L2/05-005

Re: Gautam Sengupta's proposal for assigning a distinct code point to Bangla Yaphalaa

Mike Meir
Director, Gate Seven Computers Ltd

Introduction

This document is a response to Gautam Sengupta's proposal, and should be read together with that document.

I will demonstrate that Gautam takes up what he accurately identifies as a problem with TUS 4.0, but then turns it into a case for assigning a separate code point to a part-ligature. This is demonstrated to be not justified, and several alternative possibilities are explored.

What is the problem being addressed by Gautam?

The problem with encoding Yaphalaa is its unexpected appearance in এ্যা and অ্যা. These are relatively neologistic full vowels used to represent non-Bengali vowel sounds, such as the a- in English "ado". Apart from this Yaphala is textually, if not in pronunciation, straightforward.
The suggested encoding for these anomalous entities in TUS 4.0 is intuitively inappropriate, as Gautam points out - in the case of অ্যা the recommended encoding is literally:

1. Take A
2. Delete A from it
3. pronounce Ya, which by now is Ja, or may perhaps be nothing.
4. pronounce Aa

This formulation seems to generate an instant feeling of "kludgedom" (technically known as "cognitive dissonance") in people coming across it, though they have been quite happy to understand what এ্যা and অ্যা imply in print, even though they are not in accordance with the normal rules of Aksara formation.

Does Yaphala behave differently to Ya/Yya?

Gautam considers that the behaviour, pronunciation-wise of Ya and Yaphala are quite different. However, he does not take account of the fact that Yya was distinguished from Ya only after 1855, and that the combined character, even when not conjuncted, has very odd properties. Yaphalas are really related to the combined Ya/Yya letter, and are related to Ya alone only by convention.
There is some confusion in the literature with regard to the letter which gives rise to Yaphala, with Chatterji, a very eminent scholar, (in Bengali Self Taught, 1927, p15), stating that Yaphala is the conjunct form of Yya, which in fact would make more sense, except that Vidyasagar, who introduced the distinction into the alphabet, definitely lists it as being associated with Ya (Varna Parichaya, 1855).
While the pronunciation (or more accurately, the non-pronunciation) of Yaphala is indeed odd, the question needs to be asked whether this is characteristic of Yaphala, or whether it has been inherited from Ya/Yya.

Chatterji, in "The Origin and Development of the Bengali Language" (1926) deals with the history of the use of Ya/Yya. *The quotes omit the phonetic transliterations in the original.*

> "Bengali orthography in late MB. (Middle Bengali) and NB. (New Bengali) times looked upon with disfavour the juxtaposition of vowels, as Sanskrit did not allow it: hence spellings like ধুআ, হআ, হওআ, খাআ, খাওআ fell into disfavour with the Pandits, and especially a spelling like হওা, খাওা where the vowel ও « ō » was treated like a consonant, with the "matra" vowel added to it. The use of য (য়) on a large scale as a letter avoiding hiatus was thus fully established in the standard form of Bengali from late MB. times: thus ধুয়া, হওয়া, খাওয়া. Further, য (য়) had become a colourless letter, a mere vowel carrier, in MB. ...
>
> "In ordinary NB. pronunciation there is not much of a deliberation, and the « y, w » glide is not ordinarily an audible sound except between low vowels (e, o, adot, abar)... There has thus been a tendency towards diphthongisation and contraction, - words like MB. শিয়াল ... being reduced to শেল, শ্যাল ....
>
> (p410):
>
> "Bengali [æ] when it comes from [e] is written এ. The subscribed «-y-» followed by «-ā-» অ্যা, য়্যা ্যা is otherwise employed. In (semi-tatsamas) post-consonantal «-y-» of Sanskrit, which became [ɛa] in MB. is written as « -yā- »; also post-cosonantal «-yă-» in initial syllables. The tendency in writing NB. standard colloquial now is to employ lavishly the য-ফলা + আ = « -yā » = ্যা: e.g. ছ্যাখে, ... for দেখে ...
>
> (p 533):
> "The letter য (য়) is much used in Bengali orthography, but it does not often indicate any sound ... The English sound of «j» as in York is unknown to Bengali, and the Bengali substitute is [i] : ইয়োর্ক্, iork, ইয়েস্, ies. The Sanskrit spelling with য় as in য়োর্ক, য়েস would not emphasise the semivowel.
>
> ... the dropping of the subscribed « -y-» in pronunciation of Sanskrit was the way in the beginning of the 12th century: but in the 7th century the « -y- » was fully pronounced: witness the spellings আর্য্য, বীর্য্য ... not আর্য্য, বীর্য্য ..."

