetical listing.

The silencer is used sparingly. Speakers are expected to know that «na» comes out as /n/ in «bhinadesi» /bhindeʃi/ ‘foreigner’ but as /na/ in «anara» /onoʃ/ ‘unmoving’ and /no/ in «anabarata» /nobaroto/ ‘continuously’. No known rules reliably map the written «a» to these and other phonemic correlates and back in Bangla orthography. This single biggest problem in the system has been compounded by the transition from the High to the Low norm in Bangla writing when diglossia was renegotiated in the twentieth century. The High system was conceptualized on a Sanskrit foundation and used graphemic «o» sparingly, making it obligatory to write /dito/ ‘used to give’ as «dita», for instance. But the Low system suggests a closeness to speech that has led many authors to switch to «dito» in such cases. This switchover has not been complete or consistent in anyone’s usage, and remains an area of contestation. Another contested area is a spelling reform measure eliminating the speech-unfaithful long «ï» in favour of short «i» wherever this can be done without directly confronting the tatsama doctrine. Some authors have been resisting this reform in the case of feminines like «māmâ» ‘mother’s brother’s wife’ (derived from «māmā» ‘mother’s brother’), for which the long «ï» is an established feminine marker. The fact that feminines ending in «» are frequent in the tatsama sublexicon, a fact that underlies this resistance, is left untouched by spelling reforms that do not tackle the tatsama doctrine itself. In such contested areas, the dust is unlikely to settle in the foreseeable future.

Apart from these areas of systemic uncertainty, there are a few unpredictable and thus truly exceptional speech-writing correlations as in cases like «padma» /poddo/ ‘lotus’ (the non-nasalized final vowel is unpredictable), «gardabha» /gordhob/ ‘donkey’ (one would have expected /gordob/ with a possibly weakened final aspiration, instead of this throw-back of the aspiration to the next plosive to the left), «smita» /smito/ ‘smiling’ and «smārta» /smarto/ ‘Smriti-erudite’ (the only words with an initial /sm/ left intact in speech; the regular pattern is shown in «smāraka» /ʃārok/ ‘memorial’). These are not a problem. The areas of systemic uncertainty are large enough to give rise to a persistent malaise.

We conclude with some formulas to bridge the orthography-speech gap caused mainly, but not exclusively, by the vagaries of tatsama words. In the statements we provide, positions (initial, medial, final) are defined within the orthographic word; for this purpose, a word remains a word even if a prefix is attached to it.
«Cb», for any C[onsonant] except «g, m, r, h», is single /C/ if initial and geminated /CC/ if medial. For «g, m, r», «Cb» is simply /Cb/. For «hb» /v/, see the fricatives subsection, 1.6. Some cases of «db» are to be read as /db/ on the basis of idiosyncratic lexical knowledge, but in such cases most writers today use (or casually omit) a silencer «dť» and not the «db» conjunct consonant.

«Cya, Cyă» for any C other than «b, h» are respectively /Co, Cæ/ initially (the /o/ but not the /æ/ is liable to be raised by a vowel harmony trigger) and /CCo, CCæ/ medially. For the purpose of this formula – and other purposes, for that matter – «jnV» counts as if it was «gyV». The formula holds for a medial sequence with C = «b»; but initial «bya, byă» are both /bæ/, with the further embedded exception that «bya» /bæ/, but not «byă» /bæ/, raises to /be/ if a trigger follows, as in /bekti/ realizing «byakti». The principle that prefixation leaves a word-initial boundary intact is mitigated only to the extent that gemination applies, as in «bikhyāta» /bikhaeto/ ‘renowned’ which differs only in this respect (not in its vocalism) from the unprefixed «khyāta» /khaeto/ ‘famous’. Compare the vocalism with «ākhyā» /akha/ (not /akkha/) ‘epithet’. For C = «h», the formula works if C is initial. If it is medial, «hy» is always realized as /jjh/: «sahya» /jōjhol/ ‘tolerance’.

Final «a» observes what one might call Indic Silence if preceded by a non-cluster, non-«h» consonant, unless it forms part of a tatsama past participial «(i)ta» affix (the general deletion principle overrides this exception clause if the word is a proper noun).

