ISO/IEC JTC1/SC2/WG2 N4494
L2/13-204
2013-10-28

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
L̥	A7B0	LATIN CAPITAL LETTER L WITH RING BELOW
l̥	A7B1	LATIN SMALL LETTER L WITH RING BELOW
L̥̄	A7B2	LATIN CAPITAL LETTER L WITH RING BELOW AND MACRON
l̥̄	A7B3	LATIN SMALL LETTER L WITH RING BELOW AND MACRON
R̥	A7B4	LATIN CAPITAL LETTER R WITH RING BELOW
r̥	A7B5	LATIN SMALL LETTER R WITH RING BELOW
R̥̄	A7B6	LATIN CAPITAL LETTER R WITH RING BELOW AND MACRON
r̥̄	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 — r̥, r̥̄, l̥, l̥̄ and their upper case forms — are used, respectively, for transliterating ऋ U+090B DEVANAGARI LETTER VOCALIC R, ॠ U+0960 DEVANAGARI LETTER VOCALIC RR, ऌ U+090C DEVANAGARI LETTER VOCALIC L, ॡ 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:

Ḷ	U+1E36 LATIN CAPITAL LETTER L WITH DOT BELOW
ḷ	U+1E37 LATIN SMALL LETTER L WITH DOT BELOW
Ḹ	U+1E38 LATIN CAPITAL LETTER L WITH DOT BELOW AND MACRON
ḹ	U+1E38 LATIN SMALL LETTER L WITH DOT BELOW AND MACRON
Ṛ	U+1E5A LATIN CAPITAL LETTER R WITH DOT BELOW
ṛ	U+1E5B LATIN SMALL LETTER R WITH DOT BELOW
Ṝ	U+1E5C LATIN CAPITAL LETTER R WITH DOT BELOW AND MACRON
ṝ	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 *ṛ* and *ḷ* for ऋ and ऌ conflicts with the use of these dot-below forms for the retroflex flaps ड़ *ṛa* and ढ़ *ṛha* and the retroflex lateral flap ळ *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, r̥̄ 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 r̥̄́ 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 r̥̄́ is <r̥̄ *LATIN SMALL LETTER R WITH RING BELOW AND MACRON, ◌́ U+0304 COMBINING ACUTE ACCENT>.

Third, the sequences of combining signs currently needed for representing r̥, r̥̄, l̥, l̥̄ 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 r̥, r̥̄, l̥, l̥̄ 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.[1]

Fifth, there have been numerous discussions in the Indological community over the past several years regarding the representation of r̥, r̥̄, l̥, l̥̄ 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).

