Review of Bengali Khanda Ta and PRI-30 Feedback

Peter Constable, Microsoft Corporation
2004-06-15

Abstract

Public Review Issue 30 (PRI-30) discussed the question of how to encode Bengali khanda ta. This document reviews some key issues and feedback that was received on PRI-30, which included new information that has bearing on arguments for or against different alternatives. In light of the feedback, a strong case for encoding khanda ta as a separate character is presented.

Introduction

The need for PRI-30 arose due to a problem in the description of the encoded representation of khanda ta in Unicode 4.0 and Indic FAQ #19. The problem was due to a description that did not fully specify interpretation of all possible sequences related to the khanda ta, and inappropriate association of khanda ta with half forms.

When software developers at Microsoft and elsewhere implemented khanda ta like a half form, incorrect cluster behaviour resulted; for instance, the e-kaar matra in a sequence like < TA, VIRAMA, ZWJ, TA, VOWEL SIGN E > would reorder around the khanda ta, where the desired behaviour was to have the e-kaar position between the khanda ta and the following ta. To resolve this, developers devised an encoding sequence of < TA, VIRAMA, ZWJ, ZWNJ > to represent a non-clustering khanda ta. Because this was not specified in Unicode, however, this led to efforts to clarify what the intended encoded representation and behaviour for khanda ta is or should be.

PRI-30 presented four alternate encoded representations (along with a complete specification for rendering and clustering behaviour for each), which covered the alternate possibilities that seemed to be worth of consideration at the time. It also presented arguments for and against each, based on current understanding of the issues. The four alternative models presented were:

A. Khanda ta represented as < TA, VIRAMA > with ZWJ for explicit override

B. Khanda ta represented as < TA, VIRAMA > with ZWJ, ZWJ + ZWNJ overrides

C. Khanda ta represented as < TA, VIRAMA, ZWJ >

D. Khanda ta encoded as a separate character

Model A is the model that, it appears, was currently intended by the specification in section 9.2 of Unicode 4.0 for representing khanda ta. After considering pros and cons of each alternative, PRI-30 presented model A and the recommended solution. Of course, the public review issue

gave opportunity for this straw-man position to rebutted, or for entirely new alternatives to be presented.

Feedback on PRI-30

There was not a lot of feedback received; this may be due in part to a general problem of South Asians not being engaged (for whatever reason) in the process of developing the standard. Nevertheless, feedback was received from the Indian Ministry of Information Technology (a full-member organization) and some other informed contributors.

Feedback fell into two groups:

- There was two contributions from outside South Asia that supported the proposed solution, model A. One added an additional argument against model D related to data conversion between ISCII and Unicode: a new khanda ta character would not have a corresponding distinct representation in ISCII.

- Feedback from South Asians (submitted via the submission mechanism, as a UTC doc, or unofficially on the Indic list) has been very strongly in favour of model D, encoding a separate khanda ta character.

With regard to the former feedback, it should be noted that the claim regarding ISCII is incorrect: ISCII reportedly does have a distinct representation for khanda ta: < TA, VIRAMA, VIRAMA >. If anything, then, an argument based upon conversion to or from ISCII actually favours models such as D that give khanda ta a distinct representation.

The request from South Asians for a separate khanda ta character is not new: it goes back at least three years, to the feedback that the Indian government gave on Unicode 3.0 (L2/01-304), and has been a recurring topic on email discussion lists. Regrettably, most of the requests for a new khanda ta character have not been accompanied by a technical justification for why existing representations are inadequate and a new character is needed.

On this occasion, though, feedback from one contributor, Gautam Sengupta (L2/04-192) did present some technical argumentation for encoding a new character. A key element of the case made for a new character is that the khanda ta is graphemically distinct from other forms of ta. This is new information that was not previously available, and has some bearing on how alternatives might be evaluated.

In light of the feedback on PRI-30, then, there are primarily two alternatives in focus: model A, the current specification (albeit poorly documented), and model D.

