Proposal to Encode Alternate BA for the Bengali Language

Vinodh Rajan vinodh@virtualvinodh.com

Deepro Chakraborty deepro@ualberta.ca

27-July-2022

Introduction

Current mainstream Bengali orthography does not possess the equivalent of the Brahmic letter VA[1] (in Unicode parlance). Both VA and BA in Sanskrit-origin words are orthographically spelled using U+09AC ব (BENGALI LETTER BA). As a result, there is no way to distinguish them in the resulting text. While this isn't an issue for writing the Bengali language (as it lacks /v/ entirely), the conflation presents a problem with Sanskrit texts, where they are distinct phonemes and require clear orthographic distinction.

(The Bangla Academy recommends [7] the orthographic differentiation of, at least, the clusters using an explicit virama. For instance, দ্ব /d̪bɔ/ vs দ্ব /d̪[ː]ɔ/[2] (< /*d̪ʊɔ/). Seemingly, this convention is not widely followed [4]).

This lack of a traditional letter to represent the letter VA is common to Assamese, Oriya, and Bengali. The former two eventually innovated new letters to differentiate the two phonemes by using slightly modified forms of BA to represent VA. Assamese innovated U+09F1 ৱ (ব BA with lower diagonal) and Oriya U+0B35 ଵ (ବ BA with nukta).

Differentiating VA and BA in the Bengali orthography

In a similar vein, some Bengali publications[3], particularly those printing Sanskrit works using the Bengali orthography, also wanted to distinctly denote the two phonemes in their texts. As a result, they innovated a new letter to distinguish VA from BA in their publications. Unlike Assamese and Oriya, this was however done by using a distinct letter for BA: ৱ (ব with an inner diagonal stroke) and ব was relegated to only represent VA.

From [4]:

---

[1] It is variously realized as /v/ or /w/ depending on the region

[2] Word-initially /d̪ɔ/ but word-internally /d̪ːɔ/

[3] At least one linguistic publication [5] [8] has used ভ় (BHA + nukta) to denote /v/. But this has not been widely adopted by publishers. One plausible reason might be that an average reader will likely ignore the nukta and will simply read it as /bha/.

জ্যায়সী চেৎ কর্মণস্তে মতা বুদ্ধির্জনার্দন ।
তৎ কিং কর্মণি ঘোরে মাং নিয়োজয়সি কেশব

jyāẏasī cet karmaṇaste matā buddhirjanārdana .
tat kiṁ karmaṇi ghore māṁ niẏojaẏasi keśava

It has been claimed such distinction can be traced back to the pre-modern/medieval era.

From [6]:

ba and va

Rule: 23: In the earlier language, both the shape and pronunciation of the antyastha (sic) va were different than the vargīya ba. In Sanskrit texts handwritten by pandits the shape of a vargīya ba is written as ব. But in the Bengali printed alphabet, both ba and va are identical and because of their similar shape, their pronunciation has also become similar. In order to prevent this undesired thing, I have emended the letter ব. Afterwards, the letter ব should be pronounced like the English letter V and ব should be pronounced like the English letter B."

This is contentious, as there is no evidence in the manuscripts to claim the same. This usage was most likely inspired by the modern Devanagari letter forms - BA (ब) vs VA (व).

This new alternate BA shows analogous behavior to that of ব in terms of vowel sign placements and the shapes of the various conjuncts as seen below. (The default behavior of the similar-looking U+09F0 is also shown for comparison)

	U+09AC	Bengali Alternate BA	U+09F0
	ব	ৰ	ৰ
Cu	বু	ৰু	ৰু
Cū	বূ	ৰূ	ৰূ
rCa	র্ব	র্ৰ	র্ৰ
Cra	ব্র	ৰ্র	ৰ্র
Cru	ব্রু	ৰ্রু	ৰ্রু
Cya	ব্য	ৰ্য	ৰ্য
Cda	ব্দ	ৰ্দ	ৰ্দ
Cdha	ব্ধ	ৰ্ধ	ৰ্ধ
mCa	ম্ব	ম্ৰ	ম্ৰ
CCa	ব্ব	ৰ্ৰ	ৰ্ৰ

There are several Sanskrit publications that have used this character to print their Sanskrit texts. There are also few blogs that advocate using this distinction for the Bengali language. Clarifying the representation of the character in Unicode will facilitate the digitization of those published works. It will also allow anyone to digitally represent Sanskrit texts in the Bengali orthography without any loss of textual fidelity.