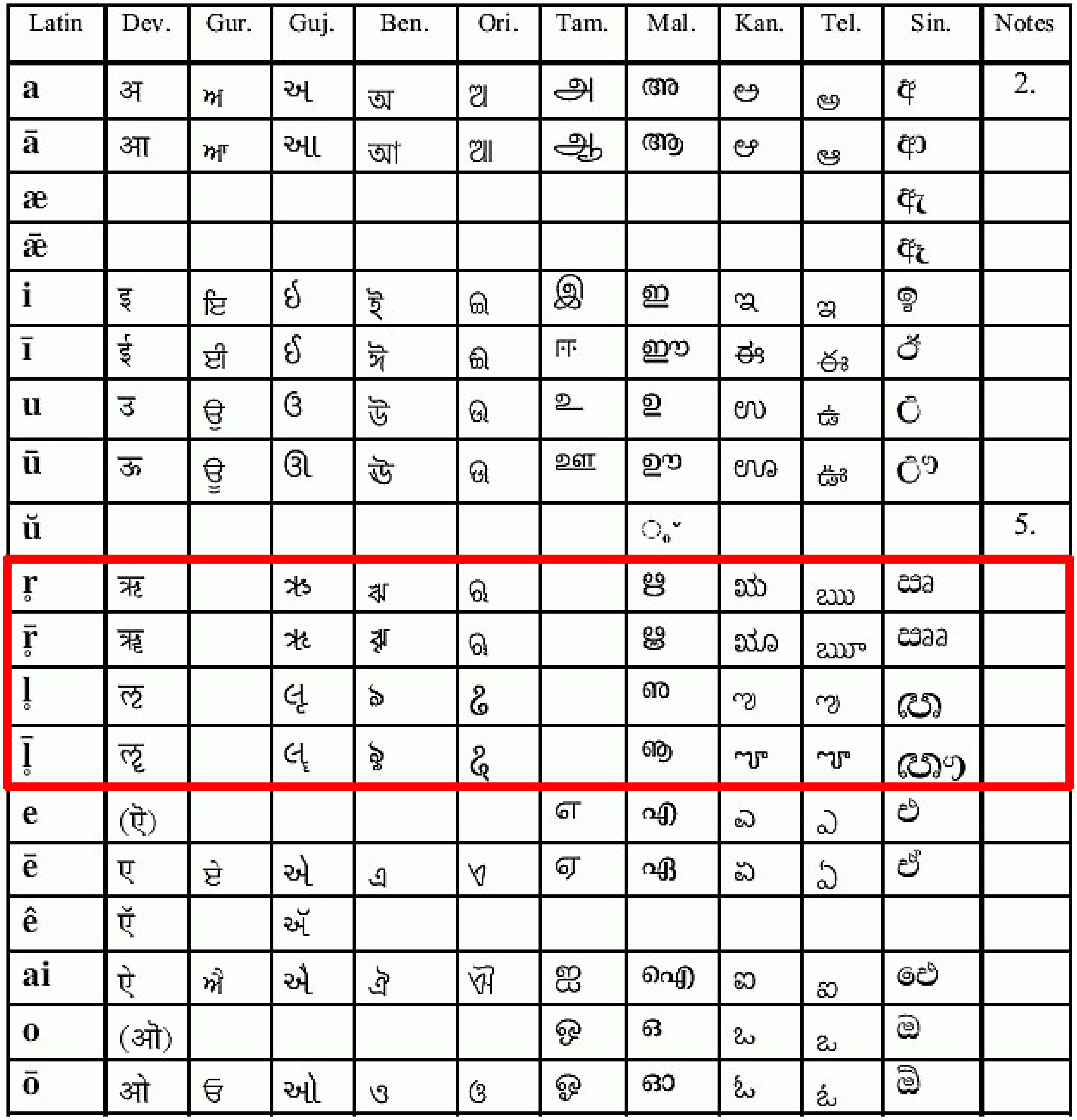
Pandey, Anshuman. 1998. "Romanized Indic and LaTeX". *TUGboat*, vol. 19, no. 4, pp. 417–418. `http://www.tug.org/TUGboat/tb19-4/tb61pand.pdf`

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

[1] The only capital-letter ring-below character in CSX+ is 'R̥', encoded at the position of ASCII character 187. The decision to include only 'R̥' was based upon space limitations and the determination that, of the four capital-letter forms, only 'R̥' occurs frequently in word-initial position and a capital-letter form of r̥ may be needed for usage in titles, eg. *R̥gveda*.