Further evidence for this is to be found in Lambert (Introduction to the Devanagari Script, OUP, 1953), in which 3 pages are devoted to the pronunciations of Yya and Ya, without, the authors say, being exhaustive.
I think it is reasonably safe to conclude that the behaviour of Yaphala is not its own; it inherits from Ya/Yya, and what might be felt to be a series of kludges perpetrated some centuries ago, possibly becoming a little more eccentric itself along the way. So the argument does not apply.

It is to be noted that if Yaphala is to be encoded on this basis, then so should Baphola, which has similar pronunciation issues.

Allograph or display variant of YA?

Gautam states:

In TUS 4.0 the Bangla YA-phalaa is assumed to be an allograph or display variant of the abstract character YA,

This is based on the principle that in the pair, for example ট্য, the Yaphala can be separated meaningfully from the Tta, somehow carrying the virama with it, or possibly not having one, in Gautam's interpretation, and therefore maintaining its shape. The only indication of this is the forms এ্যা and অ্যা.

A phola is conceived as a part of a whole, not as a thing in itself; the term is used to identify regularities in the morphologies of ligatures, not always ligatures of consonants, and not always the final element.

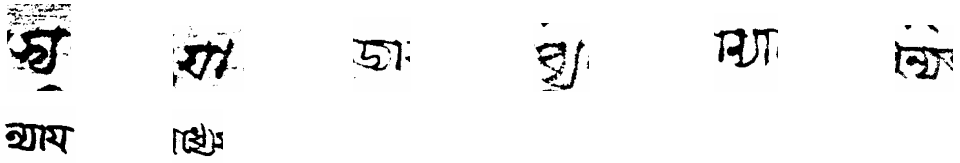
There tends to be a confusion between the concept of a conjunct ligature, and the means of realising it in print. Just because a ligature form is arrived at by, effectively, stamping standardised bits in a relationship to one another, more-or less distorting an ideal or traditional representation, does not really give the standardised parts an independent existence in their own right, outside the world of the compositor, these days usually a machine. There are signs that this confusion is not entirely resolved in the Unicode documentation.

It is therefore unwise to make assumptions about the nature of a character based purely on how it looks in print. The forms that characters take in print may be a function of the limitations of the technology involved in doing the printing.

Bengali printing has always been limited by its technology, never more so than recently, when the script has, on the whole, had to be approximated by around 200 disparate, often inadequate, elements.

The current form of the Yaphala, as a spacing, non-headline breaking, non-interactive, wiggly line placed next to a full consonant is not apparent prior to around 1860. The form developed in printing from spacing forms which are closer to the appearance of the manuscript forms from which they were derived.

Below are some Yaphalas extracted from manuscript sources, which date from the 12th to the 18th century AD. There is also a 12th century Ya for comparison:



These forms make it clear that Yaphala was, for some 600 years, stably formulated in conjunct ligatures, very similarly to the current forms ম্য, ন্য, ট্য; that is to say the differences which makes the difference are preserved while other elements may be modified. (In the case of Ya it is the beak to the left, which does not join to the upright to the right, bur rather to the usually superscripted part to the left; and the "upright" to the right which does join to the headline. The first part of the conjunct may be modified, usually by making it smaller, and actively joining it with the Yaphala).

The separated form of Yaphala is not found in the manuscripts I have studied. The phola always touches the rest of the ligature.

The ligated forms continued to be in use in high quality Bengali typesetting, based on movable type, until after 1978, and my yet be revived, now that there is no problem with having many sorts. There are good typographical reasons for using them.

The following extract from Varna Parichaya illustrates the use of both forms as alternatives. In some cases we have the traditional ligated version in the font, in some cases not, and where the font does not have the traditional form, the ligature is synthesised from two glyphs by adding the spacing wiggly line. The forms in brackets, containing the proposed character, are exactly equivalent to the traditional versions.

