

## 12. Encoding of Indic Script : ISCII & Unicode

ASCII is a well known encoding scheme for (Latin script that is used for) English. In order to represent Indic languages with their own scripts different from Latin script, efforts were made at national level during 1960s and 1970s. During 1980s, ISO made provision for evolving multi-octet code for simultaneous representation of world-languages in addition to English. This article discusses intricacies of ISCII (Indian Script Standard Code for Indic Scripts) and their representation in Unicode. Unlike Latin, Indic scripts are non-linear, and hence rendering mechanism becomes important. Indic Scripts are Brahmi Script based and phonetic in nature. There is Alphabetic similarity among the Indic scripts. To illustrate, we cite examples in Devanagari script. A character represents a phoneme. The grapheme (the shape of character) may vary from one language to another.

### History of Indic Scripts:

All Indic scripts derived from Brahmi Script. The Indian scripts may be further divided into two major linguistic groups, Indo-European languages in the north and Dravidian languages in the south of India. The Indo-European group languages are Devanagari, Gurumukhi, Gujarati, Bengali, and Oriya. The Dravidian languages include Telugu, Tamil, Kannada and Malayalam.

### Sounds of Indic languages:

Sounds in Indic languages are divided into four groups Vowels, Consonants, Nasals and Conjuncts. Vowels are pure sounds, Consonants are combination of one sound and a vowel, Nasals are nasal sounds along with vowels & consonants. The conjuncts are combination sounds of two or more sounds.

Indian languages are read as they are written because Script used by Indic languages is a Syllabic alphabet representation. The Syllabic alphabet is basically divided into Swar (Vowel) and Vyanjan (Consonant). Swars may be short Swars or long Swars. The Vyanjans are classified into vargs (Tables).

The uniqueness of Indic language is that each phonemes has been assigned a letter and this the reason the alphabet tends to be quite large. The table on next page shows a superset of Indic consonant sounds.

### Some important facts about Indian Languages:

- ♦ In Indian languages, retroflex variants of a basic sound are found,
- ♦ All the plosive sounds have an aspirated and un-aspirated version,
- ♦ In Indian languages the recognise combinations of one or more sound (character) is considered to be a single unit for sorting purposes.

### Characteristics of Indic Scripts:

India is a multi-lingual country with 22 constitutional languages and 11 different scripts. Twenty two constitutional Indian Languages are mentioned as follows with their scripts within parentheses: Hindi (*Devanagari*), Konkani (*Devanagari*), Marathi (*Devanagari*), Nepali (*Devanagari*), Sanskrit (*Devanagari*), Bodo (*Devanagari*), Dogri (*Devanagari*), Mathili (*Devanagari*), Sindhi (*Devanagari/Urdu*), Kashmiri (*Devanagari/Urdu*), Santhali (*Santhali*), Assamese (*Bangali*), Manipuri (*Bangali*), Bangla (*Bangali*), Oriya (*Oriya*), Gujarati (*Gujarati*), Punjabi (*Gurumukhi*), Telugu (*Telugu*), Kannada (*Kannada*), Tamil (*Tamil*), Malayalam (*Malayalam*) and Urdu (*Urdu*). These are in vogue in different states.

There are less than 5 percent people in India who can work in English. In spite of the plurality of Indian languages and scripts, their script grammar and language grammars are quite similar, and they have 40 to 80 percent vocabularies in common.

Indic Scripts are very complex in nature however they have a common phonetic nature. All major Indic scripts run left to right, although some vowel sign (matra) may appear to the left or right, above or below the base consonant, and sometimes surrounds the base consonant on more than one side appear to the left of their base character for display. In a number of scripts, characters commonly have a headstroke (know as shirorekha) and a high baseline (know as kana). Such characters fix at top from the line when written. The various characteristics of Indic scripts are described below.

