

Proposal to Encode 6 Characters in the Brahmi Block

Vinodh Rajan vinodh@virtualvinodh.com
Shriramana Sharma jamadaqni@gmail.com

This document proposes the addition of six characters to support the Old Tamil orthography in the Brahmi block of the UCS.

1. Introduction

The Tamil-Brahmi script is one of the earliest variants of the pan-Indic Brahmic script that was adapted to write Old Tamil in Southern India. As early Brahmi was inadequate to express the phonology of Old Tamil, several additions were made to the script repertoire and the complex conjunct behavior was replaced with a simple visible Virama, effectively resulting in the Tamil-Brahmi variant.

2. Brahmi unification

The Brahmi block was originally conceived to be a unified block for all varieties of Brahmi, unifying multiple orthographies and variations spanning several centuries across multiple geographical areas into one single set. As a result, only a minimal set that deviated from the Brahmic repertoire was encoded separately for Tamil Brahmi: the consonants LLLA, RRA and NNNA. Other variant characters were unified with the existing characters. This is understandable as the prevalent trend a decade ago was that of unification, unlike now, where individual encoding of atomic characters is preferred for minority scripts to facilitate wide-spread font and rendering engine support. This is evidenced by the inclusion of variant characters (e.g. Siddham), atomic encoding of Repha (e.g. Masaram Gondi) and the recommendation to disunify Virama into an explicit Virama and a conjunct-forming control character in historical scripts (e.g. Tulu).

In this background, the present proposal seeks to address the unification of the Tamil-Brahmi Virama with the generic Brahmi Virama and the Tamil-Brahmi LLA with that of the generic Brahmi LLA.

3. Issues with Brahmi unification

Before the existence of Universal Shaping Engine (USE) or the new HarfBuzz (HB), it was a reasonable thing to assume that such unification would have worked as everything was relegated to the font level. There was no wide-spread vendor support for minority scripts and the effects of unifications were still unknown. However, with Windows and Android shipping Brahmi fonts now and the existence of generic shaping engines that handle shaping of minority and historic scripts, it is found that the Tamil-Brahmi orthography is not being properly treated across platforms and

applications. In the next sub-sections, we will see the major issues that arise due to Brahmi unification.

3.1 Rendering/Clustering Issues

The first case in point is the Tamil-Brahmi Virama that was unified with the generic Brahmi Virama. It is graphically and functionally distinct from that of the Generic Brahmi Virama. This is because it usually takes the form of a dot. Further, Tamil Brahmi orthography does not form conjuncts, hence a Virama character should not form them in this context. The generic Brahmi Virama on the other hand is usually shaped as a horizontal bar above and forms conjuncts. This is also reflected in the clustering behavior, where Generic Brahmi and Tamil-Brahmi cluster differently

<i>Syllable</i>	<i>Generic Brahmi</i>	<i>Tamil-Brahmi</i>
k	𑀓	𑀓
kka	𑀓𑀓	𑀓𑀓
putta	𑀧𑀭𑀭𑀭	𑀧𑀭𑀭𑀭

Notice how generic Brahmi forms conjuncts, whereas Tamil-Brahmi does not. In a font that does not support Tamil-Brahmi properly, one must resort to ZWNJ to block conjunct formation. It puts additional constraints on the users and rendering engines. Also, the word /putta/ produces two graphemic clusters in the former, whereas the latter has three.

The unification also deems that the Brahmi Virama act as a vowel reducer to form the Old Tamil vowels short /e/ and short /o/.

(Vowel | Vowel Sign) E + Brahmi Virama = (Vowel | Vowel Sign) Short E

(Vowel | Vowel Sign) O + Brahmi Virama = (Vowel | Vowel Sign) Short O.

This vowel reduction property creates specific problems in rendering engine support. Typically, in many rendering engines, the definition of a valid Indic syllable does not include the combination of a vowel or vowel sign with Virama. This results in such combinations being considered illegal and, consequently, producing incorrect rendering. This is particularly the case with USE and was also with HB. For the latter, a bug had to be filed¹ to fix the issue. But we think that the fix is more or less an ad-hoc measure and a long-term measure is required.