In an even-numbered or final syllable (but only if the preceding syllable nucleus is not /i/), «a» reduces to /o/, and a «Cra» sequence anywhere is rendered as /Cro/, both of these reductions subject to the mitigator «ya» /e/ which, if it follows an «a» liable to reduce to /o/, overrides the reduction and makes it come out intact, as /a/: hence unmitigated «apada» /apod/ ‘trouble’ (but «bipada» /bipod/ ‘danger’) and «āśrama» /asrom/ ‘ashram, hermitage’, but mitigated «āśya» /ajǝ/ ‘receptacle’, «āśraya» /asro/ ‘asylum’.

An «a» followed (not across word boundary, and this includes a boundary between a prefix and a recognizable word) by a syllable with a high vowel or by a «Cy» cluster or by «kṣ» raises to /o/: «ati» /oti/ ‘very’, «gadya» /goddo/ ‘prose’.

A «CmV» sequence where V is any vowel and C is a coronal (dental, retroflex, palatal) plosive or affricate comes out as initial /CV/ and medial /CCV/, provided that a preceding vowel enables the gemination to surface. One case where the absence of a vowel prevents it is «bartma» /bortô/ ‘road’. These and other vowel
nasalization markings are subject to the steady erosion of nasalized vowels throughout western speech under the influence of eastern varieties that lost it long ago.

In careful speech, «VCr» with graphologically vocalic «r» places a syllable boundary between V and C, whereas «VCri» or «VCrī» with the graphological consonant «r» places it after C (some speakers exaggerate this by gerninating a C in every «CL» cluster, L any liquid). It remains to be seen if this type of care can survive such demographic realignments as the community of literates is now experiencing.

A conjunct consisting of a nonretroflex sibilant plus «t, d, n, r, l, r» implements the sibilant as /s/, though spelling pronunciation for some speakers preserves the palatal sibilant if conjoined with «n, l» (but the switch to /s/ is absolute before «r» and «r»). A conjunct of «s» with «p, ph, k, kh» tends to realize the sibilant initially as /s/, with /ʃ/ a marginal variant, and medially as a consistent /ʃ/. A conjunct of «s» with a retroflex always implements it as /s/. The retroflex sibilant is always rendered as /ʃ/, except in the cluster «kʃ», which maps into /kh/ initially and /kʃ/ medially. That cluster plus «m» have varying fates, but always lose the /m/: «lakṣmī» /lokkhī/ ‘Lakshmi’ shows no nasalization, «pakṣma» /pokkhō/ ‘eyelash’ does, and with «lakṣmaṇa» [lokkhōn] ‘Lakshmana’ it is unclear, due to the interference of the phonetic nasalization imposed anyway by the following [n], if phonetic [d] here counts as phonological /d/.

1.10. Observations on Asamiya

We begin with some obvious remarks. Graphological sibilants are rendered in Asamiya phonology as the velar fricative /x/. Default postconsonantal «a» after a single consonant at the end of a non-tatsama word always corresponds to phonological zero, instead of, as in Bangla, sometimes corresponding to /o/. We have already made the point that both dental and retroflex graphemes correspond to a phonologically merged series that is neither retroflex nor dental but alveolar, and the point that there is no separate /t/ phoneme (and that therefore «t» is rendered as /s/).

We are indebted to Pallavi Borah for providing us with the following data, on the basis of which we can make some preliminary observations going beyond these generalities about where the speech-writing interface in Asamiya differs from that in Bangla.