### Manner of articulation

		Voiceless Unaspirated	Voiceless Aspirated	Voiced Unaspir.	Voiced aspirated	Nasals
Place of articulation	Velars	क (ka)	ख (kha)	ग (ga)	घ (gha)	ङ (Nha)
	Palatals	च (cha)	छ (chha)	ज (ja)	झ (jha)	ञ (nha)
	Retroflexes	ट (Ta)	ठ (Tha)	ड (Da)	ढ (Dha)	ण (Na)
	Dentals	त (ta)	थ (tha)	द (da)	ध (dha)	न (na)
	Labials	प (pa)	फ (pha)	ब (ba)	भ (bha)	म (ma)
	Semi vowels		य (ya)			र (va)
	Liquid			र (ra)	ल (la)	
	Sibilants		श (sha)	ष (Sa)	स (sa)	
	Glottal		क्ष (kSa)	ह (Ha)	त्र (tra)	ज्ञ (jnha)

There are vowels and consonants in each Indian script. The vowels for Devanagari script are as given below:

#### (a) Vowels:

Vowels are independent letters in Indic scripts but these Independent vowels always come as first Akshar or after a preceding vowel with no intervening consonant. The following table illustrates their correspondant the most common independent vowels and vowel signs in Devanagari:

अ (A), आ (AA), इ (I), ई (II), उ (U), ऊ (UU), ऋ (VOCALIC R), ॠ (VOCALIC L),  
 ऐ (CANDRA E), ऐ (SHORT E), ऐ (E), ए (AI), ओ (CANDRA O), ओ (SHORT O), ओ (O),  
 औ (AU), ऋ (VOCALIC RR), ॠ (VOCALIC LL)

Unicode name	A	AA	I	II	U	UU	vR	E	EE	AI	O	OO	AU
Independent vowel	अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ	ऐ	ओ	औ	औ
Vowel sign	-	ा	ि	ी	ु	ू	ृ	े	ै	ँ	ो	ौ	ँ

Difference in Devanagari between की [kee] and कई[kai].

की follows the base consonant with a vowel sign and कई with an independent vowel. In the second case only, the independent vowel sound is retained immediately after the initial [k], and the sequence is pronounced with two distinct vowel sounds.

#### (b) Consonants:

The word consonant is used to refer to letters of an alphabet that denote a consonant sound. In Indic scripts a consonant character carries an inherent vowel with it. In Indic scripts consonant characters represent a consonant

+ vowel syllable. Each consonant is followed by inherent vowel sound and inherent vowel can vary in pronunciation from script to script.

**For example:** Character क in Hindi (Devanagari script) is pronounced [ka] rather than just [k]. The [a] sound is the inherent vowel, and is usually transcribed as a .

At places where the inherent vowel sound is not required, the inherent vowel sound is killed by use of *Halant*. *Halant* is a small diacritic mark, attached with the consonant to kill its inherent vowel sound. It is called a Virama in the Unicode Standard.

### Examples:

Devanagari, क [k] (क [ka])

Gujarati, ક [k] (ક [ka])

Tamil, க [k] (க [ka])

Telugu, క [k] (క [ka])

In these examples consonants pronounced without the inherent vowel.

Some important facts about inherent vowels:

- ♦ Inherent vowel is not always pronounced.
- ♦ Inherent vowels are needed for round trip script conversion (transliteration).
- ♦ Inherent vowels are not trivial case. They are not contained in every consonant cluster even when that cluster does not contain a visual dependent vowel.
- ♦ Inherent vowels are useful for preserving phonetic value in dictionaries or text-to-speech applications.

### (c) Vowel signs:

The sound of a consonant is modified with the use of Vowel signs. Vowel signs are better known as *matra* in Hindi. This vowel sign overrides the inherent vowel with a different sound. In Indic scripts vowel signs are all combining characters. A consonant can only support one vowel sign at a time. A vowel sign may appear to the left or right, above or below the base consonant, and sometimes surrounds the base consonant on more than one side.