¹ <https://github.com/harfbuzz/harfbuzz/issues/1102>

ஃஃஃஃ

Tamil-Brahmi font rendering of kē kō (1) with USE in MS Word 365. They should have been rendered ஃ ஃ

This also creates problems with clustering as rendering engines incorrectly cluster sequences involving short vowels with the succeeding syllable due to the presence of a Virama.

3.2 Font-dependant graphemic identity

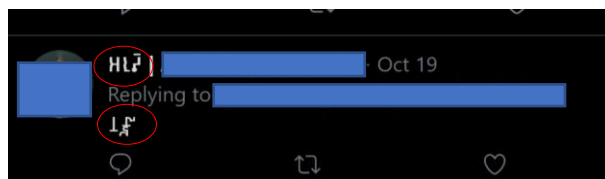
The second case in point is the Tamil-Brahmi LLA. It has a distinct shape compared to the generic Brahmic LLA.

Syllable	Generic Brahmi	Tamil-Brahmi
!	ஃ	ஃ

The generic variant is derived from ஃ /ḍa/, whereas the latter is derived from ஃ /la/. These are clearly different graphemic entities with different origins. The unification results in a complicated font-dependent graphemic identity that is not ideal, expected or even wanted (For instance, see below).

3.3 Insufficient font support

Many users in Social Media (as shown below with sample screenshots rendered in Chrome under Windows 10) have recently begun to post in Tamil-Brahmi. Currently, due to the unification, the default font in Windows namely Segoe Historic UI (and Noto Sans Brahmi in Android) display their text incorrectly. The fonts render the Tamil-Brahmi text using generic behavior.



The name of the above twitter user has been rendered ஃஃஃ and not properly as ஃஃஃ. Ignoring the minor glyphic variations in the first two syllables, notice the use of a completely different character in the last syllable along with the generic Virama. It also has an incorrect rendering of /naṇri/ ஃஃஃ as ஃஃஃ using conjuncts (due to unification).



Here too, the Tamil-Brahmi text forms non-existing Tamil-Brahmi conjuncts in the supposed Tamil-Brahmi text. The commentator noticing this discrepancy, requests the poster to paste the text as an image, which has the correct rendering.

Rationale for disunification

L2/12-226 initially proposed two of the characters for separate encoding by the second author of this document. And this was opposed by the first author in L2/12-233. But since those two documents were written, several things have changed such as the improved vendor support for minority scripts and as a result wide-spread use of Brahmi (and Tamil-Brahmi) across the internet. Back then, one would not have imagined someone posting on Facebook or Twitter in Brahmi, let alone Tamil-Brahmi. So only now we come to see the problems created by the unification. In the light of these issues as discussed earlier, we propose the following complete disunification of these Tamil-Brahmi characters from their corresponding generic Brahmi characters.

Tamil Brahmi short vowels

Firstly, to remove the need for any special rendering behavior in vowel + Virama combinations and thus simplify shaping requirements, we propose to disunify the Tamil short vowels and vowel signs as atomic characters. This would be in line with the character encoding principles of Indic scripts. This would involve four new characters: two independent vowels short /e/ and short /o/ and their corresponding vowel signs.

Tamil Brahmi virama

Secondly, the Tamil-Brahmi Virama itself must be disunified from the generic Brahmi Virama due to its different shape and behavior. This will have the desired effect of improved clustering of Tamil-Brahmi texts.

While it is noted the proposed Brahmi Old Tamil Virama character has the same general dot-like shape as the Brahmi Anusvara, its properties would be different as the proposed character is a Virama character with GC=9 and Indic_Syllabic_Category=Virama against the GC=0 and Indic_Syllabic_Category=Bindu of the Anusvara. Therefore, the Old Tamil Virama is justified to be distinctly encoded.

Tamil Brahmi LLA

Finally, Tamil-Brahmi LLA must be disunified from its generic Brahmi variant based on both the characters' graphemic dissimilarity. This is similar to that of the Bengali block, which has two characters to represent the phoneme /ra/ namely রা for Bengali and ৰ for Assamese. The current unification is technically contradictory to the encoding principles of Unicode, as it was solely based on the phonetic realization of the graphemes and not their graphemic identity.