য ফলা - ্য

ক	য	**ক্য**	ঐক্য, বাক্য, মাণিক্য
খ	য	**খ্য (খ্য)**	মুখ্য, অখ্যাতি, উপাখ্যান
গ	য	**গ্য (গ্য)**	ভাগ্য, ঘোগ্য, আরোগ্য
চ	য	**চ্য**	বাচ্য, বিবেচ্য, পদচ্যুত
জ	য	**জ্য**	রাজ্য, বিভাজ্য, জ্যোতিষ
ট	য	**ট্য**	নাট্য, কাপট্য, নৈকট্য
ঠ	য	**ঠ্য**	পাঠ্য, শাঠ্য
ড	য	**ড্য**	জাড্য, তাড্যমান
ঢ	য	**ঢ্য**	আঢ্য, ধনাঢ্য
ণ	য	**ণ্য (ণ্য)**	পুণ্য, অরণ্য, লাবণ্য
ত	য	**ত্য**	নিত্য, সত্য, হত্যা, মৃত্যু
থ	য	**থ্য (থ্য)**	তথ্য, পথ্য, মিথ্যা
দ	য	**দ্য (দ্য)**	অদ্য, বাদ্য, বিদ্যা, বিদ্যুৎ
ধ	য	**ধ্য (ধ্য)**	ধ্যাতব্য, ধ্যান
ন	য	**ন্য (ন্য)**	অন্য, ধন্য, শূন্য, অন্যায়
প	য	**প্য (প্য)**	রৌপ্য, আলাপ্য, আপ্যায়িত
ভ	য	**ভ্য**	লভ্য, সভ্য, অভ্যাস
ম	য	**ম্য (ম্য)**	রম্য, অগম্য, বৈষম্য
য	য	**য্য (য্য)**	অজয্য, আতিশয্য, শয্যা
ল	য	**ল্য (ল্য)**	বাল্য, তুল্য, মূল্য, কল্যাণ
ব	য	**ব্য (ব্য)**	নব্য, দিব্য, তালব্য, অব্যাহতি
শ	য	**শ্য (শ্য)**	অবশ্য, আবশ্যক, শ্যামল
ষ	য	**ষ্য (ষ্য)**	দৃষ্য, পোষ্য, শিষ্য,
স	য	**স্য (স্য)**	নস্য, শস্য, আলস্য, ঔদাস্য
হ	য	**হ্য**	সহ্য, বাহ্য, লেহ্য

Note from the above that it simply is not true, as Gautam says, that ... "(Yaphala) does not ligate with a preceding consonant"; it obviously does - or more accurately, it *always* forms a part of a conjunct, which *may* then be presented as a traditional ligature in which the pieces join, or it *may* form a ligature which has the superficial

appearance and actual realisation in composition of two separate characters side by side. It is the changes in shape of arbitrary parts of the combination such that the *whole* is recognised as a conjunct which makes the difference.

Attachment of other elements indicates Yaphala is a separate character?

With regard to the point that subscripted vowel signs and rephs do not attach to the Yaphala, this really needs to be established in relation to manuscript sources. I have been unable to find any instances in which vowel signs are attached to these conjuncts in the samples of manuscripts I have, so I can draw no concussions, but others with better knowledge and access may wish to explore this.
Rephs in Bengali, unlike those in Devanagari, are traditionally attached above the character for which it is to be realised - i.e., the first part of the conjunct (though the rule is rather loose), as in র্ব, so one would not expect to find them on the Yaphala, especially as the Yaphala is not realised as a consonant.

Summary: Significance for encoding:

The grounds given by Gautam for change do not seem to be correct and/or sufficient. Excepting that a solution is needed to encode অ্যা and এ্যা the current encoding is a proper representation of the conjunctions in the text, and can best be left alone.
The pros of this approach are that the acknowledged problem can be investigated in its own right, and can be dealt with at its own level, without the imposition of a global solution which goes against historical and conventional practice.
I demonstrate below that the anomalous forms can be encoded in a way which respects their identity. I cannot think of any cons for doing this, if my reasoning is accurate.

Use of Yaphala in অ্যা এ্যা.