If «asama» /əxɔm/ ‘Assam’ and «āma» /əm/ ‘mango’ show that the behaviour of «a» is basically as in Bangla, reasons for caution become apparent in «ajñāna» /əg yan/ [əggyan], where we begin to propose interpretations, such as phonological single /g/ for this
particular kind of phonetic gemination, still wondering about the proper treatment of phonological /Cy/ sequences. Further caution is appropriate when we reach this pronunciation of a Bhupen Hazarika song line: «dakṣyasthala hala tezapura» /lokhyosthā hol tezpur/ [loikhōsthōl hol tezpūr] ‘the target is Tezpur’, which shows us that the negative prefix «a» /a/ behaves differently from the initial akṣara «a» preceding a «CCy» sequence, where C is any consonant. In general, such a sequence comes out phonologically as /CCy/ but phonetically as [iCC], and induces vowel harmony effects, so that we find [loikhō] and not the expected /loikhō/. That «Cy» induces an automatic epenthetic preconsonantal [i] is confirmed in the example «kāmākhyā» /kamaikkha/ ‘Kamakhya’. What requires further investigation is why «kābya» /kaibba/ ‘poetry’ and «bhāgya» /bhagga/ ‘fate’ pattern this way, while «ājñā» /agya/ [aggya] does not. Tentatively, we hypothesize that /Cy/ with a single /C/ does not come out as [iC] whereas /CCy/ does surface as [iCC], and at the speech-writing interface that «jñā» is rendered as /gy/. Pursuing «kṣ» further, we find some evidence that there is some yod element to its interpretation, as in «rākṣasa» /rakyox/ [raikkox], cf Bangla /rakkhoch/; but, this may turn out to be a lexeme-specific matter, for other Asamiya examples belie the impression that the y-glide could be a regular concomitant of «kṣ»: «rakṣā» /rōkkha/ [rokkha] and «dakṣmī» /lokkhī/ [lokkhi] are entirely straightforward and show no signs of /y/ or epenthetic /i/.

We turn to the area of velar and glottal fricatives. The regular /h/ seems not to be an issue: «hasta» /hōsto/ ‘hand’. It contrasts with the velar /x/ that matches orthographic sibilants: «saru» /xoru/ ‘thin’, «sāta» /kat/ ‘seven’, «śrābaṇa» /xrabōn/ ‘the month of Shravan’, «śrabāṇa» /xrōbon/ ‘listening’, «śruti» /xruti/ ‘scriptures’. There is some allophony that seems phonologically conditioned: «śiyāla» ‘fox’ surfaces as phonetic [cśial], which we speculate is to be phonemicized as /xial/. There are also some cases that indicate that there is more to be said about sibilants at a lexical level: «aśīlā» /ośīl/ ‘obscene’, «śrī» /sri/ ‘Shri’, «śrṣṭi» /srīstī/ ‘creation’, «śrotā» /srot/ ‘current’ indicate some pockets of sibilant retention (to speak diachronically) where further investigation is needed if we are to disentangle the phonological regularities from the lexical idiosyncrasies.

Finally, the Sanskrit-hugging phonology for «padma», /podma/, indicates that some of the sound changes do not affect the phonology of tatsamas as much in Asamiya as they do in Bangla, where «padma» surfaces as /poddo/. However, the irregular «gardabha» /gardhōb/ in Asamiya is heart-warmingly identical to the irregular Bangla form.
«gardabha» /gôrdhob/, apart from the slight adjustment of the vowel due to factors discussed in the synthesis section above.

2.0. An Interlude

It is time for a little diversion, a brief interlude characterized by a willing suspension of disbelief, during which we look at our mundane world through tinted glass and twisted mirrors that make familiar everyday objects appear to be what they are not. It is no surprise that from this distorted perspective even the familiar alpha-syllabaries (or abugidas) of the Indic scripts appear to be perfectly well-behaved alphabets, predominantly phonemic in nature, and far more regular in this respect than many of the more familiar alphabets.

Imagine a script called Bangla-Asamiya’ (BA’) identical with Bangla-Asamiya (BA) in every respect except for the fact that each grapheme in BA’ represents a single phoneme. Each consonantal grapheme, in particular, stands for a single consonantal phoneme rather than a /CV/ sequence. Assume further that BA’, unlike BA, has a visible mātrā allograph of the first vowel «a» in its alphabet. In other words, while BA has only an independent allograph of «a>, in BA’ «a» has both an independent allograph <a> and a mātrā allograph <a>. In actual orthographic practice, the salient difference between the two scripts lies in the fact that a word like /koi/ ‘a kind of fish’ is transcribed as <ki> in BA but as <kai> in BA’. Words like /ki/ ‘what’ however, are written as <ki> in both scripts.