Benefits of disunification

There will ever be only a handful of fonts for Brahmi and, optimistically speaking, only a limited support for Brahmi itself. As such the only existing Unicode font for Tamil-Brahmi was created by the authors of this document (in collaboration with Udhaya Sankar) in 2012. There haven't been any other fonts that render Tamil-Brahmi text appropriately. Both Google's Noto Sans Brahmi font and Microsoft's Segoe Historic UI, possibly the only two other Unicode fonts for Brahmi, do not properly support the Tamil-Brahmi orthography and are unlikely to within the existing technological limitation due to unification.

By simplifying shaping and removing font-dependent behavior, the disunification and encoding of a minimal set of 6 characters will enable Tamil-Brahmi to gain wide-spread support across platforms. It will allow vendors and rendering engines to support it in a straightforward manner out of the box without any additional effort or special considerations being involved.

Proposed characters

We propose the following characters be added to the Brahmi block of the UCS to properly support the Old Tamil orthography.

11070		BRAHMI SIGN OLD TAMIL VIRAMA
11071		BRAHMI LETTER OLD TAMIL SHORT E
11072		BRAHMI LETTER OLD TAMIL SHORT O
11073		BRAHMI VOWEL SIGN OLD TAMIL SHORT E
11074		BRAHMI VOWEL SIGN OLD TAMIL SHORT O
11075		BRAHMI LETTER OLD TAMIL LLa

Character properties

```

11070;BRAHMI SIGN OLD TAMIL VIRAMA;Mn;9;NSM;;;;N;;;;
11071;BRAHMI LETTER OLD TAMIL SHORT E;Lo;0;L;;;;N;;;;
11072;BRAHMI LETTER OLD TAMIL SHORT O;Lo;0;L;;;;N;;;;
11073;BRAHMI VOWEL SIGN OLD TAMIL SHORT E;Mn;0;NSM;11042
11070;;;;N;;;;;
11074;BRAHMI VOWEL SIGN OLD TAMIL SHORT O;Mn;0;NSM;11044
11070;;;;N;;;;;
11075;BRAHMI LETTER OLD TAMIL LLLA;Lo;0;L;;;;N;;;;

```

Note that ideally speaking the two letter characters short E/O should also receive decompositions to their visual components just like their corresponding vowel signs, but in keeping with existing Indic practice (0B94 Tamil AU being the only aberration), independent vowel letters are not given decompositions. The UTC should confirm that this is what is expected.

Consonants series with Tamil-Brahmi Virama, short E and short O

k n̥ c ñ t̪ ɳ t n p m y r l v ʌ ʀ ɹ ŋ s ʃ dh (consonant + ◌̘)

[illegible]

kě ně cě ně tě ně pě mě yě rě lě vě lě rě ně sě šě dhě (consonant + ㄜ)

ትርጉሙ ርቀው ለገዢው ገንዘብ ያገኙት ገንዘብ ለገዢው ለገዢው

kõ nõ cõ ñõ tõ nõ tõ nõ põ mõ yõ rõ lõ võ lõ lõ rõ nõ sõ sõ dhõ (consonant + ㅏ)

ቶ፡ደጽ ት፡ሮቲ ለ፡፳፻፲፱ ዓ.ም. የፊት ምክር ቤት አመልካች

Indic Syllabic Category

Indic Syllabic Category=Pure Killer

11979 ; Pure Killer # Mn BRAHMI SIGN OLD TAMIL VIRAMA

Indic Syllabic Category=Vowel Independent

```
11071..11072    ; Vowel_Independent # Lo    [2]  BRAHMI LETTER OLD TAMIL SHORT
E..BRAHMI LETTER OLD TAMIL SHORT O
```

```
# Indic_Syllabic_Category=Vowel_Dependent

11073..11074 ; Vowel_Dependent # Mn [2] BRAHMI VOWEL SIGN OLD TAMIL
SHORT E..BRAHMI VOWEL SIGN OLD TAMIL SHORT O

# Indic_Syllabic_Category=Consonant

11074 ; Consonant # Lo BRAHMI LETTER OLD TAMIL LLA
```