Representing the alternate BA in Unicode

Re-using U+09F0 Assamese RA

This is probably the most straight-forward solution. The Assamese RA already encoded at U+09F0 as BENGALI LETTER RA WITH MIDDLE DIAGONAL is visually similar to that of the Bengali alternate BA, even though the behavior of the two letters varies considerably as seen above. While some of the expected behavior can be achieved using joiners, others are problematic.

	Bengali Alternate BA	Sequence
bu	ৰু	U+09F0 ZWNJ VS-U
bū	ৰূ	U+09F0 ZWNJ VS-UU
rba	র্ৰ	U+09B0 VIRAMA U+09F0
bra	ৰ্র	U+09F0 ZWJ VIRAMA U+09B0
bru	ৰ্রু	U+09F0 ZWJ VIRAMA U+09B0 VS-U
bya	ৰ্য	U+09F0 ZWJ VIRAMA YA
bda	ৰ্দ	?
bdha	ৰ্ধ	?
mba	ম্ৰ	?
bba	ৰ্ৰ	?

Issues arise when we try to represent the conjunct forms. Unicode (or at least the fonts) treats U+09F0 as an allograph of U+09B0 Bengali letter RA. As such, U+09F0 forms a *repha* at the cluster initial positions and a *ra-phalā* at cluster-final positions, imitating the behavior of U+09B0.

ৰ্ক /rka/ - U+09B0 VIRAMA KA (OR) U+09F0 VIRAMA KA

ক্ৰ /kra/ - KA VIRAMA U+09B0 (OR) KA VIRAMA U+09F0

As an allograph of RA, as per Unicode conventions, U+09F0 should produce a *repha* (C1-conjoining form) with the sequence U+09F0 VIRAMA ZWJ C2 and a *ra-phalā* (C2-conjoining from) with C1 ZWJ VIRAMA U+09F0. For those conjuncts marked in red, there aren't any reasonable sequences with ZWJ that will give the desired output.

It is possible that a font-level override might achieve the expected behavior, but it would violate the expected standard character behavior for U+09F0. Furthermore, even a font-level override would require additional joiners to cajole the rendering engines to treat it as a normal sequence and avoid unwanted re-ordering of the syllables (due to *reph* requiring special treatment). Below are sample renderings in MS Word of conjunct */bda/* using the alternate BA (with U+09F0 removed from the *reph* feature & /bda/ ligature defined in *akhn*)

্দৰ ্দৰা ্দেৰ	U+09F0 VIRAMA DA
ৰ্দ ৰ্দা ৰ্দে	U+09F0 ZWJ VIRAMA DA
ৰ্দ ৰ্দা ৰ্দে	U+09F0 VIRAMA ZWJ DA

Also, when the appropriate font is not available it will result in an entirely erroneous and misleading rendering with no way to verify it unless one checks their underlying encoding. (e.g. শব্দ /śabda/ will be rendered as শর্দ /śarda/ in the absence of an appropriate font).

In essence, it would just be a font-level hack.

Encoding a new character

This would be the cleanest option of them all. The previous solution requires a very extensive use of joiners and deviation from the established behavior of U+09F0. It is already an overloaded character that requires special processing by the rendering engines. Overloading the character with joiners to create more specialized forms is counterproductive.

Given that Unicode is trying to move away from specifying complicated joiner-based solutions, a new character would be an appropriate solution after all. Even for Bengali, Khanda TA was atomically encoded to avoid using joiners to render the special form of vowelless TA.

This will, however, result in the encoding of two characters that look similar but behave differently. This won't be the first time such characters are encoded in the UCS. Similar `characters already exist for Arabic where they are only differentiated in isolated/final forms. Bengali already has ambiguous sequences due to the existence of two RAs and the encoding

of the sequence is already decided based on the language of the text. This would be another language-specific addition to the Bengali block.