### Examples:

Use of vowel signs with the क consonant in Hindi, and the resultant sounds:

की [ki] के [ke] कू [ku]

Vowel signs may also appear to the left of the base consonant they are related to. For example:

Devanagari, क + ि → कि [ki]

Combination of base character and vowel sign produces a change in the basic shape of either base character or vowel sign or both.

### (d) Consonant clusters:

A consonant cluster is a group or sequence of consonants that appear together in a syllable without a vowel between them. There are two main approaches to indicate that the inherent vowels have disappeared, either (a) change the shape of the consonants or merge them together in some way (a *conjunct* form), or (b) use a special diacritic to indicate the absence of intervening vowels.

**For example,** स् [s] + म [ma] - स्म [sma].

A cluster is not limited to two consonants in all case, eg.

र् [r] + ग् [g] + घ [gha] -> र्घ [gargh]

### (e) Vowel signs used with consonant clusters:

The placement of the vowel sign after consonant cluster may need attention. For example in Devanagari, where a vowel sign that is normally placed to the left of a character is pronounced immediately after the cluster, in परिधान [paridhan], the ि is pronounced after the र.

### (f) Nasalisation and alternative nasal letter representations:

Nasalization is the production of a sound while the velum is lowered, so that air escapes partially or wholly through the nose during the production of the sound. The diacritics of nasalization vary from script to script. The following are a few examples of usage.

- ♦ In Devanagari, nasalisation of vowel sounds is indicated using the *candrabindu* or *anusvara* diacritics, eg. कई [kai], हैं [han].

The anusvara is written above the headstroke, at the right-hand end of the preceding character in following example:

बिंदी = बिन्दी [bindi], खंभा = खम्भा [khambha]

#### (g) Numbers:

Different scripts have different number shapes. While some scripts, such as Modern Tamil uses European numerals.

The following table shows the number symbols:

<b>European</b>	0 1 2 3 4 5 6 7 8 9
<b>Devanagari</b>	० १ २ ३ ४ ५ ६ ७ ८ ९
<b>Bengali</b>	০ ১ ২ ৩ ৪ ৫ ৬ ৭ ৮ ৯
<b>Gurmukhi</b>	੦ ੧ ੨ ੩ ੪ ੫ ੬ ੭ ੮ ੯
<b>Gujarati</b>	૦ ૧ ૨ ૩ ૪ ૫ ૬ ૭ ૮ ૯
<b>Tamil</b>	௦ ௧ ௨ ௩ ௪ ௫ ௬ ௭ ௮ ௯ (no zero)
<b>Telugu</b>	౦ ౧ ౨ ౩ ౪ ౫ ౬ ౭ ౮ ౯
<b>Kannada</b>	೦ ೧ ೨ ೩ ೪ ೫ ೬ ೭ ೮ ೯
<b>Malayalam</b>	൦ ൧ ൨ ൩ ൪ ൫ ൬ ൭ ൮ ൯

Shapes of Tamil numbers are not based on the decimal system that's why there is no zero.

#### (h) Text units & punctuation:

Words are separated with spaces. In most cases modern text uses western punctuation as (, , , etc), but some scripts use traditional punctuation also for example, the DANDA I in Devanagari and DOUBLE DANDA II in Telugu in place of (.) in English to end the sentence.

Computer is basically a number crunching machine. Information is stored in flip-flops as 0/1. Human communication necessitates a language which has alphabets. Alphabet representation is comprehensible to human being, but quite incomprehensible by computer in its original form. Alphabet needs numeric representation of fixed length for the purpose of information processing and exchange. This representation is called coding scheme. For English

alphabet 7-bit code ASCII (American Standard Code for Information Interchange) was evolved that may represent 128 symbols including alphabet, numerals, punctuation marks, special characters and control characters. Later-on 8-bit byte became the minimal unit for information storage with 256 available code-points.