Khanda ta as a grapheme

The claim of grapheme status for khanda ta in L2/04-192 was accompanied with some linguistic evidence, though how this evidence supports the claim was not made clear. Such evidence is not regularly required for encoding proposals, though this case might be considered exceptional by some since there had been a prior understanding that khanda ta was not a distinct grapheme

from ta. There is some evidence from a few different sources that is reasonably convincing, though:

First, independent feedback from two Bengali experts (both in university positions) clearly indicated that khanda ta is considered a distinct grapheme. This was also affirmed by several other Bengalis.

Secondly, recent dictionaries published by the Bangla Academy list khanda ta separately in the list of consonants, and have begun to sort khanda ta differently than ta:

বর্তমান অভিধানে ৎ-কে স্বতন্ত্র বর্ণের মর্যাদা দিয়ে ত এবং ত-এর যুক্তবর্ণের পরে বিন্যাস করা হয়েছে।

("In this dictionary, khanda-ta has been treated as a unique/independent (swatantra) alphabet (barna), and it has been positioned/sorted after ta and the conjuncts of ta.")[1]

বর্ণানুক্রম
(alphabetical order)

সাধারণ বর্ণানুক্রম
অ আ ই ঈ উ ঊ ঋ এ ঐ ও ঔ   ক খ গ ঘ ঙ চ ছ জ ঝ ঞ ট ঠ ড ঢ ণ ত থ দ ধ
ন প ফ ব ভ ম য র ল শ ষ স হ ড় ঢ় য় ৎ ং ঃ ঁ

Figure 1. Ta and khanda ta listed as separate consonants in the alphabetical order (Bangla Academy 1994b)

Thirdly, there have been linguistic changes that, as I understand them, might be conducive to creation of a grapheme distinction:[2] in at least word-final position, the inherent vowel is regularly dropped, though spellings have not been revised to reflect this. This results in contrasts in identical environments in the sound-symbol relationships of ta and khanda ta:

মত /mOt/ 'opinion'
মৎ /mOt/ 'my'

Figure 2. Ta and khanda ta: sound-symbol contrast in identical environments

Since in cases like the first word in Figure 2 the consonant ta is already phonologically vowelless, it becomes a less appropriate description to say that "khanda ta is the dead form of ta". This would seem to provide motivation, therefore, for considering khanda ta a distinct grapheme.

Finally, the relationship between khanda ta and ta-halant (or the /t/ component in t-C conjuncts) is similar to that of anusvar and nga-halant: these are graphemically-distinct spellings for the same sound,[3] and each occupies a distinct position in the "alphabetical" order. For instance, Figure 3 shows words having alternate spellings, using anusvar versus ng-C conjuncts; yet the

---

[1] Bangla Academy (1994a), p. 12. Quotation and translation courtesy of Sayamindu Dasgupta.

[2] This hypothesis is based on my limited knowledge of Bengali language and script, and there has not been adequate opportunity to have these ideas reviewed by Bengali experts.

[3] Unlike its counterpart in some other Indic scripts, the Bengali anusvar does not stand for a homorganic nasal, but rather only for a velar nasal, /ŋ/.

pronunciations of these representations is the same; Figure 4 shows the different order for nga and anusvar used by the Bangla Academy dictionaries:

Figure 3. Alternate spellings using anusvar versus ng-C conjuncts (Bangla Academy 1994b, p. 769)

Figure 4. Nga and anusvar listed as separate consonants in the alphabetical order (Bangla Academy 1994b)

Clearly anusvar and nga have long been considered distinct graphemes. Given the similarities, it should come as no surprise that khanda ta is considered a distinct grapheme.

Encoded representation of graphemes and khanda ta

While some users have on occasion requested the encoding of text elements on the basis that they are considered graphemes for a particular language, grapheme status has never been considered a sufficient condition for encoding a text element as a distinct, atomic character. Rather, it is assumed in general that software implementations should be able to provide appropriate behavior that reflects user expectations regarding what text elements are considered graphemes. For instance, by giving the sequence < c, h > a single sort weight greater than that for the character < c >, a Spanish sorting implementation can provide behaviour that reflects a user expectation of "ch" as a distinct grapheme without requiring that the grapheme be encoded as an atomic character.

At the same time, Unicode has also provided means to ensure that graphemes can be given distinct encoded representations if needed to avoid ambiguities. For instance, if a grapheme "ch" needs to be distinguished from coincidental occurrences of "c" followed by "h", the character U+034F COMBINING GRAPHEME JOINER can be used in the sequence < c, CGJ, h > to indicate the grapheme "ch", providing a distinct encoded representation from < c, h >.