Bengali	ক্র /rkra/	U+09B0 U+09CD U+0995 U+09CD U+09B0
Assamese	ক্র /rkra/	U+09F0 U+09CD U+0995 U+09CD U+09F0
Bengali	ৰ্ব /barba/	U+09BA[4] U+09B0 U+09CD U+09BA
Assamese	ৰ্ব /rarra/	U+09F0 U+09F0 U+09CD U+09F0

Therefore, it is recommended that a new character BENGALI ALTERNATE BA be encoded with the character properties as shown below.

<div align="center">

ৰ   09BA   Bengali Letter Alternate BA

</div>

```
09BA;BENGALI LETTER ALTERNATE BA;Lo;0;L;;;;;N;;;;;
```

The following change needs to be made to the *IndicSyllabicCategory.txt* file:

Replace:

```
09B6..09B9    ; Consonant # Lo   [4] BENGALI LETTER SHA..BENGALI LETTER HA
```

With:

```
09B6..09BA    ; Consonant # Lo   [5] BENGALI LETTER SHA..BENGALI LETTER ALTERNATE BA
```

As a rarely used letter, we recommend that the letter be excluded from domain names for security purposes.

Collation

Normal Collation

As a rarely used letter it can be the last consonant in the collation following YA.

[…] < YA < ALTERNATE BA < AVAGRAHA < […]

Tailored collation for Sanskrit

For Sanskrit, it should collate as:

[…] < PHA < ALTERNATE BA < BHA < […] < YA < RA < LA < BA < BA WITH LOWER DIAGONAL < […]

---

[4] Proposed codepoint

Attestations

মন্ত্রৱলৈর্ৱসন্তি /mantrabalairvasanti/ [1]
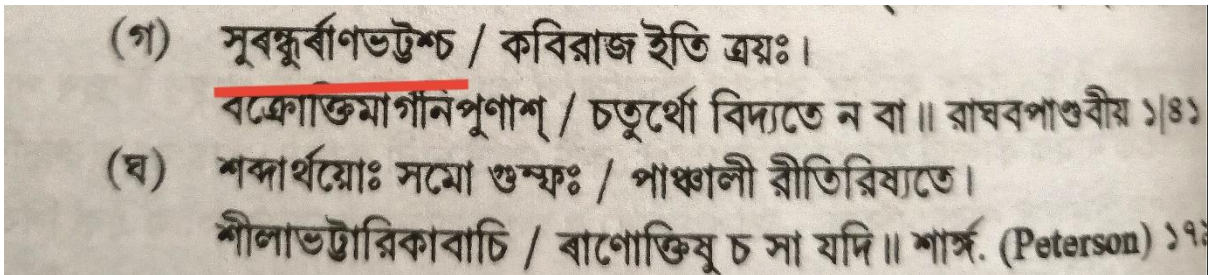