The vowel graphemes of BA’ have regular contextual variants or allographs, exactly like those of BA, except for the difference noted above. Each vowel grapheme has an independent or base form and a combining mātrā form. The mātrā form occurs after consonants; the base form elsewhere. The consonants too display systematic allography: each consonant has a base allograph C that appears before vowels, and a combining allograph C_d (dead consonant) that appears elsewhere. Hence, a word like stri ‘woman/wife’ is represented as <s_t_r_i> at a level of representation showing allographic distinctions. BA’ is undoubtedly an alphabetic system, indeed an alphabetic system par excellence in which each grapheme corresponds to one and only one phoneme.

It is quite obvious that all occurrences of <a> in BA’ are completely redundant. Imagine a script BA”, identical with BA’, but in which all mātrā allographs of «a» are systematically deleted. In other words, whereas BA’ has <a>, BA” has only a null mātrā allograph of «a». It would be quite obvious to native users of BA’ that there was a null mātrā allograph of «a» hidden between the <b _> and the <i _> of <b_i _>. Otherwise what would explain the appearance of <i _> rather than <i f _> after a consonantal grapheme? Similarly, it would be quite
clear that there was no null allograph of «a» lurking between the two graphs in <bì बिं> from the occurrence of <ı इ> rather than <ı इ> after the <b ब>.

At this point let us pause and ask whether BA’’ is also an alphabetic script like BA’. The answer seems to be an unequivocal “yes”. If BA’ is an alphabetic script, then so is BA’’. It is quite unlikely that an almost perfect alphabetic system would turn into an alpha-syllabary merely by systematically dropping a completely redundant allograph. But if BA’’ is an alphabetic script, then so is BA; for the two are indistinguishable, and hence identical.

In the present scheme of things, consonant sequences and word-final consonants pose a practical problem for users of the Bangla-Asamiya script. How do they figure out whether there is or there isn’t a null mātrā allograph of «a» hidden between two successive consonants or after the final consonant graph in a word? In other words, how do they figure out whether the sequence of graphs <kṣ> represents «kaṣa» «kaṣa» «kaṣa» or «kṣ» in order to be able to pronounce it correctly? This boils down to the question: for each of the consonants in the sequence, how do they figure out whether it is a C or a Cd? If it is a Cd then it cannot be followed by a null allograph of «a»; otherwise, it must be. The script adopts two strategies to deal with this situation. It marks a Cd either by placing a virāma on it, or by conjoining it to the following consonant, if any, forming a yuktaśara. The latter process is recursive, and allows up to three consonants to be conjoined in this manner. As pointed out earlier, sometimes neither strategy is adopted, leaving the script user to fend for herself. This is what we find in <bl> /bollo/. But this is a quirk of Bangla orthographic practice, rather than a property of the script qua script. In the introductory remarks it was pointed out that any study of the properties of a script should ideally distinguish these properties from those of the various orthographies that apply the script to various languages. Whether the Roman script is alphabetic in nature is a question that needs to be investigated independently of the orthographic practices of languages such as English, Italian or Finnish that use the script. A comparative study of scripts qua scripts would make it necessary to abstract away from language-specific orthographic practices, without prejudice to the claim that the latter also constitute a legitimate and significant area of inquiry. Fortunately, the history of Sanskrit writing in South Asia provides a window of opportunity for a comparative study of the scripts derived from Brāhmī in abstraction from the orthographic practices of the major NIA languages associated with these scripts. Contrary to what we are often led to believe, Nāgarī was never the exclusive script for writing Sanskrit. "... Sanskrit and the Prakrits, at
different times and places were written in a vast number of forms and
derivatives of Brāhmī. In the premodern period, in other words, these
languages would be written by a given scribe in whatever happened to
be the current local script; there was no fixed definitive form for the
written language, at least as far as the outward shapes of the
characters were concerned.” (Salomon 2003, p. 70). The existence of
Sanskrit texts transcribed in the Bangla-Asamiya script – such as
many of the manuscripts of the Navya Nyāyā literature which were
originally written down in the Bangla-Asamiya script – permit us to
study the properties of the script in isolation from the orthographic
practices of Bangla and Asamiya. Orthographic anomalies such <bll>
/bollo/ are totally absent from these texts.

