UTIL: UNIFIED TRANSLITERATION OF INDIC LANGUAGES

:NOTE: Download PDF of this article. You need good Unicode fonts to read the article. I recommend Noto font family from Google. You can `apt-get` it. IPA symbols are within square brackets, like [ʂ] and transliterated symbols are within slashes, like /ṭ/.

UTIL is a romanization scheme for Indic languages. It is designed as pan-Indian transliteration scheme. It covers 20+ languages: Bengali, Dogra, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Lepcha, Limbu, Manipuri (Meitei), Maithili, Malayalam, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Santali, Sindhi, Sinhala, Tamil, Telugu, Urdu and probably many more.

So, why yet another scheme?

1. IAST is insufficient. It serves Sanskrit and Pali but is incomplete for pretty much everything else (e.g. Bengali, Gujarati, etc.).
2. ISO–15919 is also insufficient. It ignores Kashmiri and Sindhi, which are integral Indian languages. Plus, it lacks symbols for newly-assigned Unicode codepoints (e.g. ज़ or ষ). Also कृष्ण /Kṛṣṇa/ is typographically more consistent than /Kr̥ṣṇa/.
3. ALA-LC is designed as a single-language model ignoring the inherent similarity of Brahmic scripts. This leads to inconsistencies. For example, Tamil ழ /l̤/ and Kannada ಱ /l̤/ correspond to same character and sound ("retorflex approximant") and yet have different representations. Conversely, the same symbol /ṣ/ represents Hindi ष [ʂ] and Urdu ص [sˤ] even though they're completely different sounds.
4. Other schemes like Hunterian or Gretil are as bad as the above or even worse sometimes.

So, how is UTIL better?

- Covers the entire character set of ISO15919 plus more (Kashmiri, Sindhi)
- Long vowels always have macron above (ā, ē, ...)
- Aspirated consonants always have 'h' as second letter (kh, gh, ...)
- Minimum number of diacritical marks:
  - Only four diacritics are used: "dot above", "dot below", "macro above", "macron below" (or their combination).
  - Only three diacritics are needed for Sanskrit, instead of IAST's five.

- Prefers precomposed characters in Unicode repertoire, but not required.

VOWELS

Primary vowels and diphthongs:

अ	आ	इ	ई	उ	ऊ	ऋ	ॠ	ऌ	ॡ	ऐ	ए	ऐ	ओ	ओ	औ
a	ā	i	ī	u	ū	ṛ	ṝ	ḷ	ḹ	e	ē	ai	o	ō	au

Additional ones, all have a dot below:

अं	आं	ऍ	ॲ	ऑ	अु	अू
ạ	ạ̄	ẹ	ọ	ọ̄	ụ	ụ̄

CONSONANTS

Consonants with their Sanskrit names:

	Plosives				Nasal	Implosives	Fricatives		Vibrants		Approximants	
	स्पर्श				नासिक	विस्पर्श	ऊष्मन्		द्रव		अन्तस्थ	
कण्ठ्य	क	ख	ग	घ	ङ	ग़	ख़	ग़			ह	
Velar	k	kh	g	gh	ṅ	g̱	k̲h̲	ġ			h	
तालव्य	च	छ	ज	झ	ञ	ज़	श	झ			य	य़
Palatal	c	ch	j	jh	ñ	j̱	ṡ	zh			y	ẏ
मूर्धन्य	ट	ठ	ड	ढ	ण		ष		ड़	ढ़	ळ	ऴ
Retroflex	ṭ	ṭh	ḍ	ḍh	ṇ		ṣ		ṛ	ṛh	l̤	l̤̇
दन्त्य	त	थ	द	ध	न		स़				ल	
Dental	t	th	d	dh	n		s̱				l	
वर्त्स्य	च़	छ़			ऩ	ड़	स	ज़	र	ऱ		
Alveolar	ċ	ċh			n̤	d̤	s	z	r	r̤		
ओष्ठ्य	प	फ	ब	भ	म	ब़	फ़				व	व़
Labial	p	ph	b	bh	m	ḇ	f				v	w

Affricate glide ষ (‘JJYA’) is transcribed /j̄/.

OTHER SYMBOLS

Anusvāra: ṃ	Anunāsika: ँ	Avagraha: ’
Visarga: ḥ	Jihvāmūlīya: x̣	Upadhmānīya: ẋ

Vedic Udātta: ˊ	Svarita (independent): ˋ	Anudātta: ˍ
Arabic hamza ʾ: ’	Arabic ain ʿ: ‘	
Rising tone: ˊ	Falling tone: ˋ	

Udātta and svarita use *combining* grave and acute accent respectively. Whereas hamza and ain use *non-combining* modifier letters U+02BC and U+02BD respectively. Tone modifiers are used in Maithili, Dogra and other Pahari languages.