চুম্ষিআইং /cumbiāiṁ/ [2]
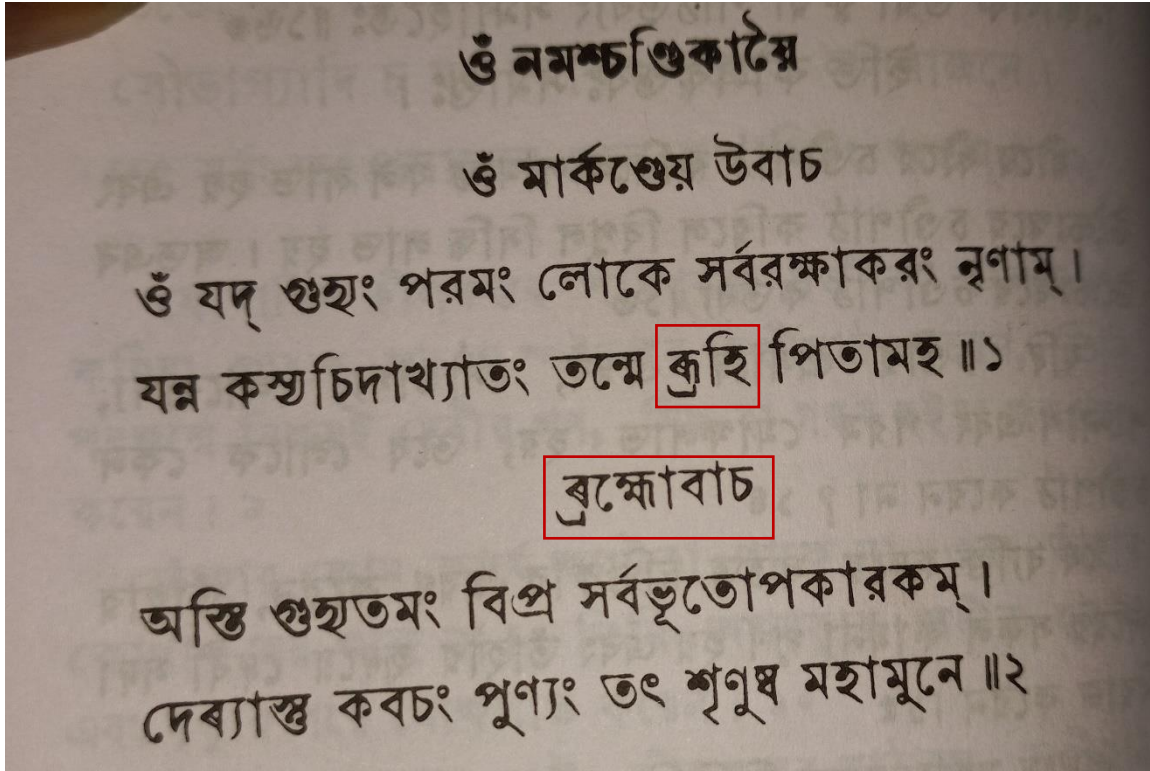

নাট্যবেদাব্ধিমালোড্য / nāṭyavedābdhimāloḍya/ [2]
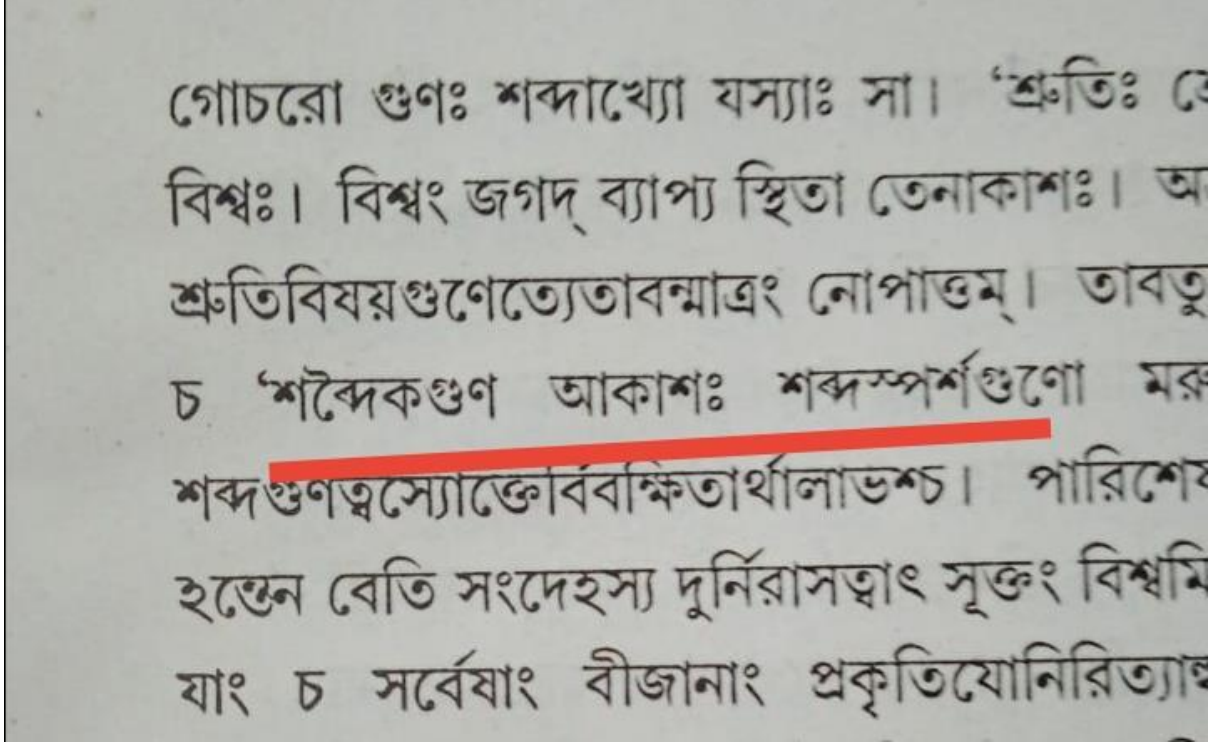
শব্দৈকগুণ আকাশঃ শব্দস্পর্শগুণো /śabdaikaguṇa ākāśaḥ śabdasparśaguṇo / [2]
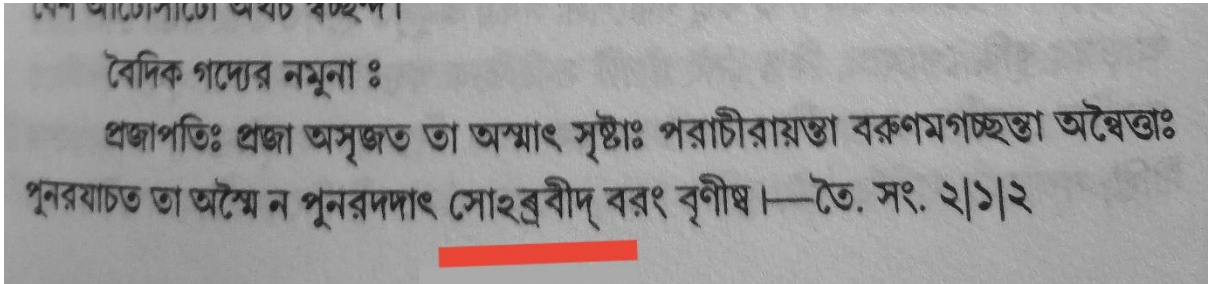
সোঽব্রবীদ্ /so'bravīd/ [3]