Let us assume that initially virāma insertion is the only strategy
adopted to distinguish between Cs and CdS. Application of the rules of
vowel allography and virāma insertion5 yields an intermediate level of
representation consisting of abstract characters. Akṣara formation is
a rather superficial, low-level and late applying process – with no
systemic import and comparable to rendering algorithms in font
technology – that applies to sequences of abstract characters to yield
graph(eti)cal output. Its sole purpose is to lump alphabetic characters
into larger display units for purely functional reasons of the sort just
discussed.

Let us rephrase and summarize what has been said so far about the
consonantal graphemes in the Indic scripts in general and the Bangla-
Asamiya script in particular. Each consonantal grapheme has an
allograph – the so-called independent or full form of the consonant –
that appears before vowels, and another with a virāma on it that
appears elsewhere. Let us call them C and Cd (dead consonant)
respectively. Graphemic «CCCV» becomes <CfCdCV> by the rules of
allography discussed earlier. The graphemic sequence underlying the
word strī ‘wife/woman’ «strī» becomes <s,d,rt>. The rest is akṣara
formation or rendering. The akṣara formation algorithm proceeds left
to right recursively as follows:

- each Cd (simple or composite) is composed with the consonant
  following it, if any, resulting in a composite C if the following
  consonant is a C, or a composite Cd if the following consonant
  is a Cd,
- each C (simple or composite) is composed with the following
  vowel mātrā, if any.7

Several features of the analysis presented here are noteworthy. First,
and foremost among them is the implicit claim that the virāma is not a
primitive of the graphological system underlying Brāhmī or any of its
derivative scripts. It is neither a grapheme nor an allograph of any grapheme of the system. It is merely a representational modification introduced by a rule of consonantal allography, and as such, can have no content, no semantics such as ‘inherent vowel killer/silencer’ associated with it. This accords well with the fact that none of the alphabets derived from Brāhmī list the virāma as an independent character in its alphabet.

We depart from received opinion and traditional wisdom in two crucial respects: (1) we maintain (on sustainable grounds) that the Indic scripts in general and the Bangla-Asamiya script in particular, are essentially alphabetic systems like Latin, Greek and Cyrillic, and (2) we assume that the characteristics of these scripts that have prompted scholars to refer to them as alpha-syllabic or aksara-phonemic are superficial rather than systemic, concerned solely with rendering and display on visual media. Our analysis accords no special status to the vowel «a» apart from granting it a null allograph that is completely predictable and regular in distribution. It seems unlikely that a perfectly well-behaved alphabetic system would turn into an alpha-syllabary merely by dropping a completely predictable allograph from its representations at the etic level. No graphological rule refers exclusively to this grapheme either, a state of affairs that might have, had it obtained, raised doubts about the correctness of our analysis.

If we are on the right track then it would appear that all of the Brāhmī derivatives retain the essential underlying architecture and dynamics of the ancient script. In fact, it would then be worth considering whether Brāhmī and its derivatives constitute a single script system associated with several distinct sets of glyphs. The differences between the Oriya and the Nāgarī alphabets, for example, would then be only at the level of glyphs and rendering.