In order to understand and evaluate the case made in L2/04-192, we need to consider the current encoded representation of khanda ta in light of the foregoing discussion of encoded representations for graphemes. Let us take it as given that khanda ta is considered a distinct grapheme in Bengali (which appears to have become the case).

As currently specified in Unicode and proposed (as model A) in PRI-30, a khanda ta may be represented by the sequence < TA, VIRAMA, ZWJ >. It may also be represented as < TA, VIRAMA >, but (and this is the critical point) only in some contexts: it depends upon what characters follow. In other words, if a process needs to operate on khanda ta, distinguishing it from other graphemes, it cannot do so without enumerating all possible contexts.

To this extent, the situation is similar in some respects to the example of coincidental combinations of "c" followed by "h". There are significant differences, however.

First, for languages that have digraph graphemes such as "ch", it is generally the case that coincidental occurrences of the separate graphemes juxtaposed would be fairly rare. In contrast, while the Bengali khanda ta is not used with a very high frequency, it is hardly rare.

Secondly, for cases like Latin digraphs, we have seen that CGJ can be used to disambiguate from otherwise-similar grapheme sequences. Khanda ta is not a digraph of base characters, however: under model A, it is represented as a combining character sequence. Use of CGJ within combing-mark sequences has been sanctioned in Unicode to prevent re-ordering of combining marks under canonical reordering, but that is not the issue here. There is no function for CGJ defined that can be used to disambiguate different graphemes otherwise represented with the sequence < TA, VIRAMA >.

As mentioned in the previous section, the relationship between khanda ta and ta is analogous to that between anusvar and nga. Just as anusvar and nga have distinct and unambiguous encoded representations, so it is felt by users in South Asia that the same is needed for khanda ta and ta.

The appeal for a khanda ta character

The basic functional requirement that is being asked to be met, therefore, is to have a distinct encoded representation for khanda ta that reflects its status as a distinct grapheme. The user community has indicated a need in order to facilitate text processes:

- L2/04-192 mentioned specifically the need to search for khanda ta without getting matches on other graphemes. This can be achieved without an unambiguous representation, but only by creating a complex query using Boolean logic to suppress all of the unwanted contexts.

- While not mentioned in L2/04-192, there is a need for the khanda ta to sort differently from "regular" ta. Again, this can be achieved without an unambiguous representation, but only by assigning atomic weights to every sequence beginning with < TA, VIRAMA >.

- In feedback on PRI-30 submitted by Manoj Jain (Ministry of Information Technology, India), there was reference to the need for a distinct representation (in fact, a distinct character) in order to facilitate various types of processing "in both scientific (eg NLP, Computational linguistics) and commercial applications". While no specific details are provided, it is not unreasonable to suggest that a distinct grapheme may require a unique representation to facilitate certain processes. (Of course, it is not hard to imagine that a distinct, atomic character might be considered more convenient to work with.)

If it is accepted as a reasonable requirement that a unique representation should be available for the grapheme khanda ta, the question then remains as to what unique representation may be provided. Various possibilities exist, but the one preferred within the user community by far is to encode a new, atomic character for khanda ta.

Alternatives for unique representation of khanda ta

Encoding a distinct khanda ta character is not the only possibility, and alternatives should be evaluated.

In relation to the models presented in PRI-30, models A and B do not (without further mechanisms) provide a means to uniquely represent khanda ta. They will therefore not be considered further here. (A re-evaluation of model A will be discussed in a later section.)

One alternative is to have the khanda ta encoded always and only using the sequence < TA, VIRAMA, ZWJ >. This is similar to model C in the PRI-30 document, though that presentation introduced a contrived half form of ta which is unknown in actual usage.[4] As this innovation is neither needed nor wanted, it can be dispensed with.