সুবন্ধুর্বাণভট্টশ্চ /subandhurbāṇabhaṭṭaśca/ [3]

ব্রূহি /brūhi/

ব্রহ্মোবাচ /brahmovāca/ From [4]

Acknowledgements

References

1. *সেনগুপ্ত, আচার্য্য জ্যোতি (2016). মহাকবিশ্রীহর্ষদেবকৃত রত্নাবলী.* সংস্কৃত বুক ডিপো।

   Senaguptā, Ācāryya Jyoti (2016). Mahākabiśrīharṣadebakṛta Ratnābalī. Saṁskṛta Buka Ḍipo.

2. চক্রবর্তী, সত্যনারায়ণ (2007). *মহাকবি -কালিদাস-প্রণীতম্ অভিজ্ঞান-শকুন্তলম্.* Sanskrit Pustak Bhandar.

   Cakrabartī, Satyanārāẏaṇa (2007). Mahākabi-Kālidāsa-Praṇītam Abhijñāna-Śakuntalam. Sanskrit Pustak Bhandar.

3. বন্দ্যোপাধ্যায়, ধীরেন্দ্রনাথ (2000). সংস্কৃত সাহিত্যের ইতিহাস. পশ্চিমবঙ্গ রাজ্য পুস্তক পর্ষৎ (West Bengal State Book Board).

   Bandyopādhyāẏa, Dhīrendranātha (2000). Saṁskṛta Sāhityera Itihāsa. Paścimabaṅga Rājya Pustaka Parṣat (West Bengal State Book Board).

4. https://github.com/virtualvinodh/aksharamukha/issues/173
5. https://fr.wikipedia.org/wiki/%E0%A6%AD%E0%A6%BC
6. সান্যাল, দুর্গাচন্দ্র (সন ১৩১৬). ভাষা বিজ্ঞান নামক বাঙ্গালা ভাষার ব্যাকরণ. হিতবাদী লাইব্রেরী.

   Sānyāla, Durgācandra (sana 1316). Bhāṣā Bijñāna Nāmaka Bāṅgālā Bhāṣāra Byākaraṇa. Hitabādī Lāibrerī.

7. No author (1997). বাংলা বানানবিধি: পশ্চিমবঙ্গ বাংলা আকাদেমি গৃহীত. পশ্চিমবঙ্গ বাংলা আকাদেমি. কলকাতা

   No author (1997). Bāṁlā bānānabidhi: paścimabaṅga bāṁlā ākādēmi gṛhīta. Paścimabaṅga Bāṁlā Ākādēmi. Kalakātā.