2.1. Encoding Bangla-Asamiya in Unicode

The Unicode Standard classifies the Indian scripts derived from Brāhmī as alpha-syllabaries or abugidas. It does not recognize akṣaras. This is evident from the fact that the Bangla orthographic word <হিন্দু>, made up of two akṣaras (separated by the vertical bar), is encoded using six characters from the Universal Character Set (UCS). The closest that the standard gets to recognizing akṣaras in encoding practice is in the assignment of single code points to Indic consonants that are construed as representations of /CV/ sequences. UCS encodings of Indic texts appear to be very similar to their representations at the intermediate level of abstract characters under the scheme proposed here. But there are significant differences lurking behind this apparent similarity. Consider the following Bangla words
along with their abstract character representations (according to the analysis proposed here [DS]) and UCS encodings:

<table>
<thead>
<tr>
<th>DS</th>
<th>UCS</th>
</tr>
</thead>
</table>
| strī ‘woman, wife’ | <
c\r]|[\sdtdrt] |<s	\rt> |
| sthāna ‘place’ | <s\thān> |<s\thān> |
| tyāga ‘renunciation’ | <t\yāg> |<t\yāg> |

The only visible difference between the two sets of representations is in the encoding of dead consonants. In DS they are individual segments that are allographs of their corresponding consonantal graphemes, while in the UCS encodings they are pairs consisting of a consonant followed by a virāma. In fact, the Unicode representations could have been read as notational variants of the representations we adopt here but for the fact that the Unicode Standard recognizes the virāma as a distinct graphological entity with specific semantic properties, including the property of an inherent vowel killer/silencer. In the following section dealing with the UCS encoding of *yaphālā* we run into further problems relating to this interpretation of the virāma.

2.2. Bangla *yaphālā* in Unicode

Two *yaphālā*- and virāma-related problems arise in the UCS encoding of Bangla. It is sometimes overlooked that both are rooted in the advent of a novel graphism in the orthographic practice of Bangla and Assamiya. Consequently, attempts to address either of the problems without reference to the other turns out to be ad hoc at best.

The first of these has to do with the UCS sequence <ॉ> which is ambiguous between <ॉ> and <०>. The Unicode Technical Committee recently accepted a proposal by Nelson (2003) to disambiguate the sequence by inserting a zero width non-joiner <ZWNJ> after the <ॉ> to encode <०>, and allowing the unaltered sequence to stand for <ॉ> and <ॉ> alone.

The second problem has to do with the way <आ> and <ा> are encoded in UCS, with a virāma after the vowel in each case, as <ा> and <ा> respectively. Given the semantics assigned to the virāma in the Unicode Standard, namely that of an inherent vowel killer or silencer, a virāma after a vowel makes no sense whatsoever. In these contexts it is no more than graphic glue, an ad hoc rendering instruction with no graphological significance whatsoever. Note that Nelson’s solution, if adopted would leave this second problem untouched.

Both problems are rooted in the advent of the graphism <ा> (*op cit*) as a mātrā or combining form standing for the vowel /æ/, and <ा> or <ा> as composite independent graphs denoting the same. These are the only instances in the script of composite vowels, but instances can
be found in other Brāhmī derivatives such as Gurumukhi. With this graphism in place, it is no longer possible to construe the UCS sequence <ṛ'y> unambiguously as <ṛ>, without adopting some additional encoding convention such as the one proposed by Nelson (2003). This is so because with the advent of the novel graphism <ṛ'y> can also used to depict <ṛ迦> in <ṛ迦> encoded as <ṛ'y迦>.

From our point of view, yaphalā (ि) is not an abstract character representation of «y ि», but rather a display or rendering variant of <y ि> created in the process of akṣara formation. We may wish to look upon it as a combining form of <y ि>: the shape that <y ि> assumes when it forms a non-initial part of an phonological syllable. It cannot be a rendering variant of the allograph <y’d ि > since it must, and <y’d ि > cannot, occur before a vowel. A rendering variant is not an abstract character, and as such ought not to be included in the Unicode Character Set (UCS). Rather, the occurrence of such a variant in an encoded text should be flagged by something like a ZWJ or ZWNJ. Nelson’s proposal is thus on the right track. But in our scheme of things there is no virāma. Indeed there is none in the alphabet of Brāhmī or of any of the scripts derived from Brāhmī. We can only make sense of a virāma that is immediately preceded by a consonant by construing the pair as a notational convenience used to denote a C_d. Hence, the sequence <ṛ> <ZWJ><ṛ<y> is ill-formed in our scheme of things. However, another UCS sequence that Nelson suggests, <ṛ><y><ZWJ><ṛ>y>, for a graphism not attested in Bangla (ঃ, Reph followed by ি rather than Reph on ি) would serve our purpose quite well. Canonical rendering displays <ṛ’y> as ् retaining the full form of <y ि> rather than reducing it to its combining or rendering variant ।. The presence of a zero width joiner (ZWJ) immediately to the left of the <y ि> would block this by signalling a reduction of the <y> to its combining form ।. Once we recognize । as a reduced rendering variant of <y ि> the problem of encoding आ and आ also disappear. We encode them as <a><ZWJ><ṛ>y><a>, and <e><ZWJ><ṛ>y><e> respectively, without using a semantically anomalous virāma after the vowel.