In the revised model C, then, interpretation of difference sequences would be as summarized in the following table:

Pattern	Sample text sequence	Display	Note
< ta, virama, Cj, ekaar >	< ত, ্, ত, ে >	ো	
		ত্ে	fallback rendering
< ta, virama, Cn, ekaar >	< ত, ্, ক, ে >	ত্ে	
< ta, virama, # >	< ত, ্, # >	ত্ #	
< ta, virama, ZWNJ, * >	< ত, ্, ZWNJ, * >	ত্ *	
< ta, virama, ZWJ, * >	< ত, ্, ZWJ, * >	ৎ *	

Table 1. Interpretation of sequences under revised model C.

The main argument in favour of this solution is that it supports the necessary distinctions and behaviour without requiring new characters. It has some significant drawbacks, however:

- It involves an exception to prototypical behaviour of ZWJ in Indic scripts in that ZWJ is used to represent a letterform that has a following obligatory cluster break. Of course, the functions of ZWJ in Unicode in general are not without exceptional cases. Even so, giving it contrary behaviour, particularly in scripts where it is so heavily used, is not a wise choice since it can lead to confusion and incorrect implementations .(A case in

---

[4] This was done for no other reason that to provide completely uniform behaviours for various types of character sequence.

point is the implementations mentioned in the introduction that ended up with incorrect cluster behaviour.) The special cases within Indic scripts are already complicated for implementers to understand. Therefore, rather than increasing the number of special-cases, there should be a conscious effort to keep such proliferation in check.

- It uses a three-character sequence that, for most implementations, would require three presses of a backspace key to delete by backspacing, and the intermediate steps would convert the khanda ta into various forms of "regular" ta, which may create confusion for many, perhaps even most, users.

- The three-character sequence including ZWJ would be used for all instances of khanda ta. The khanda ta is not a high-frequency character, but it also is not at all rare. Heavy use of ZWJ would be prone to lead to common usability issues for users,[5] and should be kept to more exceptional cases. (This point was made by one of the contributions in response to PRI-30 in favour of model A.)

- It does not completely eliminate the problem of distinct representation. For instance, whereas before < TA,VIRAMA > could represent khanda ta but would not exclude other text elements, so now < TA,VIRAMA > would represent ta-halant, but would not exclude khanda ta.

A further variation of model C would have khanda ta represented always and only as < TA, VIRAMA, ZWJ, ZWNJ >. This would avoid the problem of ZWJ used in a sequence that is followed by an obligatory cluster break, but at the cost of using a very long sequence with multiple control characters for a not-infrequently-used text element. The arguments just made in relation to usability stand all the more strongly against this alternative.

Another alterative is to represent khanda ta using COMBINING GRAPHEME JOINER as part of a sequence with < TA, VIRAMA >. As mentioned earlier, this is intended for use in creating grapheme distinctions in encoded representation. As also noted earlier, though, the usage that has been specified for CGJ does not include base + mark sequences, and the specification would need to be expanded to additional sequence types. Moreover, in this case the CGJ would not be "joining" two elements of a digraph to form grapheme; it would simply be a flag whose presence or absence provided a Boolean distinction between two graphemes. This would be a radical departure from the current functional semantics of CGJ and could lead to problems in implementations as well as confusion for users.

Failing these alternatives, then, the only other obvious possibility is to encode a separate khanda ta character. The entire discussion thus far has constituted a case for this solution. What remains is to re-evaluate the arguments in PRI-30 for or against models A and D.

---

[5] For a somewhat hypothetical but analogous comparison, imagine that computers had existed when the innovation of graphemic distinction between "i" and "j" occurred, and that representation of "j" required a three-character sequence. Software engineers would certainly work to shield users from unpleasant side effects, but users likely would still be negatively impacted.

Re-evaluation of arguments in PRI-30

As mentioned, of the alternatives presented in PRI-30, it is models A and D that different parties have considered viable. It is appropriate, then, in view of feedback received on PRI-30 to re-evaluate the arguments made for and against each of those models in that document.

First, one of the arguments in PRI-30 for model A and against model D has to do with avoiding change with the current intended specification. This should be put into better perspective:

- There is no debate that the currently specification is insufficiently clear and would, in any event, need to be revised editorially if nothing else.