8. চৌধুরী, জামিল (2016). বাংলা একাডেমি আধুনিক বাংলা অভিধান. বাংলা একাডেমি. ঢাকা.

   Choudhury, Jamil (2016). Bāṁlā ekāḍēmi ādhunika bāṁlā abhidhāna (Bangla Academy Modern Bangla Dictionary). Bangla Academy. Dhaka.
   (https://xeroxtree.com/pdf/adhunik_bangla_ovidhan.pdf)

ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646[5]
Please fill all the sections A, B and C below.
Please read Principles and Procedures Document (P & P) from http://std.dkuug.dk/JTC1/SC2/WG2/docs/principles.html
for guidelines and details before filling this form.
Please ensure you are using the latest Form from http://std.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html.
See also http://std.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html **for latest Roadmaps.**

A. Administrative

1. **Title:**	*Proposal to Encode Alternate BA for the Bengali Language*
2. Requester's name:	*Vinodh Rajan & Deepro Chakraborty*
3. Requester type (Member body/Liaison/Individual contribution):	*Individual*
4. Submission date:	*27th July 2022*
5. Requester's reference (if applicable):	
6. Choose one of the following:	
This is a complete proposal:	*Y*
(or) More information will be provided later:	

B. Technical – General

1. Choose one of the following:
   a. This proposal is for a new script (set of characters):
       Proposed name of script:
   b. The proposal is for addition of character(s) to an existing block: *Y*
       Name of the existing block: *Bengali*
2. Number of characters in proposal: *1*
3. Proposed category (select one from below - see section 2.2 of P&P document):

A-Contemporary	B.1-Specialized (small collection)	B.2-Specialized (large collection) *Y*
C-Major extinct	D-Attested extinct	E-Minor extinct
F-Archaic Hieroglyphic or Ideographic		G-Obscure or questionable usage symbols

4. Is a repertoire including character names provided?
       a. If YES, are the names in accordance with the "character naming guidelines"
           in Annex L of P&P document?   *Y*
       b. Are the character shapes attached in a legible form suitable for review?   *Y*
5. Fonts related:
       a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard?
           Vinodh Rajan
       b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.):
           Vinodh Rajan vinodh@virtualvinodh.com
6. References:
       a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided?   *Y*
       b. Are published examples of use (such as samples from newspapers, magazines, or other sources)
       of proposed characters attached?   *Y*
7. Special encoding issues:
       Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)?
           Collation

8. Additional Information:

Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at http://www.unicode.org for such information on other scripts. Also see Unicode Character Database (http://www.unicode.org/reports/tr44/) and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

---

[5]. Form number: N4502-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09, 2003-11, 2005-01, 2005-09, 2005-10, 2007-03, 2008-05, 2009-11, 2011-03, 2012-01)

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before?	*No*
If YES explain	
2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)?	*Yes*
If YES, with whom? *The co-author, Deepro Chakraborty, is a native user and manuscript scholar*	
If YES, available relevant documents:	
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included?	*Y*
Reference: *See Proposal*	
4. The context of use for the proposed characters (type of use; common or rare)	*Rare*
Reference: *See Proposal*	
5. Are the proposed characters in current use by the user community?	Yes
If YES, where? Reference: *Some Sanskrit publishers using the Bengali orthography*	
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP?	
If YES, is a rationale provided?	*Yes*
If YES, reference: *Bengali is in the BMP*	
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)?	*Y*
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence?	*Y*
If YES, is a rationale for its inclusion provided?	*Y*
If YES, reference: *See Proposal*	
9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters?	*N*
If YES, is a rationale for its inclusion provided?	
If YES, reference:	
10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to, or could be confused with, an existing character?	*Y*
If YES, is a rationale for its inclusion provided?	*Y*
If YES, reference: *See Proposal*	
11. Does the proposal include use of combining characters and/or use of composite sequences?	*N*
If YES, is a rationale for such use provided?	
If YES, reference:	
Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?	
If YES, reference:	
12. Does the proposal contain characters with any special properties such as control function or similar semantics?	*N*
If YES, describe in detail (include attachment if necessary)	
13. Does the proposal contain any Ideographic compatibility characters?	*N*
If YES, are the equivalent corresponding unified ideographic characters identified?	
If YES, reference:	