L2/22-268R

# Revised Proposal to Encode Alternate BA for the Bengali Language

Vinodh Rajan vinodh@virtualvinodh.com

Deepro Chakraborty deepro@ualberta.ca

*17-Dec-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.

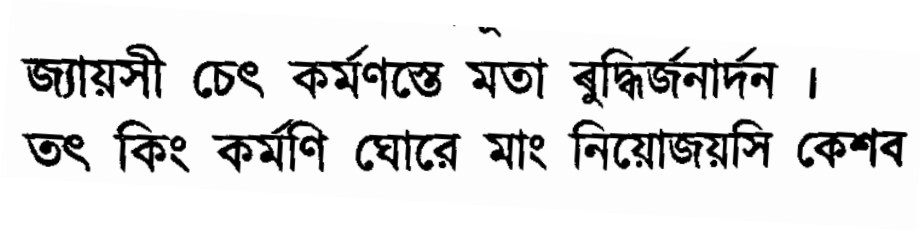
(The Bangla Academy recommends [7] the orthographic differentiation of, at least, the clusters using an explicit virama. For instance, দ্‌ব /d̪bɔ/ vs দ্ব /d̪[ː]ɔ/[2] (< /*d̪ʊɔ/). Seemingly, this convention is not widely followed [4]).

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 (ଵ BA with nukta).

Differentiating VA and **BA in the Bengali orthography**

In a similar vein, some Bengali publications[3], particularly those printing Sanskrit and Pali (see L2/22-278) 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: ৰ (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 /d̪ː ɔ/

[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 ৱুদ্ধির্জনার্দন ।
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]:

ৰ এবং ব।

২৩ সূত্র। আদি ভাষায় অন্ত্যস্থ ৰকারের আকৃতি এবং উচ্চারণ উভয়ই বর্গীয় বকার হইতে বিভিন্ন। পণ্ডিতের হস্ত লিখিত সংস্কৃত পুস্তক সমূহে বর্গীয় ৰকারের ৰ এইরূপ আকৃতি লিখিত হয়। কিন্তু বাঙ্গালা ছাপার বর্ণ মালায় ৰ এবং ব উভয়ই ব সদৃশ লিখিত হয় এবং আকৃতি তুল্যতা হেতু উচ্চারণও তুল্য হইয়া গিয়াছে। এই অনিষ্ট নিবারণ জন্য আমি ৰ কারের সংশোধন করিলাম। অতঃপর ব কার ইংরেজী V নামক বর্ণের ন্যায় এবং ৰ ইংরেজী B নামক বর্ণের ন্যায় উচ্চারণ করা উচিত।

*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 ৰ. 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 ৰ. Afterwards, the letter ৰshould be pronounced like the English letter V and ৰshould 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 | ব্র | র্র |
| Cru | ব্রু | ব্রু | র্রু |
| Cya | ব্য | ব্য | র্য |
| Cda | ব্দ | ব্দ | র্দ |
| Cdha | ব্ধ | ব্ধ | র্ধ |
| mCa | ম্ব | ম্ব | ম্র |
| CCa | ব্ব | ব্ব | র্ব |

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 | ব্র | U+09F0 ZWJ VIRAMA U+09B0 |
| bru | ব্রু | U+09F0 ZWJ VIRAMA U+09B0 VS-U |
| bya | ব্য | U+09F0 ZWJ VIRAMA YA |
| bda | ব্দ | ? |
| bdha | ব্ধ | ? |
| mba | ম্ব | ? |
| bba | ব্ব | ? |

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.

র্ক /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 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*)

| | |
|---|---|
| ্দৰ্ দৰা্ দেৰ | 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 শর্দ/śarda/ in the absence of an appropriate font). In essence, it would just be a font-level hack

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 | ৰৰ/barba/ | U+09FF[4] U+09B0 U+09CD U+09FF |
| Assamese | ৰৰ /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.