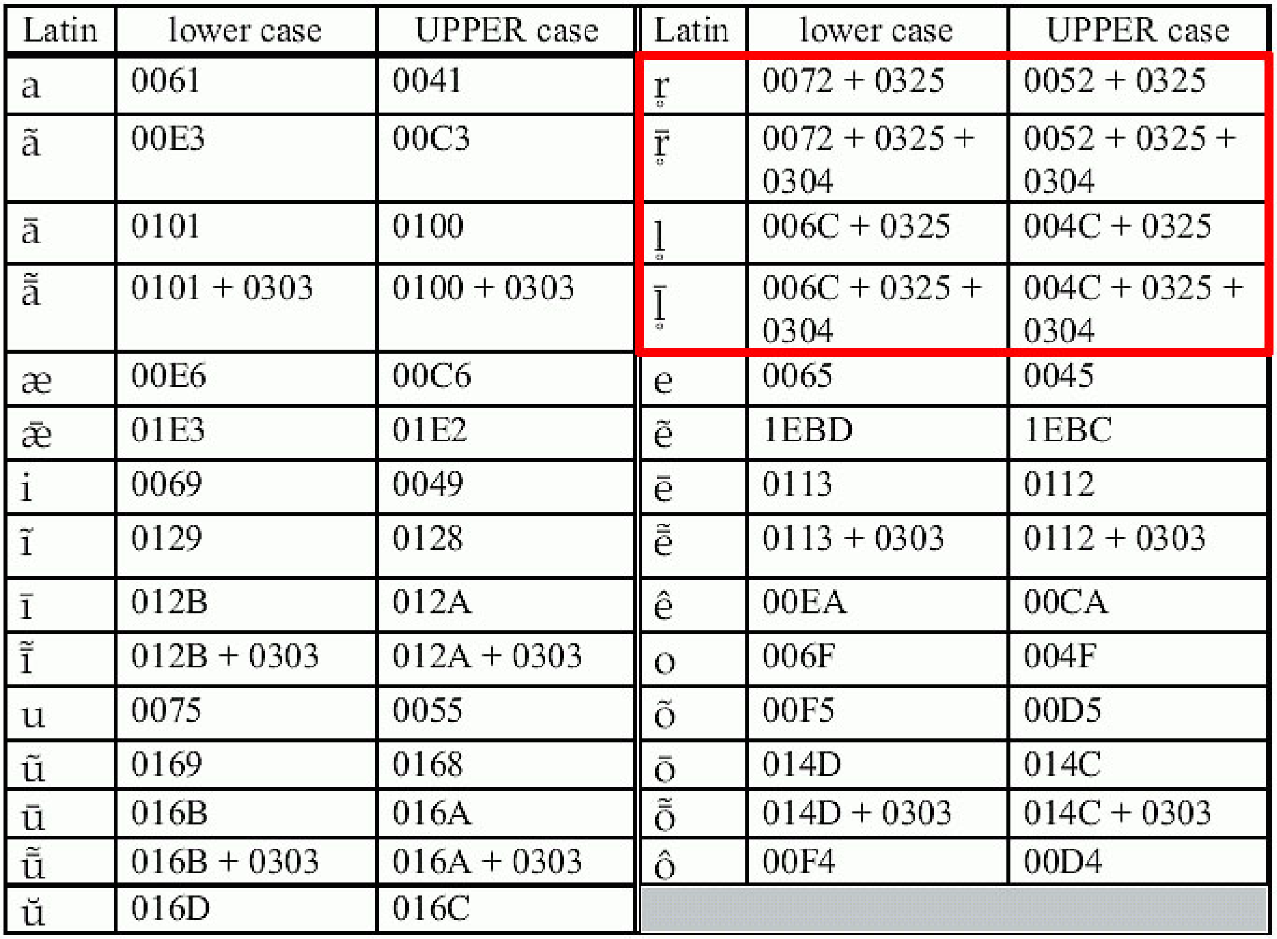
128	Ç	147	ô	165	Ñ	184	ī̀	203	Ǣ	222	—	241	ṭ
129	ü	148	ö	166	l̃	185	ē	204	k̲h̲	223	“	242	Ṭ
130	é	149	ò	167	ṁ	186	ō	205	ġ	224	ā	243	ḍ
131	â	150	û	168	ā̆	187	R̥	206	ĉ	225	ß	244	Ḍ
132	ä	151	ù	169	ī̆	188	ẏ	207	ṝ́	226	Ā	245	ṇ
133	à	152	ǣ	170	ū̆	189	ū́	208	ã	227	ī	246	Ṇ
134	å	153	Ö	171	ā̃	190	ū̀	209	ĩ	228	Ī	247	ś
135	ç	154	Ü	172	ī̃	191	r̆	210	ũ	229	ū	248	Ś
136	ê	155	ŭ	173	ṉ	192	ō̃	211	ẽ	230	Ū	249	ṣ
137	ë	156	ē̃	174	r̥̄	193	m̐	212	õ	231	ṛ	250	Ṣ
138	è	157	r̥	175	l̥	194	ṯ	213	ĕ	232	Ṛ	251	”
139	ï	158	á	176	l̥̄	195	Ē	214	ŏ	233	ṝ	252	ṃ
140	î	159	ṟ	177	ŕ̥	196	Ō	215	ḻ	234	Ṝ	253	Ṃ
141	ì	160	*space*	178	r̥̀	197	n̆	216	ū̃	235	ḷ	254	ḥ
142	Ä	161	í	179	r̥̄́	198	ṛ́	217	Ġ	236	Ḷ	255	Ḥ
143	Å	162	ó	180	m̆	199	ṛ̀	218	Ĉ	237	ḹ		
144	É	163	ú	181	ā́	200	K̲h̲	219	ẖ	238	Ḹ		
145	æ	164	ñ	182	ā̀	201	ḵ	220	ḫ	239	ṅ		
146	Æ			183	ī́	202	*space*	221	–	240	Ṅ		

