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, দৃব /d̪ɔ/ vs দ্ব /d̪ʼɔ/. 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 ൵ (ଉ 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: ব (ব with an inner diagonal stroke) and ব was relegated to only represent VA.

From [4]:

জ্যায়সী চেৎ কর্মণেতে মতা বুদ্ধিজ্ঞানার্দন।
তৎ কিং কর্মণি ঘোরে মাং নিয়োজ্যসি কেশব

\(^1\) 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/.
It has been claimed such distinction can be traced back to the pre-modern/medieval era.

From [6]:

\[
\text{ba and va}
\]

**Rule: 23:** In the earlier language, both the shape and pronunciation of the antyastha (sic) va were different than the vargiya ba. In Sanskrit texts handwritten by pandits the shape of a vargiya ba is written as \( \text{ব} \). 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 \( \text{ব} \). Afterwards, the letter \( \text{ব} \) should be pronounced like the English letter V and \( \text{ব} \) 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 (ব) vs VA (ব).

This new alternate BA shows analogous behavior to that of \( \text{ব} \) 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)
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.
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.

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

ক্র/ 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 form) 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*)

| দেবদেবদেব | U+09F0 VIRAMA DA |
| ব্দ ব্দা ব্দে | U+09FO ZWJ VIRAMA DA |
| ব্দ ব্দা ব্দে | U+09FO 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 শৰ্ব/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.

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Bengali</td>
<td>ৰ /barba/</td>
<td>U+09BA U+09CD U+0995 U+09CD U+09B0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assamese</td>
<td>ৰ /barba/</td>
<td>U+09F0 U+09CD U+0995 U+09CD U+09F0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bengali</td>
<td>ব্র /barba/</td>
<td>U+09BA U+09B0 U+09CD U+09F0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assamese</td>
<td>ব্র /barba/</td>
<td>U+09F0 U+09CD U+09F0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

ব 09BA Bengali Letter Alternate BA

09BA;BENGALI LETTER ALTERNATE BA;Lo;O;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 < [...]
Attestations

মন্ত্রলৈবসন্তি /mantrabalairvasanti/ [1]

চুম্বিঃ /cumbiāṁ/ [2]

নাট্যবেদাক্ষিণালোভ / nāṭyavedābdhimalodya/ [2]
শব্দকণ্ঠ আকাশঃ শব্দস্পর্শগুণ /śabdaikaguna ākāśaḥ śabdasparśaguno / [2]

সোঘ্বাদী /soˈbraːdi/ [3]

সুবক্ষুর্বার্ণভট্ট /subandhurbānabhatṭaśca/ [3]
Acknowledgements

Thanks to Milind Chakraborty for providing additional attestations for the character and clarifying its usage.
References

1. সেনগুপ্ত, আচার্য জ্যোতি (2016). মহাকবিশ্রীবরং রত্নাবৈবী. সংস্কৃত বুক ডিপো।


5. https://fr.wikipedia.org/wiki/%E0%A6%AD%E0%A6%BC

6. সান্তাল, দুর্গাচন্দ্র (সন ১৩১৬)। ভাষা বিজ্ঞান নামক বাঙ্গালা ভাষার ব্যাকরণ। হিতবাদী লাইব্রেরী।

7. No author (1997). বাংলা বানানবিদ্যা। পশ্চিমবঙ্গ বাংলা অকাদেমি গৃহীত। পশ্চিমবঙ্গ বাংলা অকাদেমি। কলকাতা।

8. চৌধুরী, জামিল (2016)। বাংলা একাডেমি আধুনিক বাংলা অভিধান। বাংলা একাডেমি। ঢাকা।
### PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646

Please fill all the sections A, B and C below.


---

#### 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:**  
   1. This is a complete proposal: **Y**  
   2. (or) More information will be provided later:  

---

#### B. Technical – General

1. **Choose one of the following:**
   1. This proposal is for a new script (set of characters):
   2. The proposal is for addition of character(s) to an existing block: **Y**
   
2. **Number of characters in proposal:** 1  
   
3. **Proposed category (select one from below - see section 2.2 of P&P document):**
   1. A-Contemporary  
   2. B.1-Specialized (small collection)  
   3. B.2-Specialized (large collection) **Y**  
   4. C-Major extinct  
   5. D-Attested extinct  
   6. E-Minor extinct  
   7. F-Archaic Hieroglyphic or Ideographic  
   8. G-Obscure or questionable usage symbols  

4. **Is a repertoire including character names provided?**
   1. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document? **Y**
   2. Are the character shapes attached in a legible form suitable for review? **Y**

5. **Fonts related:**
   1. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard? **Vinodh Rajan**
   2. 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:**
   1. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? **Y**
   2. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? **Y**

7. **Special encoding issues:**
   1. 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](http://www.unicode.org) for such information on other scripts. Also see Unicode Character Database ([http://www.unicode.org/reports/tr44/](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.
### C. Technical - Justification

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