---

ৰ 09FF Bengali Letter Alternate BA

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

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

```
09FF        ; Consonant # Lo     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***

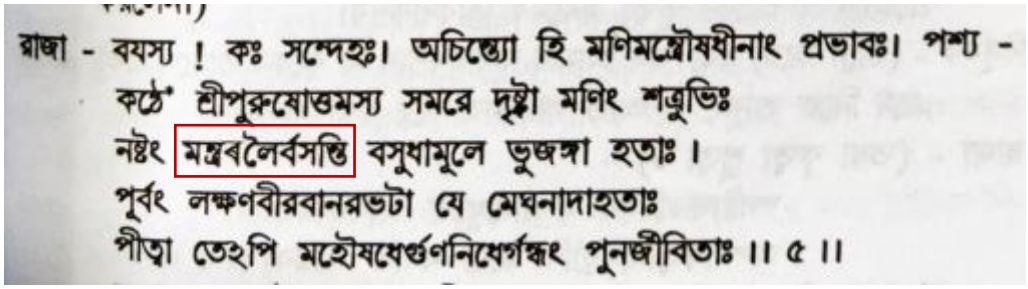
For Sanskrit, it should collate as:

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

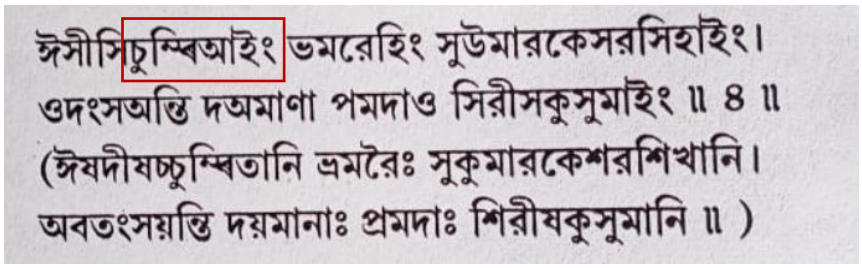
[…]

[4] Proposed codepoint

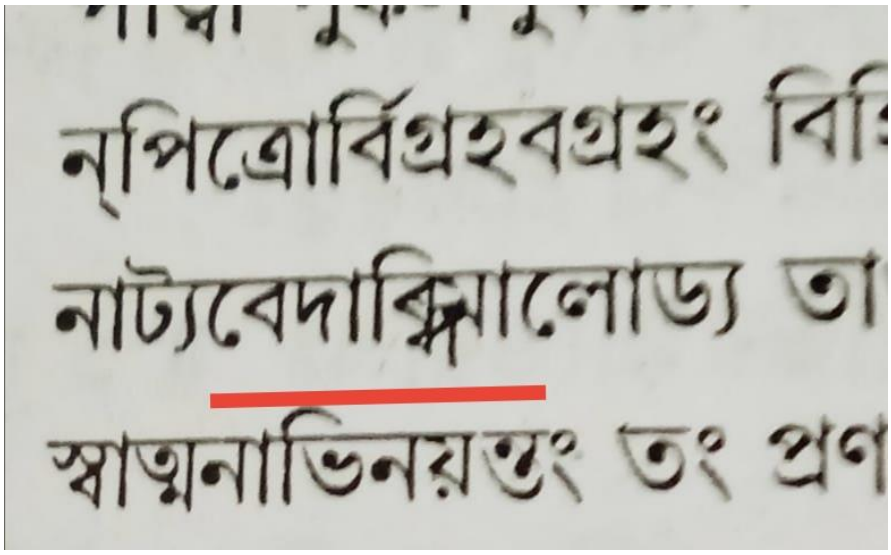
## Attestations


রাজা - বয়স্য ! কঃ সন্দেহঃ। অচিন্ত্যো হি মণিমন্ত্রৌষধীনাং প্রভাবঃ। পশ্য -
কণ্ঠে শ্রীপুরুষোত্তমস্য সমরে দৃষ্ট্বা মণিং শত্রুভিঃ
নষ্টং মন্ত্রবলৈর্বসন্তি বসুধামূলে ভুজঙ্গা হতাঃ।