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



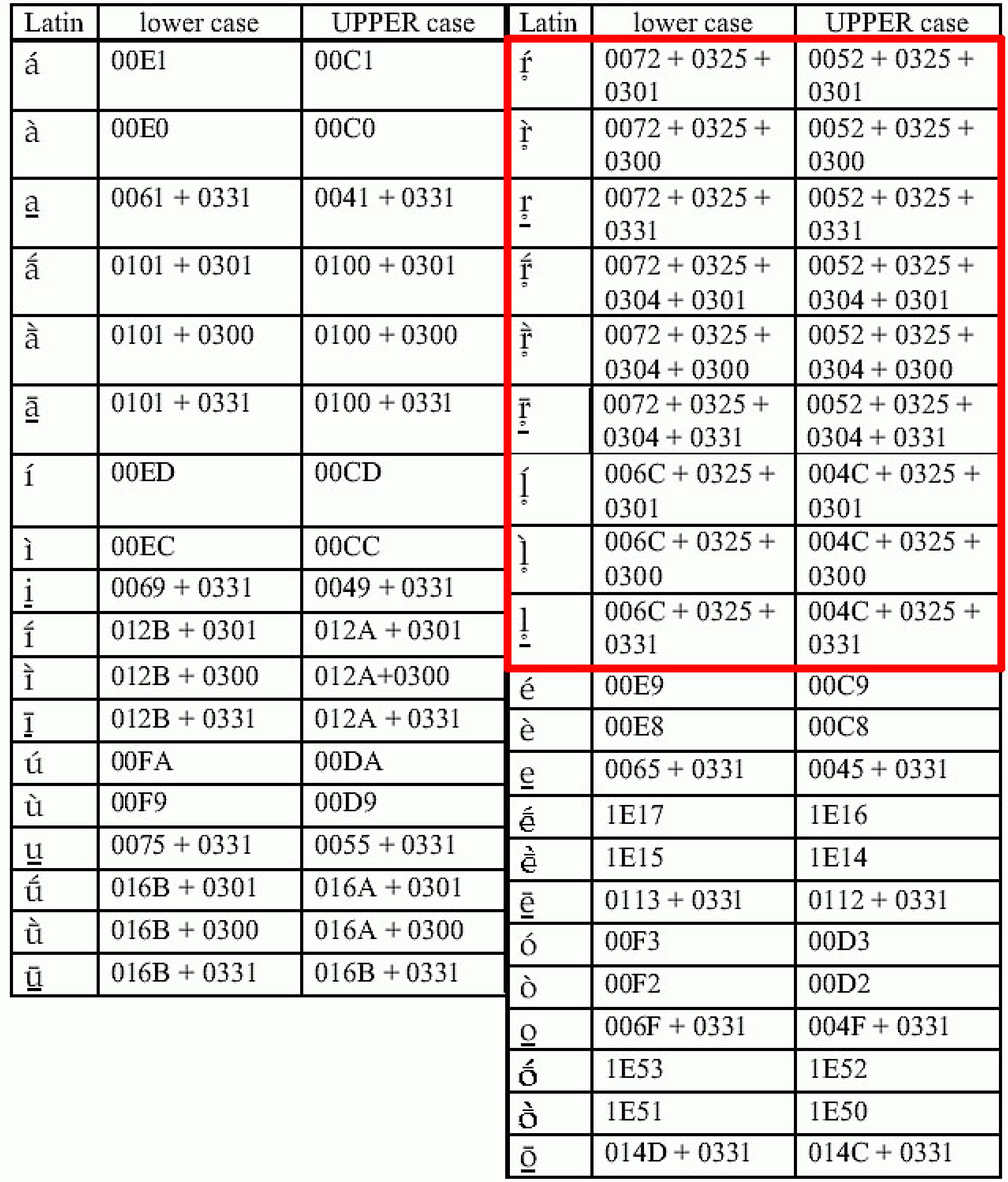
Latin	Dev.	Gur.	Guj.	Ben.	Ori.	Tam.	Mal.	Kan.	Tel.	Sin.	Notes
a	अ	ਅ	અ	অ	ଅ	அ	അ	ಅ	అ	අ	2.
ā	आ	ਆ	આ	আ	ଆ	ஆ	ആ	ಆ	ఆ	ආ	
æ										ඇ	
ǣ										ඈ	
i	इ	ਇ	ઇ	ই	ଇ	இ	ഇ	ಇ	ఇ	ඉ	
ī	ई	ਈ	ઈ	ঈ	ଈ	ஈ	ഈ	ಈ	ఈ	ඊ	
u	उ	ਉ	ઉ	উ	ଉ	உ	ഉ	ಉ	ఉ	උ	
ū	ऊ	ਊ	ઊ	ঊ	ଊ	ஊ	ഊ	ಊ	ఊ	ඌ	
ŭ							◌ു്				5.
r̥	ऋ		ઋ	ঋ	ଋ		ഋ	ಋ	ఋ	ඍ	
r̥̄	ॠ		ૠ	ৠ	ୠ		ൠ	ೠ	ౠ	ඎ	
l̥	ऌ		ઌ	ঌ	ଌ		ഌ	ಌ	ఌ	ඏ	
l̥̄	ॡ		ૡ	ৡ	ୡ		ൡ	ೡ	ౡ	ඐ	
e	(ऎ)					எ	എ	ಎ	ఎ	එ	
ē	ए	ਏ	એ	এ	ଏ	ஏ	ഏ	ಏ	ఏ	ඒ	
ê	ऍ		ઍ								
ai	ऐ	ਐ	ઐ	ঐ	ଐ	ஐ	ഐ	ಐ	ఐ	ඓ	
o	(ऒ)					ஒ	ഒ	ಒ	ఒ	ඔ	
ō	ओ	ਓ	ઓ	ও	ଓ	ஓ	ഓ	ಓ	ఓ	ඕ	

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 highlighted.



Latin	lower case	UPPER case	Latin	lower case	UPPER case
a	0061	0041	r̥	0072 + 0325	0052 + 0325
ã	00E3	00C3	r̥̄	0072 + 0325 + 0304	0052 + 0325 + 0304
ā	0101	0100	l̥	006C + 0325	004C + 0325
ā̃	0101 + 0303	0100 + 0303	l̥̄	006C + 0325 + 0304	004C + 0325 + 0304
æ	00E6	00C6	e	0065	0045
ǣ	01E3	01E2	ẽ	1EBD	1EBC
i	0069	0049	ē	0113	0112
ĩ	0129	0128	ē̃	0113 + 0303	0112 + 0303
ī	012B	012A	ê	00EA	00CA
ī̃	012B + 0303	012A + 0303	o	006F	004F
u	0075	0055	õ	00F5	00D5
ũ	0169	0168	ō	014D	014C
ū	016B	016A	ō̃	014D + 0303	014C + 0303
ū̃	016B + 0303	016A + 0303	ô	00F4	00D4
ŭ	016D	016C			