2.3 Bangla anusvar and khaṇḍa-ta in Unicode

The Bangla anusvar ऋ and khaṇḍa-ta ऋ never occur before a vowel. Each must, therefore, be a rendered form or akṣara variant of a <C_d>, presumably of <ṛ_d ु> and <t_d े> respectively. But <ṛ_d ु> and <t_d े> have other akṣara variants as well, as evident from the second akṣara ु in बस <bṛṛ_u> and the third akṣara े in चरित <cṛṛit_u>. From our formulation of the akṣara formation algorithm in section 2.0 it might appear as though it is able to compose every C_d with the following...
consonant (C or $C_d$). In practice, this is not the case. Certain character compositions are blocked by the absence of corresponding conjunct akṣaras in the repertoire of the script. Thus, while $\hat{n}_d$ and $g$ in $<\hat{n}_d g>$ are composed together into the akṣara छ yielding the two-akṣara word सच $<\hat{n}_d g>$, this cannot happen with $\hat{n}_d$ and $s$ in $<\hat{n}_d s>$. Instead of the expected $b|\hat{n}_d s>$ we get बङ $<b|\hat{n}_d s>$. This is so, because there is no conjunct akṣara in the repertoire of the Bangla-Asamiya script corresponding to the sequence $<\hat{n}_d s>$. The facts are the same for $t_d$ and $s$ in $<t_d s>$. While $p\hat{a}t_d$ becomes पङ्ख $<p\hat{a}t_d>$ after akṣara formation, $p\hat{a}t_d lā$ ends up as पङ्ख्ला $<p\hat{a}t_d lā>$ because of the absence of any conjunct akṣara in the Bangla-Asamiya script corresponding to the sequence $<t_d>$.

When a $C_d$ (simple or composite) fails to form an akṣara with the following consonant due to the reason just described, it surfaces with an explicit virama on it. But if it happens to be either $\hat{n}_d$ or $t_d$ then it surfaces as ठ and भ respectively. Nothing special needs to be done to encode ठ and भ in Unicode. With a reasonably designed rendering engine, $<\hat{n}_d>$ and $<t_d>$ can be made to show up as ठ and भ respectively in the appropriate contexts just described. They are not characters but glyphs and therefore should not be included in the UCS segment for Bangla-Asamiya. The ठ $<\hat{n}_d>$ is already included in the UCS as U+0982 presumably because most, perhaps all, Indic scripts list an anusvara in their alphabets. Khandā-τa ष on the other hand is specific to the Bangla-Asamiya script, and Bangla or Asamiya orthographic practice provides no compelling reason for its inclusion in the UCS.

This clear picture is turned somewhat murky by attempts to encode unattested graphisms such as ठ $<t_d>$ in Unicode using UC sequences like $<t><VIRAMA><ZWJ><t>$. In the Bangla-Asamiya script there is a distinct conjunct akṣara for $t_d$, namely ठ. Hence, akṣara formation is never blocked in this context and one never encounters ठ. Be it as it may, what strategy should be adopted if the option to encode ठ needs to be retained?