পূর্বং লক্ষণবীরবানরভটা যে মেঘনাদাহতাঃ
পীত্বা তেঽপি মহৌষধৈর্গুণনিধের্গন্ধং পুনর্জীবিতাঃ ।। ৫ ।।

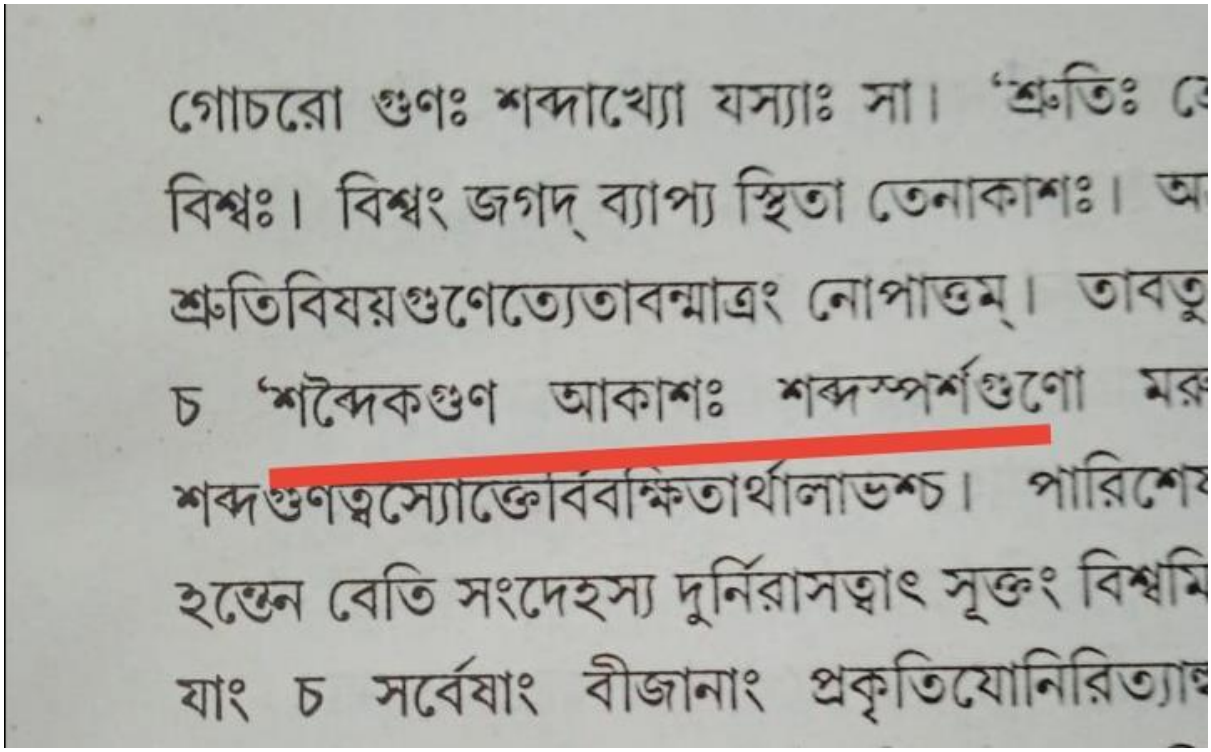
মন্ত্রবলৈর্বসন্তি /mantrabalairvasanti/ [1]


ঈসীসিচুম্বিআইং ভমরেহিং সুউমারকেসরসিহাইং।
ওদংসঅন্তি দঅমাণা পমদাও সিরীসকুসুমাইং ॥ ৪ ॥
(ঈষদীষচ্চুম্বিতানি ভ্রমরৈঃ সুকুমারকেশরশিখানি।
অবতংসয়ন্তি দয়মানাঃ প্রমদাঃ শিরীষকুসুমানি ॥ )

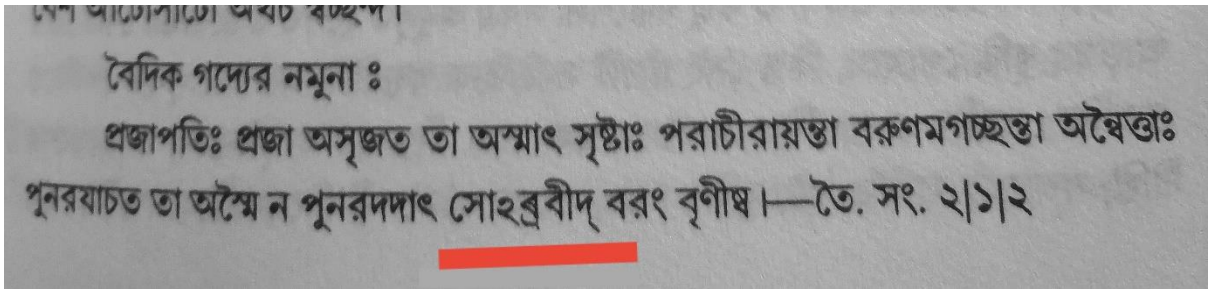
চুম্বিআইং /cumbiāiṁ/ [2]


নৃপিত্রোর্বিগ্রহবগ্রহং বি
নাট্যবেদাব্ধিমালোড্য তা
স্বাত্মনাভিনয়ন্তং তং প্রণ

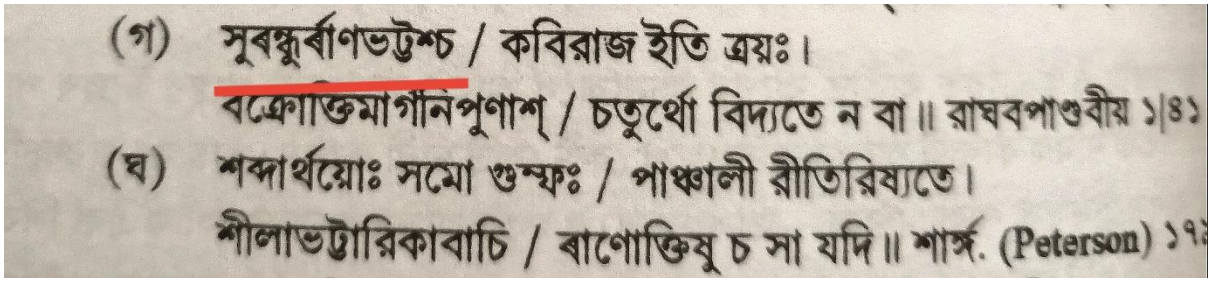
নাট্যবেদাব্ধিমালোড্য / nāṭyavedābdhimāloḍya/ [2]


গোচরো গুণঃ শব্দাখ্যো যস্যাঃ সা। 'শ্রুতিঃ
বিশ্বঃ। বিশ্বং জগদ্ ব্যাপ্য স্থিতা তেনাকাশঃ।