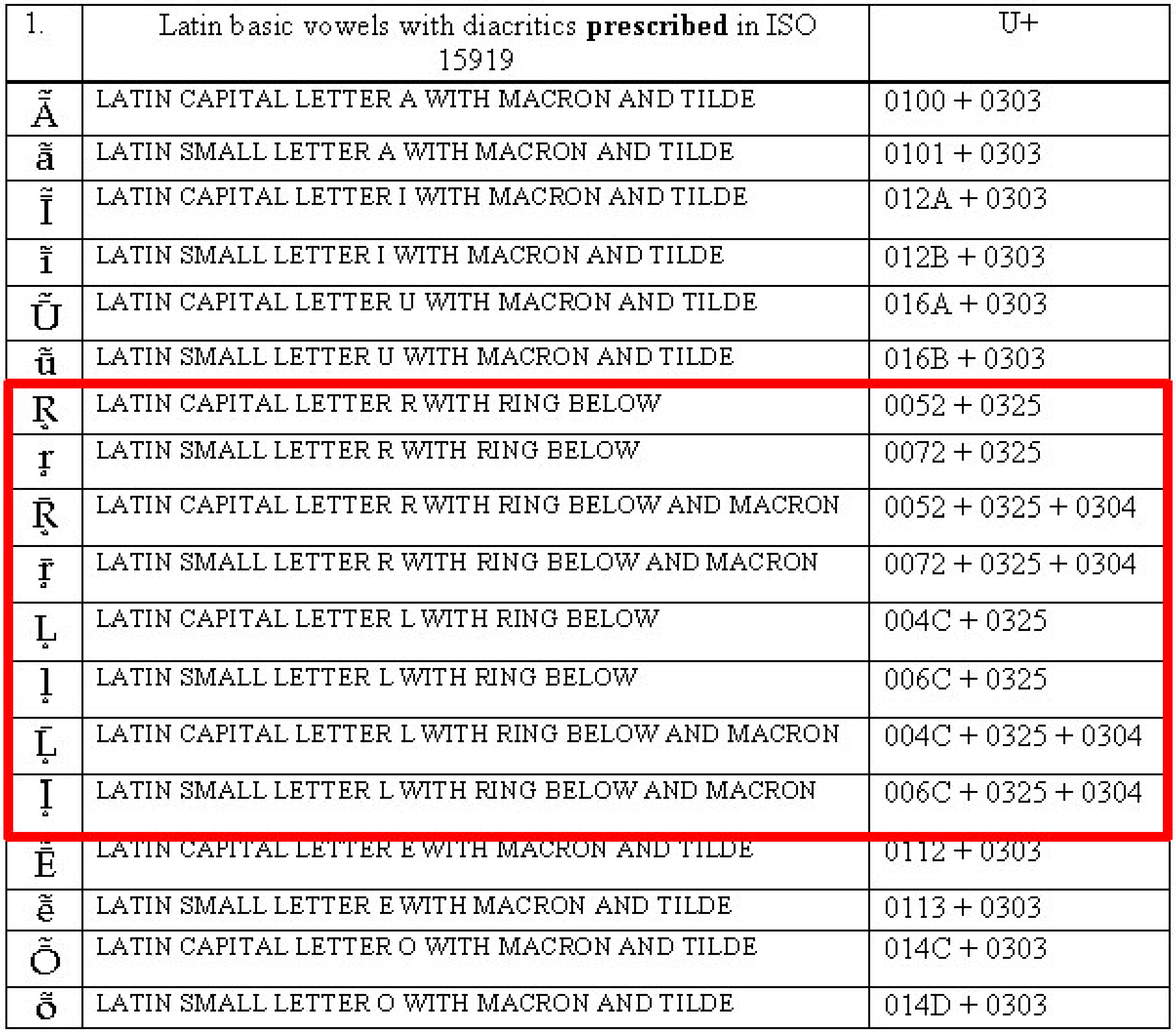
Figure 3: Chart showing transliteration of basic and nasalized vowels in ISO 15919 (source: `http://homepage.ntlworld.com/stone-catend/triunltv.htm`).