We have already observed that ठ and भ enjoy the same status in the script. Neither exclusively represents a grapheme or even an abstract character in the system. They are display variants of graphs, created by the process of akṣara formation. There is no reason why one should have a distinct code point assigned to it, but not the other. The Bangla anusvara is distinct from its other Indic counterparts by virtue of being a symbol reserved exclusively for the velar nasal and none other. It does not represent a homorganic nasal. More often than not, transliterating the Nāgarī anusvar as ठ yields unacceptable results, as in
‘saint’ transliterated as সন্ত (rather than the expected সখ). In this sense too, the Bangla anusvar ൺ is distinct from its Indic counterpart and is specific to the Bangla-Asamiya script, just as khaṇḍa-ta is. Either both should be assigned distinct code points or neither. If both ൺ and ڦ were left out of the UCS then one would still have the option of forcing the appearance of ڦ (instead of the customary ڦ) in ছর্ণ <aṅg> by encoding the word as <a><h><VIRAMA><ZWJ><g>, and the option of forcing the appearance of ڦ (instead of the customary ڦ) in ছিল <citt> by using the UCS sequence <c> <i> <t> <VIRAMA> <ZWJ> <t>, options that one would probably never have to exercise in practice.

3. Summing Up

We have proposed here a three-level model of graphological description consisting of a top level of abstract graphemes, an intermediate level of what may be called abstract characters, and a level of akṣara formation. The dividing line between the last two levels is admittedly fuzzy. This tripartite organization is necessitated partly by the nature of the Indic scripts and partly by the need to keep our terms of reference close to those of the Unicode Standard, so that the latter can be examined within the parameters of the model proposed here.

The Universal Character Set encodes representations at the level of abstract characters rather than at the level of graphemes. This is evident from the presence of characters in the code set representing vowel mātrās. An alternative strategy of encoding scripts at the level of graphemes might have served the purpose of encoding just as well, or perhaps – as we are inclined to believe – even better, an issue that needs to be investigated in future irrespective of whether the findings can possibly affect actual encoding practice. Alternative models of encoding might also have been adopted, the Tibetan model, for example, in which root and subjoined consonants are distinctly represented in the UCS and the virāma is never used to trigger akṣara formation. Perhaps it would have done just as well to represent only vowel mātrās in the UCS, allowing their independent counterparts to be formed by attaching these to a dummy character. In the foregoing we observed that this strategy is not alien to Brāhmī-derived scripts associated with NIA languages. Gurmukhī uses it extensively and it might be just this phenomenon that is evident in the manner in which অ, আ and এ are formed in the Bangla-Asamiya script, using অ and ڦ as dummy characters carrying the substantive mātrās /a/ and /a/. The consequences of adopting alternative models, such as the Tibetan model, for encoding scripts like the Bangla-Asamiya need to be further investigated.
Even a cursory glance at Brāhmī and the NIA scripts derived from Brāhmī conveys the impression of a single script system realized through multiple glyph repertoires. The ISCII Standard was informed by this insight; Unicode seems to have thrown it overboard. It is necessary to explore whether an Indic unification is possible along the lines of the Han unification, covering a much smaller segment of the UCS, but many more scripts. We hope that these and other related issues will be raised and perhaps resolved at the seminar.

References

1 Names of authors given in alphabetical order of surnames.
2 Not in the traditional order.
3 In this section we look at the architecture of Indic scripts from a different perspective and this is reflected in all subsequent transliterations both at the graphemic and the graph(et)ic levels.
4 Sanskrit term for Bangla /hasanta/ /hajonto/, Hindi /halanta/ /hailont/. Inherent vowel silencer/killer.
5 We take virāma insertion to be a rule of consonantal allography.
6 This is the level of representation encoded in Unicode. Henceforth we use <> ambiguously to mark representations at this and other sub-graphemic levels.
7 This statement of the algorithm is somewhat simplified, but will do for the present purposes. We ignore several aspects of it here, in particular the fact that it has to be capable of backtracking.
8 This is a rather desirable result from the perspective of Cartesian linguistics which tries to account for the rapidity and ease with which language is acquired by assuming underlying similarities between superficially distinct language architectures.
10 Roughly a set different glyphs representing the same character. This is the sense in which the term is used in the Unicode Standard.
11 In practice, no character, once introduced, can be removed from the UCS. So it would stay, since it is there already, but its use would be deprecated.