- The argument assumes a need to limit unnecessary impact on existing implementations, yet it is known that the problems in wording of the current specification have led to novel implementations that do not conform to the specification. Even if it is clear what an conformant implementation should be like, there is no indication that such an implementation exsists, or of any implementation that would not have to undergo changes by preserving model A

One of the arguments presented in PRI-30 against model D but which supported model A was based on an assumption that it would be problematic to break phonological and graphological connections between khanda ta and "other forms of ta". This argument, however, could only have been put forth in the absence of any knowledge that khanda ta is, in fact, considered to be a distinct grapheme from ta. In light of information received in response to PRI-30, this argument is undermined since its premise has been called seriously into question.

PRI-30 presented two arguments *against* model A (though not directly for model D). One had to do with problems in rendering implementations related to font-internal information, yet as mentioned in the appendix to PRI-30 that issue exists and is in need of a solution independently of the khanda ta question. The other argument is still applicable, though: model A requires an exception to prototypical behaviour of ZWJ in Indic scripts in that ZWJ is used to represent a letterform that has a following obligatory cluster break. This problem was shown above to exist for model C (and its variants), and the elaboration of that argument provided above is equally applicable here. It should be noted that the problems in implementation of khanda ta that immediately led up to the current discussion arose precisely because exceptional behaviour for ZWJ was specified. It seems clear, then, that it would be preferable to avoid exceptional behaviour such as this whenever possible.

Finally, it was observed that both models A and D provide the necessary visual distinctions and clustering behaviour, but that model A does so without introducing a new character. It must be noted, though, that this is possible only at the expense of failing to provide an unambiguous representation for khanda ta, as has been argued is needed in light of its graphemic status (a point that was not known to me when the PRI-30 document was written). The difference here is whether visual distinctions alone are required, or whether graphemic as well as visual distinctions are required.

This is the primary argument for model A in opposition to model D. The strength of the case for model A, then, lies in rejecting the graphemic status of khanda ta, or of the need to provide a

distinct representation for the grapheme khanda ta. Turning the issue the other way around, the strength of the case for encoding a distinct khanda ta character and against model A relies on acceptance that khanda ta is considered a distinct grapheme and that, as such, it requires an unambiguous encoded distinction (as is available if needed for graphemes in European alphabetic scripts). The evidence for distinct graphemic status of khanda ta is of a nature that could not be easily contested. To attempt to do so would require a rather longer discussion, and a debate that would, I think, ultimately come down to the practical consequences, namely, whether or not there is a need for adequate text processing to provide a distinct, unambiguous representation for the khanda ta. The case for such a need has been presented in earlier sections of this document.

On re-evaluation, then, the case for model A in opposition to model D presented in PRI-30 was not as strong as it may have appeared. Moreover, it is significantly affected by feedback in response to PRI-30 pointing out the distinct graphemic status of khanda ta. In this new light, then, the case between model A and model D hinges on one issue: is a distinct and unambiguous representation for the grapheme khanda ta needed in order to facilitate adequate support for khanda ta in various text processes? The familiar answer to such a question in usual cases is to say that other mechanisms exist for that purpose. As has been shown, however, this is not a usual case. Alternative mechanisms have been considered yet found to have shortcomings.

Perspective on the case for encoding khanda ta

In this document, I have tried to speak on behalf of the Bengali user community, providing a case for encoding a distinct khanda ta character, which they deeply desire, as best I know how. Of course, there is always a "flip side" to a proposed change. In this section, I will attempt to bring a, hopefully, fair perspective on the case for encoding khanda ta by briefly considering the "flip side" of the matter.

The case for encoding a khanda ta character hinges on the khanda ta being a distinct grapheme. Some may question how important a need this is, particularly given that the prevailing perception until recently was that the khanda ta is no more than "the halant form of ta". It was shown earlier, though, that it is given distinct status in influential and authoritative sources such as the Bangla Academy's dictionaries. Some may yet question how great the need can be to distinguish it given that it is clearly historically related to and derived from the halant form of ta. Of course, though, there are well-known cases from both Indic and European scripts of grapheme-genesis; and at some point, we need to give some consideration to the combined weight of multiple native-speaker experts telling us that this is really the current cultural understanding.