GENERAL NOTES

- Anunāsika is denoted by a combining candrabindu. Note the difference between हंस (swan) /haṃsa/ and हँस (laugh) /hǎs/. Diacritic only on second letter in a digraph. Example: हैं /haĭ/
- A colon is used to denote vowel hiatus or resolve ambiguity. Example: बई /ba:i/ (not /bai/)

SCRIPT NOTES

- ॠ /ṝ/, ऌ /ḷ/ and ॡ /ḹ/ are used only in Sanskrit.
- ऐ /e/ = short ए in Southern scripts (ఎ, ఎ, ಎ, എ)
- ओ /o/ = short ओ in Southern scripts (ఒ, ఒ, ಒ, ഒ)
- ऍ /ẹ/ = Gujarati ઍ, Sinhala ඇ, pronounced [æ] as in “bat”
- ऑ /ọ/ = Gujarati ઑ, pronounced [ɔ:] as in “ball”
- ड़ /ṙa/ = Bengali ড়, Punjabi ੜ, Oriya ଡ଼ (RRA, “retroflex flap”)
- ढ़ /ṙha/ = Bengali ঢ়, Oriya ଢ଼ (RHA, “aspirated retroflex flap”)
- ळ /l̤a/ = used in Marathi, Tamil ள, Malayalam ള, Kannada ಳ, Telugu ళ (LLA, “retroflex lateral approximant”)
- ऴ /l̤̇a/ = Tamil ழ, Malayalam ഴ, Kannada ೞ, Telugu ౞ (LLLA, “retroflex approximant” = zha)
- न /n̤a/ = Tamil ன , Kannada ನ, Malayalam ന (NNNA, “alveolar n”)
- ऱ /r̤a/ = Tamil ற, Malayalam റ, Kannada ಱ, Telugu ఱ (RRA, “alveolar r”)
- र् /r̤/ = repha in Marathi
- य़ /ẏa/ = য in Bengali and Oriya, while য /y/ = য
- व़ /wa/ = Urdu ۏ, Assamese ৱ, Oriya ୱ
- ख़ /k̤ha/ = used in Urdu, Punjabi ਖ
- ग़ /ġa/ = used in Urdu, Punjabi ਗ਼
- च़ /ċa/ = used in Kashmiri, Telugu ఀ
- ज़ /za/ = Urdu ز, Gurmukhi ਜ਼, Bengali জ়, Kannada ಜ಼, Telugu జ਼ [d͡z]
- ज़, झ़ /zh/ = Urdu ژ, Gujarati ઝ, Avestan uses ज़
- Kashmiri vowels ạ, ā , ọ, ō , ụ, ū are pronounced [ə], [əː], [ɔ], [ɔː], [ɨ], [ɨː].

Another vowel form, ऑ , is sometimes used for [ɔ] (and [ɔː] is skipped altogether). These symbols are taken from ALA-LC as it is and follows Wikipedia. Kashmiri consonants: च़ [t͡s], छ़ [t͡sʰ] and ज़ [z]

- Sindhi implosives: ग़ /g̱/, ज़ /j̱/, ड़ /ḏ/ and ब़ /ḇ/
- Sinhalese nasals: ඦ /ṉ̱ga/, ඤ /ṉ̱ja/, ඬ /ṉ̱da/, ඳ /ṉ̱ḍa/ and ඹ /ṁ̱ba/
- Sinhala long vowel ෳ and Devanagari vowel sign candra long E (U+0955), used in Avestan, are transliterated /ē̤/
- The six Malayalam chillu characters represent dead consonants (without implicit vowel). As such, the are simply transliterated without adding an a next to the consonant. Hence, ൿ, ൽ, ൾ, ൻ, ൺ and ർ are respectively transliterated as /k-/, /l-/, /ll-/, /n-/, /nn-/ and /rr-/.

URDU

Perso-Arabic characters are chosen in a non-conflicting way with the Brahmic scripts. Urdu introduces six sounds [f, z, ʒ, q, x, ɣ] on top of Hindi (see Hindustani phonology). Note that [f, z, x, ɣ] are fricatives, just like ष [ʂ], स [s], ह [h]. Excluding these IPA signs, the ones in the below table are indicative only.

Urdu	ق	ح	خ	ء	ع	غ	ط	ظ	ز	ذ	ض	ص	ث	ش	ژ	ف	و
UTIL	q	ḥ	ḳẖ	’	‘	ġ	ṭ	ẓ	z	ʐ	ż	s̄	ʂ	ṡ	zh	f	w
IPA	[q]	[ɦ]	[x]	[ʔ]	[ʕ]	[ɣ]	[tˤ]	[zˤ]	[z]	[ð]	[dˤ]	[sˤ]	[θ]	[ʃ]	[ʒ]	[f]	[w]
Devanagari	क़	ह़	ख़	ʔ	ʕ	ग़	त़	ज़	ज़	ज़	ज़	स़	स़	श	झ	फ़	व़

INPUT METHODS FOR IME:S

Of course, a transliteration scheme is not fruitful if it cannot be entered into a computer, for which Input Method Editors (IMEs) are used. This can be thought of as an ASCII transliteration of UTIL.