শ্রুতিবিষয়গুণেত্যেতাবন্মাত্রং নোপাত্তম্। তাবতু
চ 'শব্দৈকগুণ আকাশঃ শব্দস্পর্শগুণো ম
শব্দগুণত্বস্যোক্তেরবিবক্ষিতার্থালাভশ্চ। পারিশে
হন্তেন বেতি সংদেহস্য দুর্নিরাসত্বাৎ সূক্তং বিশ্বমি
যাং চ সর্বেষাং বীজানাং প্রকৃতিযোনিরিত্যা

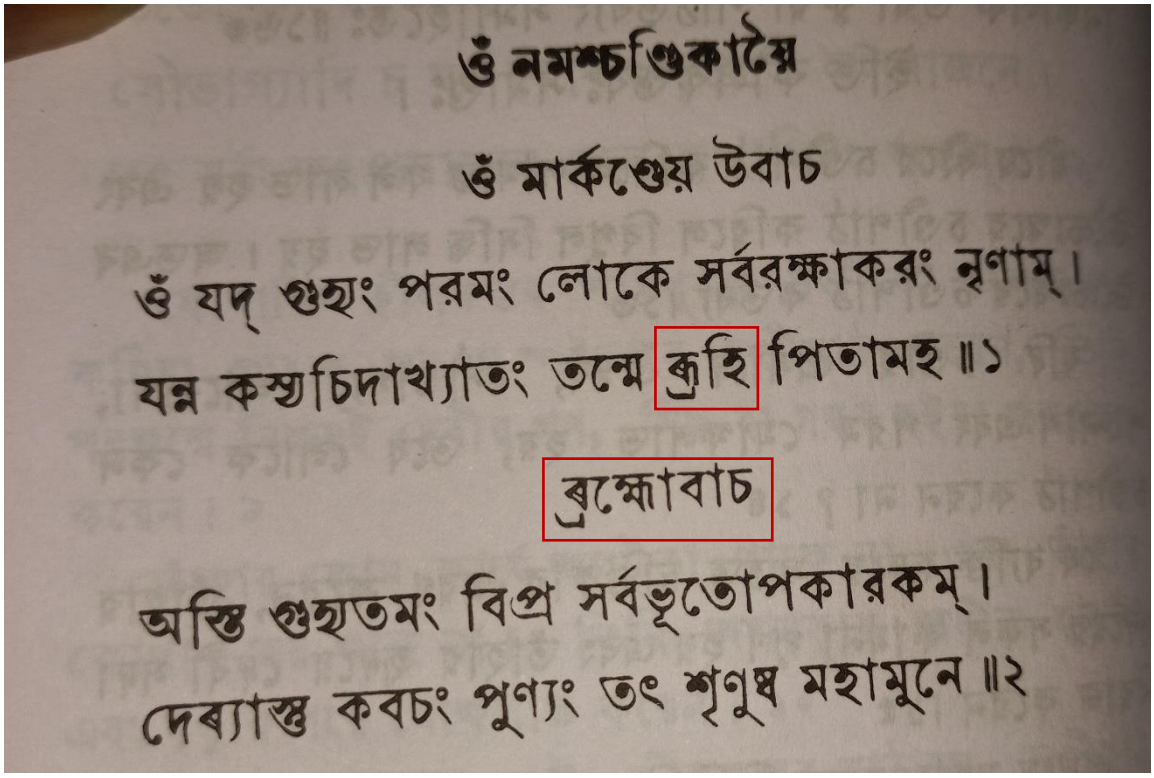
শব্দৈকগুণ আকাশঃ শব্দস্পর্শগুণো /śabdaikaguṇa ākāśaḥ śabdasparśaguṇo / [2]


বৈদিক গদ্যের নমুনা ঃ
প্রজাপতিঃ প্রজা অসৃজত তা অস্মাৎ সৃষ্টাঃ পরাচীরায়ন্তা বরুণমগচ্ছন্তা অন্বৈত্তাঃ
পুনরযাচত তা অস্মৈ ন পুনরদদাৎ সোঽব্রবীদ্ বরং বৃণীষ।—তৈ. সং. ২।১।২

সোঽব্রবীদ্ /so'bravīd/ [3]

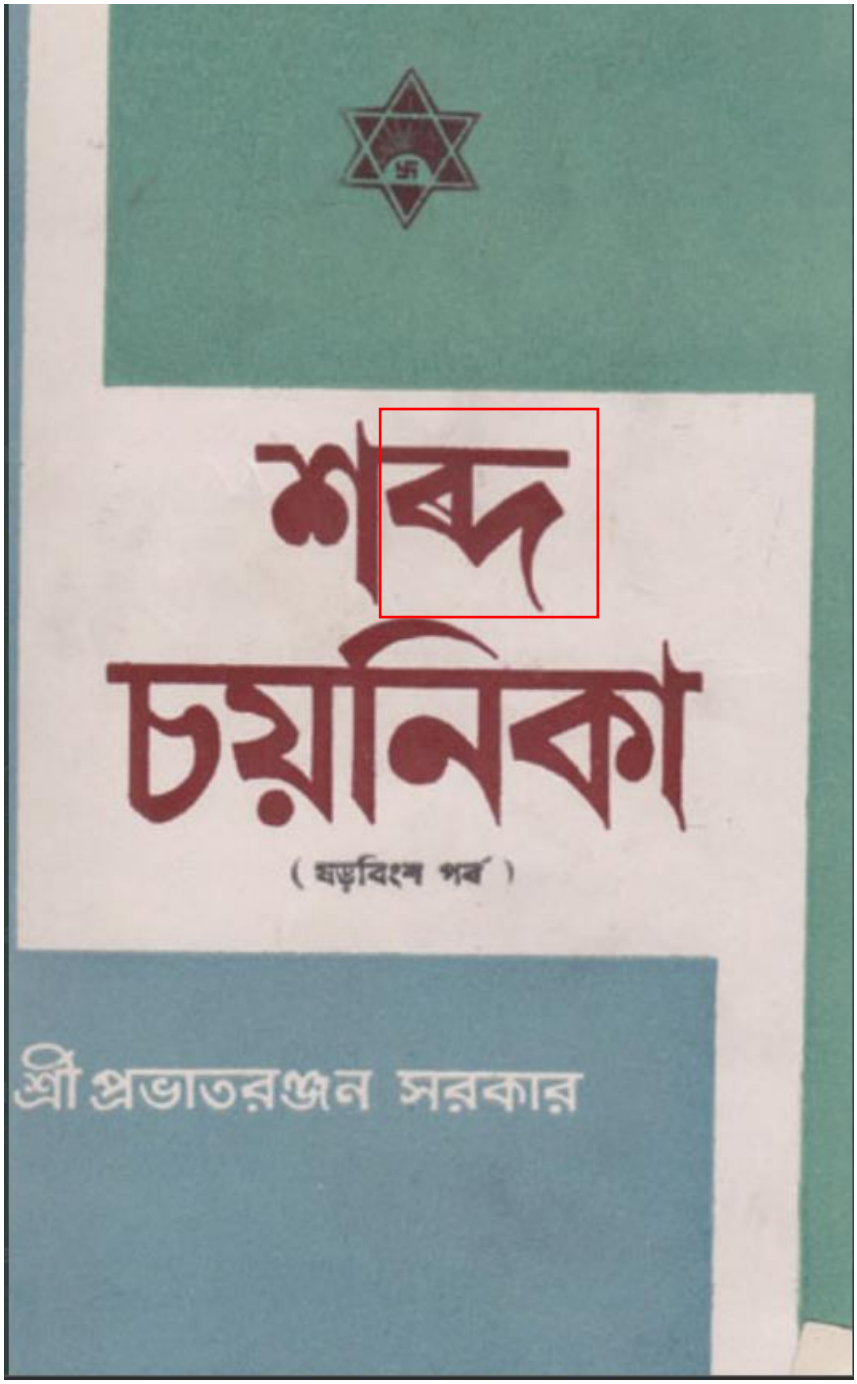

(গ) সুবন্ধুর্বাণভট্টশ্চ / কবিরাজ ইতি ত্রয়ঃ।