Indic Positional Category

```
# Indic_Positional_Category=Top

11070 ; Top # Mn BRAHMI SIGN OLD TAMIL VIRAMA
11073..11074 ; Top # Mn [2] BRAHMI VOWEL SIGN OLD TAMIL SHORT
E..BRAHMI VOWEL SIGN OLD TAMIL SHORT O
```

Collation

The short vowels and vowel signs are sorted before their corresponding long counterparts, Old Tamil Virama after the Brahmi Virama and Old Tamil LLA after Old Tamil LLLA (in accordance with the Tamil collation order)

```
U+11005 < U+11006 < U+11007 < U+11008 < U+11009 < U+1100A < U+1100B < U+1100C <
U+1100D < U+1100E < U+11071 < U+1100F < U+11010 < U+11072 < U+11011 < U+11012 <
U+11013 < U+11014 < U+11015 < U+11016 < U+11017 < U+11018 < U+11019 < U+1101A <
U+1101B < U+1101C < U+1101D < U+1101E < U+1101F < U+11020 < U+11021 < U+11022 <
U+11023 < U+11024 < U+11025 < U+11026 < U+11027 < U+11028 < U+11029 < U+1102A <
U+1102B < U+1102C < U+1102D < U+1102E < U+1102F < U+11030 < U+11031 < U+11032 <
U+11033 < U+11003 < U+11004 < U+11034 < U+11035 < U+11075 < U+11036 < U+11037 <
U+11038 < U+11039 < U+1103A < U+1103B < U+1103C < U+1103D < U+1103E < U+1103F <
U+11040 < U+11041 < U+11073 < U+11042 < U+11043 < U+11074 < U+11044 < U+11045 <
U+11046 < U+1107F < U+11070
```

Existing data

The existing data in Tamil-Brahmi is minimal (mostly in social media and few blogs) and hence any impact to existing data would be minimal. If we are to consider the long-time prospects of archival and stable rendering, such minimal disruption is unavoidable.

ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646²

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 6 Characters in the Brahmi block		
2. Requester's name:	Vinodh Rajan & Shriramana Sharma		
3. Requester type (Member body/Liaison/Individual contribution):	Individual		
4. Submission date:			
5. Requester's reference (if applicable):			
6. Choose one of the following:			
This is a complete proposal:			Yes
(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:			Yes
Name of the existing block:	Brahmi		
2. Number of characters in proposal:			6
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)	
C-Major extinct	C	D-Attested extinct	
F-Archaic Hieroglyphic or Ideographic		E-Minor extinct	
		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?			Yes
b. Are the character shapes attached in a legible form suitable for review?			Yes
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?			Yes
b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached?			Yes
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)?			Yes
	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.

² 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?	Yes
If YES explain	See L2/12-226. This is a new proposal with further disunification.
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 authors themselves belong to the user group
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?	Yes
Reference:	See Proposal
4. The context of use for the proposed characters (type of use; common or rare)	Common
Reference:	See Proposal
5. Are the proposed characters in current use by the user community?	Yes
If YES, where? Reference:	See Proposal
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP?	No
If YES, is a rationale provided?	
If YES, reference:	
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)?	
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence?	No
If YES, is a rationale for its inclusion provided?	
If YES, reference:	
9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters?	Yes
If YES, is a rationale for its inclusion provided?	Yes
If YES, reference:	See Proposal
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?	Yes
If YES, is a rationale for its inclusion provided?	Yes
If YES, reference:	See Proposal
11. Does the proposal include use of combining characters and/or use of composite sequences?	Yes
If YES, is a rationale for such use provided?	Yes
If YES, reference:	See Proposal
Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?	Yes
If YES, reference:	See Proposal
12. Does the proposal contain characters with any special properties such as control function or similar semantics?	No
If YES, describe in detail (include attachment if necessary)	
13. Does the proposal contain any Ideographic compatibility characters?	No
If YES, are the equivalent corresponding unified ideographic characters identified?	
If YES, reference:	