Latin	lower case	UPPER case
á	00E1	00C1
à	00E0	00C0
a̱	0061 + 0331	0041 + 0331
ā́	0101 + 0301	0100 + 0301
ā̀	0101 + 0300	0100 + 0300
ā̱	0101 + 0331	0100 + 0331
í	00ED	00CD
ì	00EC	00CC
i̱	0069 + 0331	0049 + 0331
ī́	012B + 0301	012A + 0301
ī̀	012B + 0300	012A+0300
ī̱	012B + 0331	012A + 0331
ú	00FA	00DA
ù	00F9	00D9
u̱	0075 + 0331	0055 + 0331
ū́	016B + 0301	016A + 0301
ū̀	016B + 0300	016A + 0300
ū̱	016B + 0331	016B + 0331

Latin	lower case	UPPER case
ŕ̥	0072 + 0325 + 0301	0052 + 0325 + 0301
r̥̀	0072 + 0325 + 0300	0052 + 0325 + 0300
r̥̱	0072 + 0325 + 0331	0052 + 0325 + 0331
r̥̄́	0072 + 0325 + 0304 + 0301	0052 + 0325 + 0304 + 0301
r̥̄̀	0072 + 0325 + 0304 + 0300	0052 + 0325 + 0304 + 0300
r̥̱̄	0072 + 0325 + 0304 + 0331	0052 + 0325 + 0304 + 0331
ĺ̥	006C + 0325 + 0301	004C + 0325 + 0301
l̥̀	006C + 0325 + 0300	004C + 0325 + 0300
l̥̱	006C + 0325 + 0331	004C + 0325 + 0331
é	00E9	00C9
è	00E8	00C8
e̱	0065 + 0331	0045 + 0331
ḗ	1E17	1E16
ḕ	1E15	1E14
ē̱	0113 + 0331	0112 + 0331
ó	00F3	00D3
ò	00F2	00D2
o̱	006F + 0331	004F + 0331
ṓ	1E53	1E52
ṑ	1E51	1E50
ō̱	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 **prescribed** in ISO 15919	U+
Ā̃	LATIN CAPITAL LETTER A WITH MACRON AND TILDE	0100 + 0303
ā̃	LATIN SMALL LETTER A WITH MACRON AND TILDE	0101 + 0303
Ī̃	LATIN CAPITAL LETTER I WITH MACRON AND TILDE	012A + 0303
ī̃	LATIN SMALL LETTER I WITH MACRON AND TILDE	012B + 0303
Ū̃	LATIN CAPITAL LETTER U WITH MACRON AND TILDE	016A + 0303
ū̃	LATIN SMALL LETTER U WITH MACRON AND TILDE	016B + 0303
R̥	LATIN CAPITAL LETTER R WITH RING BELOW	0052 + 0325
r̥	LATIN SMALL LETTER R WITH RING BELOW	0072 + 0325
R̥̄	LATIN CAPITAL LETTER R WITH RING BELOW AND MACRON	0052 + 0325 + 0304
r̥̄	LATIN SMALL LETTER R WITH RING BELOW AND MACRON	0072 + 0325 + 0304
L̥	LATIN CAPITAL LETTER L WITH RING BELOW	004C + 0325
l̥	LATIN SMALL LETTER L WITH RING BELOW	006C + 0325
L̥̄	LATIN CAPITAL LETTER L WITH RING BELOW AND MACRON	004C + 0325 + 0304
l̥̄	LATIN SMALL LETTER L WITH RING BELOW AND MACRON	006C + 0325 + 0304
Ē̃	LATIN CAPITAL LETTER E WITH MACRON AND TILDE	0112 + 0303
ē̃	LATIN SMALL LETTER E WITH MACRON AND TILDE	0113 + 0303
Ō̃	LATIN CAPITAL LETTER O WITH MACRON AND TILDE	014C + 0303
ō̃	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`).