If status as a distinct grapheme is accepted, the next line of rebuttal would be to point out that there has never been a guarantee that Unicode will provide a unique representation for every distinct grapheme. Of course, this is true. Even so, it was pointed out that mechanisms *have* been provided that can be used in familiar cases such as Latin-digraphs to disambiguate sequences, yet those mechanisms were not designed with cases like khanda ta in mind, and so

Latin-digraph cases are unequally favoured relative to a case like khanda ta. Additionally, what is being sought here is simply to provide for khanda ta the distinctness already provided to the anusvar, which has comparable graphemcial status to khanda ta.

If the need for a distinct representation for khanda ta is accepted, some will still question the choice to encode a new character. There are several reasons why this might be the case.

First, it was pointed out by Stefan Baums on the Indic list that a new character will make transliteration between scripts more difficult. Specifically, in transliterating from another script into Bengali there would be difficulty in mapping one entity in the source script onto distinct entities in Bengali script. The response from users in India is that this is a lower priority for them than support for khanda ta in other processes. I would add that, as I understand it, khanda ta was not found in older documents, and so would not normally be used in Sanskrit-language text, which would cover what is perhaps the most common transliteration scenario. (Indeed, display of Sanskrit-language text without any khanda ta would be simplified in that it would no longer require special font support.)

It was also pointed out in feedback to PRI-30 that a new character would lead to problems in conversion to or from ISCII. This is not actually true. In fact, ISCII has an unambiguous representation for khanda ta: < TA, VIRAMA, VIRAMA >. If anything, the question of conversion to or from ISCII would provide support for an unambiguous encoding of khanda ta in Unicode and, to that extent, supports the Indian case for a new character.

There is a rather more serious concern in that a new character takes time to go through the standardization process, which cannot be by-passed, yet implementations using *some* encoded representation for khanda ta will exist in the interim. In the long term, therefore, multiple implementations will need to be supported. Indians have responded that they are prepared to deal with the inconveniences this would cause. Of course, it is not necessarily they alone that would be affected: this may be a legitimate concern for any software vendor implementing Bengali support. I believe the response of users would be that lack of what they consider to be adequate support for khanda ta would be, for them, a rather greater concern.

Finally, it is of course possible that one might feel the arguments for rejecting some alternate unambiguous representation. It seems that the only likely possibility of this would be if someone were maintaining a preference for model A. As has been shown, though, the case for model A and against model D is based solely on a view that adequate support for Bengali text processing does not require distinct and unambiguous representation of khanda ta. I believe the response from South Asians, who have been doing Bengali text processing as long as anybody, is that that is not the case. The argument to that effect was presented above.

Conclusion

I have attempted to state the case for encoding a separate khanda ta as strongly as possible, and to put it into a fair perspective by also citing reservations that may exist. The issue is not a simple one, and there are valid concerns from both sides.

Up to now, the amount of careful analysis has not kept pace with the volume of words exchanged. At greatest risk has been that a decision would be made rejecting the Indian request without as best a case as possible having been made. This could only lead to a widening gulf of distrust between users in the Indian sub-continent and supporters of the Unicode program. At the very least, I hope to have shown that the case for a separate character is not completely without merit. Perhaps the analysis has revealed a case that is sufficiently convincing to grant the new character that has been requested, though I do not take that as assumed. At best, I hope to have provided a fair hearing for the Indian request such that, regardless of the outcome of a UTC decision, users in India will feel that their needs have been considered thoroughly, and that they can feel some confidence that implementations can be provided that will meet their reasonable needs.

It has been shown that the heart of the issue lies with whether one believes that a distinct and unique representation for the grapheme khanda ta is needed for it to be adequately supported in a variety of text processes. Prior assumptions that khanda ta was not a distinct grapheme have been shown not to reflect current realities, and it is important that pre-existing misconceptions not interfere with sound decisions. The Indian Minsistry of Information Technology as well as implementers and users from South Asia have indicated that the text processing issues are real, and go beyond the need to enter khanda ta as a single keystroke. Any decision on resolving PRI-30 must give careful consideration to the usability and implementation concerns that have been presented here.

References

Bangla Academy. 1994a. বাংলা একাডেমী বাংলা বানান অভিধান (Bangla Academy Bangla Banan Abidhan = 'Bangla Academy Bangla Spelling Dictionary'). Dhaka: Bangla Academy.

——. 1994b. Bangla Academy Bengali-English dictionary. Dhaka: Bangla Academy.