বক্রোক্তিমার্গনিপুণাশ্ / চতুর্থো বিদ্যতে ন বা॥ রাঘবপাণ্ডবীয় ১।৪১
(ঘ) শব্দার্থয়োঃ সমো গুম্ফঃ / পাঞ্চালী রীতিরিষ্যতে।
শীলাভট্টারিকাবাচি / বাণোক্তিষু চ সা যদি॥ শার্ঙ্গ. (Peterson) ১৭

সুবন্ধুর্বাণভট্টশ্চ /subandhurbāṇabhaṭṭaśca/ [3]




ব্রূহি /brūhi/

ব্রহ্মোবাচ /brahmovāca/ From [4]




শব্দ চয়নিকা /śabda cayanikā/ [9]

## Acknowledgements

Thanks to Milind Chakraborty for providing additional attestations for the character and clarifying its usage. Publications from Ananda Marga were provided by Biswajit Mandal.

## References


1. সেনগুপ্তা, 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)

9. https://sarkarverse.org/wiki/Shabda_Cayanika

**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:** ***Revised 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: *17th Dec 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:
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? *Y*
   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)? *Y*
   *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? *Y*
   If YES explain *This is a revised version of L2/22-268*
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.)? *Y*
   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? *Y*
   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? *Y*
   If YES, is a rationale provided? *Y*
   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: