

**Title:** Proposal to Encode Latin Letters for the Transliteration of Indic Vocalic Letters  
**Author:** Anshuman Pandey (pandey@umich.edu)  
**Status:** Individual Contribution  
**Action:** For consideration by UTC and WG2  
**Date:** 2013-10-28

## 1 Introduction

This is a proposal to encode the following eight characters as part of the “Latin Extended-D” block in the Universal Character Set (ISO/IEC 10646):

GLYPH	CODE	CHARACTER NAME
Ł	A7B0	LATIN CAPITAL LETTER L WITH RING BELOW
ł	A7B1	LATIN SMALL LETTER L WITH RING BELOW
Ł̄	A7B2	LATIN CAPITAL LETTER L WITH RING BELOW AND MACRON
ł̄	A7B3	LATIN SMALL LETTER L WITH RING BELOW AND MACRON
Ŕ	A7B4	LATIN CAPITAL LETTER R WITH RING BELOW
ŕ	A7B5	LATIN SMALL LETTER R WITH RING BELOW
Ŕ̄	A7B6	LATIN CAPITAL LETTER R WITH RING BELOW AND MACRON
ŕ̄	A7B7	LATIN SMALL LETTER R WITH RING BELOW AND MACRON

The author understands that the Unicode Technical Committee tends presently towards disapproval of proposals for encoding precomposed characters for Latin, but the matter deserves a formal review as the proposed characters are part of a widely-used, legacy scholarly encoding and an existing ISO standard.

## 2 Background

The eight proposed characters are prescribed by the international standard ISO 15919:2001 “Transliteration of Devanagari and related Indic scripts into Latin characters” for the transliteration of vocalic letters used for writing Sanskrit. These letters — Ł, Ł̄, ł, ł̄ and their upper case forms — are used, respectively, for transliterating ऌ U+090B DEVANAGARI LETTER VOCALIC R, ऎ U+0960 DEVANAGARI LETTER VOCALIC RR, ड U+090C DEVANAGARI LETTER VOCALIC L, ढ U+0961 DEVANAGARI LETTER VOCALIC LL and corresponding characters in other Indic scripts. They represent phonemically and semantically distinct letters of Indic scripts. The under-rings and macrons are inherent aspects of the transliterated forms and the combination is considered an atomic letter in these systems. Various combining marks may be added to these base letters in order to represent nasalization, accents, and other features.

In several Indic transliteration systems, the vocalic letters are represented using a dot below instead of a ring below. These dotted forms are already encoded in the UCS:

Ĺ	U+1E36 LATIN CAPITAL LETTER L WITH DOT BELOW
ĺ	U+1E37 LATIN SMALL LETTER L WITH DOT BELOW
Ľ	U+1E38 LATIN CAPITAL LETTER L WITH DOT BELOW AND MACRON
ľ	U+1E38 LATIN SMALL LETTER L WITH DOT BELOW AND MACRON
Ŕ	U+1E5A LATIN CAPITAL LETTER R WITH DOT BELOW
ŕ	U+1E5B LATIN SMALL LETTER R WITH DOT BELOW
Ř	U+1E5C LATIN CAPITAL LETTER R WITH DOT BELOW AND MACRON
ř	U+1E5D LATIN SMALL LETTER R WITH DOT BELOW AND MACRON

The ring-below letters are preferred because they offer precision in uniquely distinguishing the vocalic letters from retroflex letters, which are represented using dot below. In various conventions, the usage of *r* and *l* for ऋ and ॠ conflicts with the use of these dot-below forms for the retroflex flaps ढ़ *ṛa* and ढ़़ *ṛha* and the retroflex lateral flap ढ़ *ḷa*.

### 3 Analysis

Certainly, the proposed forms may be produced by attaching ◌◌ U+0325 COMBINING RING BELOW to either a Latin small letter ‘r’ or ‘l’ base, and additionally the ◌◌ U+0304 COMBINING MACRON for the long forms. However, there is solid rationale for encoding these letters as precomposed characters, as discussed below.

First, the proposed characters are the only basic vowel letters in ISO 15919 that do not have a one-to-one correspondence to Latin characters already encoded in ISO 10646 (see figure 3).

Second, it is necessary to treat each of the proposed characters as a base letter, which is a single Latin letter or a combination of a Latin letter and one or more Latin diacritics that corresponds to a single Indic letter. In ISO 15919, additional diacritics are used with vowels only for indicating nasalization (◌◌ U+0303 COMBINING TILDE) and Vedic accents (◌◌ U+0304 COMBINING ACUTE ACCENT, etc.) (see figures 4 and 5). At present, ṛ must be represented using a sequence of a letter and two diacritics: <r U+0072 LATIN SMALL LETTER R, ◌◌ U+0325 COMBINING RING BELOW, ◌◌ U+0304 COMBINING MACRON>; and the accented form ṝ is produced by adding a third diacritic to the base sequence. Accented and nasalized vowels are not considered basic letters in Indic scripts, so using marks to represent them in transliteration is appropriate; however, usage of combining marks for producing basic romanized letters does not properly distinguish between marks that are a structural part of a base letter and those that indicate other features. The ideal way to represent ṝ is <ṝ \*LATIN SMALL LETTER R WITH RING BELOW AND MACRON, ◌◌ U+0304 COMBINING ACUTE ACCENT>.

Third, the sequences of combining signs currently needed for representing ṛ, ṝ, ḷ, ḹ are still not rendered properly in several fonts. In such instances the combining ring below is positioned at the horizontal center of the glyph, not below the stem of ‘r’ and ‘l’, as is expected. This is particularly noticeable in italics. The results are unsightly, especially in printed materials. Precomposed letters would help to resolve this issue.

Fourth, the letters ṛ, ṝ, ḷ, ḹ are part of a widely-used 8-bit encoding standard known as Classical Sanskrit eXtended+ (CSX+). The CSX+ encoding has been used by Indologists since 1998 and was the first implementation of the transliteration convention established in ISO 15919. CSX+ is an expansion of the Classical

Sanskrit/Classical Sanskrit eXtended (CS/CSX) encoding that was developed in 1990 at the 8th World Sanskrit Conference in Vienna. It was the base encoding for electronic versions of Indic texts and served as the basis for several Indic Latin fonts used for the publication of Indological articles and books. As shown in figure 1, not only are the four small letters designated as precomposed characters in CSX+, but the accented forms are as well.<sup>1</sup>

Fifth, there have been numerous discussions in the Indological community over the past several years regarding the representation of ळ, ळ̄, ळ̅, ळ̆ as single characters in ISO 10646 on par with the dot-below forms. Until now these requests have not been formally raised.

## 4 Recommendation

These eight characters are used commonly for the transliteration of Indic scripts by general and academic user communities. An encoding for these in the UCS as precomposed characters will assist in retaining compatibility with the long-standing, but now legacy, CSX+ encoding. It will enable their usage as base letters in ISO 10646 as is the intent of ISO 15919. Moreover, these characters will ensure one-to-one synchronization between ISO 15919 and ISO 10646.

## 5 Character Data

Properties for the proposed letters given in the Unicode Character Database format are:

```
A7B0;LATIN CAPITAL LETTER L WITH RING BELOW;Lu;0;L;0052 0325;;;N;;;A7B1;
A7B1;LATIN SMALL LETTER L WITH RING BELOW;Ll;0;L;0072 0325;;;N;;;A7B0;;A7B0
A7B2;LATIN CAPITAL LETTER L WITH RING BELOW AND MACRON;Lu;0;L;0052 0325 0304;;;N;;;A7B3;
A7B3;LATIN SMALL LETTER L WITH RING BELOW AND MACRON;Ll;0;L;0072 0325 0304;;;N;;;A7B2;;A7B2
A7B4;LATIN CAPITAL LETTER R WITH RING BELOW;Lu;0;L;004C 0325;;;N;;;A7B5;
A7B5;LATIN SMALL LETTER R WITH RING BELOW;Ll;0;L;006C 0325;;;N;;;A7B4;;A7B4
A7B6;LATIN CAPITAL LETTER R WITH RING BELOW AND MACRON;Lu;0;L;004C 0325 0304;;;N;;;A7B7;
A7B7;LATIN SMALL LETTER R WITH RING BELOW AND MACRON;Ll;0;L;006C 0325 0304;;;N;;;A7B6;;A7B6
```

## 6 References

ISO 15919:2001. Information and documentation – Transliteration of Devanagari and related Indic scripts into Latin characters. Geneva: International Organization for Standardization (ISO).

Pandey, Anshuman. 1998. “Romanized Indic and L<sup>A</sup>T<sub>E</sub>X”. *TUGboat*, vol. 19, no. 4, pp. 417–418. <http://www.tug.org/TUGboat/tb19-4/tb61pand.pdf>

Stone, Anthony. 2001. “Transliteration of Indic scripts: How to use ISO 15919”. Last revised: March 2012. <http://homepage.ntlworld.com/stone-catend/trind.htm>

<sup>1</sup>The only capital-letter ring-below character in CSX+ is ‘Ṛ’, encoded at the position of ASCII character 187. The decision to include only ‘Ṛ’ was based upon space limitations and the determination that, of the four capital-letter forms, only ‘Ṛ’ occurs frequently in word-initial position and a capital-letter form of ळ may be needed for usage in titles, eg. *Ṛgveda*.

128	Ç	147	ô	165	Ñ	184	ì	203	Æ	222	—	241	ṭ
129	ü	148	ö	166	Ḻ	185	ē	204	kh	223	“	242	Ṭ
130	é	149	ò	167	ṁ	186	ō	205	ḡ	224	ā	243	ḍ
131	â	150	û	168	ă	187	Ŕ	206	ĉ	225	ß	244	Ḍ
132	ä	151	ù	169	ĩ	188	ÿ	207	ř	226	Ā	245	ṇ
133	à	152	æ	170	ũ	189	ú	208	ã	227	ī	246	Ṅ
134	ã	153	Ö	171	ã	190	û	209	ĩ	228	Ī	247	ś
135	ç	154	Ü	172	ĩ	191	ř	210	ũ	229	ū	248	Ś
136	ê	155	ů	173	ṅ	192	õ	211	ẽ	230	Ū	249	ṣ
137	ë	156	ẽ	174	ṙ	193	ṁ	212	õ	231	ṙ	250	Ṣ
138	è	157	ṙ	175	ḷ	194	ṭ	213	ě	232	Ṛ	251	”
139	ï	158	ã	176	ḻ	195	Ē	214	ö	233	Ṛ	252	ṡ
140	î	159	ṙ	177	ř	196	Ō	215	ḷ	234	Ṛ	253	Ṣ
141	ì	160	space	178	ṙ	197	ñ	216	ũ	235	ḻ	254	ḥ
142	Ä	161	í	179	ṙ	198	ř	217	Ĝ	236	ḻ	255	Ḥ
143	Å	162	ó	180	ṁ	199	ṙ	218	Ĉ	237	ḻ		
144	É	163	ú	181	ă	200	Kh	219	ḥ	238	ḻ		
145	æ	164	ñ	182	à	201	k	220	ḥ	239	ṅ		
146	Æ			183	í	202	space	221	—	240	Ñ		

Figure 1: Chart showing the CSX+ encoding (from Pandey 1998: 418). The four small letters proposed for encoding are boxed in red.

Latin	Dev.	Gur.	Guj.	Ben.	Ori.	Tam.	Mal.	Kan.	Tel.	Sin.	Notes
a	अ	ਅ	अ	অ	ଅ	அ	അ	ಅ	ఆ	අ	2.
ā	आ	ਆ	आ	আ	ଆ	ஆ	ആ	ಆ	ఆ	ආ	
æ										ඇ	
ǣ										ඈ	
i	इ	ਇ	ઇ	ই	ଇ	இ	ഇ	ಇ	ఇ	ඉ	
ī	ई	ਈ	ઈ	ঈ	ଈ	ஈ	ஐ	ಈ	ఈ	ඊ	
u	उ	ਉ	ઉ	ঊ	ଉ	உ	உ	ಉ	ఉ	ඊ	
ū	ऊ	ਊ	ઊ	ঊ	ଊ	ஊ	ஊ	ಊ	ఊ	ඊ	
ü							ു				5.
r̥	ऋ		ऋ	ঋ	ଋ		ഠ	ಠ	ఠ	ඊ	
r̄	ॠ		ॠ	ঠ	ॠ		ഠ	ಠ	ఠ	ඊ	
l̥	ऌ		ऌ	ঌ	ॡ		ഡ	ഡ	డ	ඊ	
l̄	ॡ		ॡ	ॡ	ॡ		ഡ	ഡ	డ	ඊ	
e	(ऐ)					எ	എ	ಎ	ఎ	ඊ	
ē	ए	ਏ	એ	এ	ଏ	ஏ	ഈ	ಐ	ఐ	ඊ	
ê	ऎ		ঐ								
ai	ऐ	ਐ	ઐ	ঐ	ঐ	ஐ	ഐ	ಐ	ఐ	ඊ	
o	(ओ)					ஓ	ഓ	ಓ	ఓ	ඊ	
ō	ओ	ਓ	ओ	ও	ଓ	ஓ	ഓ	ಓ	ఓ	ඊ	

Figure 2: Chart showing transliteration of vowels in various Indic scripts in ISO 15919 (source: <http://homepage.ntlworld.com/stone-catend/trimain1.htm>). The vocalic letters are high-lighted.

Latin	lower case	UPPER case	Latin	lower case	UPPER case
a	0061	0041	ṛ	0072 + 0325	0052 + 0325
ā	00E3	00C3	ṝ	0072 + 0325 + 0304	0052 + 0325 + 0304
ā	0101	0100	ḷ	006C + 0325	004C + 0325
ā̃	0101 + 0303	0100 + 0303	ḹ	006C + 0325 + 0304	004C + 0325 + 0304
æ	00E6	00C6	e	0065	0045
æ̃	01E3	01E2	ē	1EBD	1EBC
i	0069	0049	ē	0113	0112
ī	0129	0128	ē̃	0113 + 0303	0112 + 0303
ī	012B	012A	ê	00EA	00CA
ī̃	012B + 0303	012A + 0303	o	006F	004F
u	0075	0055	ō	00F5	00D5
ū	0169	0168	ō̄	014D	014C
ū	016B	016A	ō̃	014D + 0303	014C + 0303
ū̃	016B + 0303	016A + 0303	ô	00F4	00D4
ů	016D	016C			

Figure 3: Chart showing transliteration of basic and nasalized vowels in ISO 15919 (source: <http://homepage.ntlworld.com/stone-catend/triunltv.htm>).

Latin	lower case	UPPER case	Latin	lower case	UPPER case
á	00E1	00C1	í	0072 + 0325 + 0301	0052 + 0325 + 0301
à	00E0	00C0	ì	0072 + 0325 + 0300	0052 + 0325 + 0300
ā	0061 + 0331	0041 + 0331	ī	0072 + 0325 + 0331	0052 + 0325 + 0331
á̇	0101 + 0301	0100 + 0301	í̇	0072 + 0325 + 0304 + 0301	0052 + 0325 + 0304 + 0301
à̇	0101 + 0300	0100 + 0300	ì̇	0072 + 0325 + 0304 + 0300	0052 + 0325 + 0304 + 0300
ā̇	0101 + 0331	0100 + 0331	ī̇	0072 + 0325 + 0304 + 0331	0052 + 0325 + 0304 + 0331
í	00ED	00CD	ı̇	006C + 0325 + 0301	004C + 0325 + 0301
ì	00EC	00CC	ı̈	006C + 0325 + 0300	004C + 0325 + 0300
ī	0069 + 0331	0049 + 0331	ı̈̇	006C + 0325 + 0331	004C + 0325 + 0331
í̇	012B + 0301	012A + 0301	é	00E9	00C9
ì̇	012B + 0300	012A+0300	è	00E8	00C8
ī̇	012B + 0331	012A + 0331	ē	0065 + 0331	0045 + 0331
ú	00FA	00DA	ë	1E17	1E16
ù	00F9	00D9	è̇	1E15	1E14
ū	0075 + 0331	0055 + 0331	ē̇	0113 + 0331	0112 + 0331
ú̇	016B + 0301	016A + 0301	ó	00F3	00D3
ù̇	016B + 0300	016A + 0300	ò	00F2	00D2
ū̇	016B + 0331	016B + 0331	ō	006F + 0331	004F + 0331
			ó̇	1E53	1E52
			ò̇	1E51	1E50
			ō̇	014D + 0331	014C + 0331

Figure 4: Chart showing Unicode characters and sequences used for representing transliteration of Vedic accents in ISO 15919 (source: <http://homepage.ntlworld.com/stone-catend/triunlta.htm>).

1.	Latin basic vowels with diacritics <b>prescribed</b> in ISO 15919	U+
Ä	LATIN CAPITAL LETTER A WITH MACRON AND TILDE	0100 + 0303
ä	LATIN SMALL LETTER A WITH MACRON AND TILDE	0101 + 0303
Ï	LATIN CAPITAL LETTER I WITH MACRON AND TILDE	012A + 0303
ï	LATIN SMALL LETTER I WITH MACRON AND TILDE	012B + 0303
Û	LATIN CAPITAL LETTER U WITH MACRON AND TILDE	016A + 0303
ü	LATIN SMALL LETTER U WITH MACRON AND TILDE	016B + 0303
Ŕ	LATIN CAPITAL LETTER R WITH RING BELOW	0052 + 0325
ŕ	LATIN SMALL LETTER R WITH RING BELOW	0072 + 0325
Ŗ	LATIN CAPITAL LETTER R WITH RING BELOW AND MACRON	0052 + 0325 + 0304
ŗ	LATIN SMALL LETTER R WITH RING BELOW AND MACRON	0072 + 0325 + 0304
Ł	LATIN CAPITAL LETTER L WITH RING BELOW	004C + 0325
ł	LATIN SMALL LETTER L WITH RING BELOW	006C + 0325
Ł̄	LATIN CAPITAL LETTER L WITH RING BELOW AND MACRON	004C + 0325 + 0304
ł̄	LATIN SMALL LETTER L WITH RING BELOW AND MACRON	006C + 0325 + 0304
Ë	LATIN CAPITAL LETTER E WITH MACRON AND TILDE	0112 + 0303
ë	LATIN SMALL LETTER E WITH MACRON AND TILDE	0113 + 0303
Ö	LATIN CAPITAL LETTER O WITH MACRON AND TILDE	014C + 0303
ö	LATIN SMALL LETTER O WITH MACRON AND TILDE	014D + 0303

Figure 5: Chart showing suggested names for transliteration letters in ISO 15919 that are not encoded in Unicode as of version 3.0 (source: <http://homepage.ntlworld.com/stone-catend/trinoun1.gif>).