In TUS 4.0 these objects are treated as strings. It is in converting them to conventional strings that the problems arise. As strings they violate the Aksara model fairly conclusively.
An alternative interpretation however, that they are not strings, they are gestalts, indivisible wholes, which share morphological features, obviously, with other letters, and which have a functional relationship to these morphologies in their effects on pronunciation, but not actually at the textual level. The indication to the reader is to modify the sound of the full vowel "a bit like you do when you see a Yaphala in text"; in a sense it is a meta comment about the pronunciation rather than a representation of the pronunciation.
This is in fact what a Yaphala, or a Baphola is to a reader in the course of normal reading: it will be taken as advice, not of a conjunction to be pronounced in the normal way, but of a complex transformation of surrounding letters. However, Unicode is not about pronunciation, it is about text, and in text, a Yaphala as such is always an element of a textual ligature.
In fact, the অ্যা and এ্যা are understood not to be ligatures, but indivisible units. The evidence regarding this comes from the contrasting sorting behaviour of অ্যা and [consonant] ্যা. অ্যা is sorted separately. It does not sort as অ + ্ + া.
An example is to be found in Chatterji, op. cit., at page 1088, where, in the index, আহ্বান is followed by অ্যা, অ্যাদ্দিন and অ্যায়, then ই. Similarly এঁষানি, এঃ, এ্যা, ঐ. The

current practice seems to be to sort words beginning with অ্যা after ঔ, ক following. This suggests that the object is considered even more a full vowel in its own right. Both conventions are consistent with an intuitive and practical understanding that অ্যা stands alone as a unit and is not equal to অ + ্ + া. It is like আ, which is not equal to অ + া, even though in 8-bit fonts it may in practice be composed by using অ + া.

On the other hand ্যা preceded by a consonant sorts as a string; so প্যা = প + ্ + া = প + ্ + য + া. This is the normal behaviour for a conjunct with an akar attached.

The problem is therefore reduced to how to encode independent objects of the form অ্য, অ্যা, এ্যা...

(The first of these occurs in Chatterji to represent the sound of -u- as in English sun, but may not be used anywhere else; the other two are common.)

Encoding অ্যা এ্যা ; Two Possible Solutions:

Give each of these forms its own code point.

This solves the problem, while raising several more:

1) we do not know quite how many of these objects there are; we may need to add one for each full vowel, to satisfy the need for letters to be used in putative transliterations.
2) they have not been included in the "official" alphabet, if such a thing exists;
3) There is a potential problem with keyboarding if we add several new and independent characters to the repertoire. While keyboarding is not central to encoding, the likely effects of a particular encodings need to be considered.

Approach the issue more generally

The meaning of these entities to the reader is straightforward: "utter or think a modified vowel sound". Essentially the alphabet is being extended in a more-or-less fixed way. The Nukta is the character which is available to extend the alphabet, so we could establish a principle by which these extensions are *encoded* by the full form of a vowel followed by a nukta, but *presented* as desired:

অ্য = FullA Nukta

অ্যা = FullAa Nukta

এ্যা = FullE Nukta

Pros of this approach are

1) it can legitimately be applied to any full vowel,
2) it breaks nothing and
3) works now from the current keyboard layout, provided the necessary glyphs are in the font.
4) It is also in principle consistent with the work of Vidyasagar in standardising the alphabet to make it representative of Bengali - ড় is a way of indicating that ড needs to be pronounced the way Bengalis pronounce it medially or finally, and English loan words are now, effectively, part of Bengali.

Cons are

1. a Nukta is understood to be a non-spacing dot. The proposal is therefore counter-intuitive.
   Answer: It is no more counter-intuitive than the alternatives. There is only one new thing for a typist to learn, and the glyphs get produced with two keystrokes rather than at least three. Other abstract characters have presentation forms which differs from the norm: examples are seen in রু, রূ, প্রু, হ্ল, হ্ব, হ্ম, ক্ষ্ণ . Readers, and typists are not really interested in the abstract representation of what they see, which is a sequence of bits, not a dot, in fact.
2. Dotted Vowels already exist in some contexts, particularly the transliteration of Arabic.
   Answer: Only FullA dot is used for Arabic transliteration, and this can be dealt with by the convention that Char + Nukta -> Ligated form, Char + ZWNJ + Nukta -> DottedChar.

My personal preference would be for the first solution. The second is more elegant and future-proofed, but the first is more in accordance with what current dictionaries appear to be indicating.