#### ISCII:

In India since 1970s, different committees of the Department of Official Languages and the Department of Electronics (DOE) have been developing different character encodings schemes, which would cater to all the Indian scripts. In July 1983, the DOE announced the 7-bit ISCII-83 code, which complied with the ISO 8-bit recommendations. There was a revision of the ISCII code by the DOE in 1988. Bureau of Standards (BIS) adopted ISCII-88 with minor revision in 1991 and remains the prevalent national standard today [ISCII-91 or IS13194:1991].

ISCII (Indian Script Code for Information Interchange) is a fixed-length 8-bit encoding. The lower 128(0-127) code points are plain ASCII and the upper 95(160-255) code points are ISCII-specific, which is used for all Indian Script based on Brahmi script. This makes it possible to use an Indian Script along with latin script in an 8-bit environment. This facilitates 8-bit bi-lingual representation with Indic Script selection code.

The ISCII code contains the basic alphabet required by the Indian Scripts. All composite characters are formed by combining these basic characters. The alphabet in each script may vary but they all share a common phonetic structure. The differences between scripts are mainly in their written forms. ISCII encoding is completely independent from the glyphs used for display, making it possible for a script to be displayed in a variety of styles. So, ISCII has an advantage that immediate transliteration between different Indian Script is possible by just changing the script display modes.

Unicode is based on ISCII-1988 and incorporates minor revisions of ISCII-1991, thus conversion between one to another is possible without loss of information.



### Properties of ISCII: -

1. ISCII specifies a 7-bit table, which can be used in ISO-10646 compatible environments. It allows English and Indian script alphabet to be used simultaneously.
2. The ISCII characters, within a word, are kept in the same order, as they would get pronounced. Since there are variations in ordering of a few consonants between different Indian scripts, it is not possible to achieve perfect sorting in all Indian scripts. Special routines would be required when some characters like Nukta need to be ignored for the purpose of sorting. For most purposes, however, the direct sorting achieved through the ISCII code should be sufficient.
3. Dictionary sorting sequences even within a language vary. Hence soft-sequence for sort may be used.
4. By using only the basic characters in ISCII, there is only one unique way of typing a word. A word

in an Indian script can be displayed in a variety of styles depending on the conjunct repertoire used. ISCII codes however allow a complete delinking of the codes from the displayed fonts.

5. The INSCRIPT keyboard overlay has one-to-one correspondence with the ISCII code. This way, typing of word does not depend upon its displayed form.
6. The ISCII codes are rendered on the display device according to the display composition methodology of the selected script. Transliteration to another script can thus be obtained by merely redisplaying the same text in a different script.

### ISCII and ASCII Co-Existence:

ISCII is a 8-bit standard in which lower 128 characters (0-127) conform to ASCII standard. The higher 128 characters (128-255) cater to superset alphabet that can represent any one of 9 Indian scripts on switching the display mode.

### ISCII Code Table

Hex		0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
Hex Dec		0	16	32	48	64	80	96	112	128	144	160	176	192	208	224	240
0	0	NUL	DEL	SP	0	@	P	'	p				ओ	ढ	1	े	०
1	1	SOH	DC1	!	1	A	Q	a	q			ँ	औ	ण	2	े	१
2	2	STX	DC2	“	2	B	R	b	r			ं	ऑ	त	3	ै	२
3	3	ETX	DC3	#	3	C	S	c	s			ः	क	थ	4	ँ	३
4	4	EOT	DC4	\$	4	D	T	d	t			अ	ख	द	5	ो	४
5	5	ENQ	NAK	%	5	E	U	e	u			आ	ग	घ	6	ो	५
6	6	ACK	SYN	&	6	F	V	f	v			इ	घ	न	7	ौ	६
7	7	BEL	ETB	'	7	G	W	g	w			ई	ड	न	8	ॉ	७
8	8	BS	CAN	(	8	H	X	h	x			उ	च	प	9	्	८
9	9	HT	EM	)	9	I	Y	i	y			ऊ	छ	फ		र्	९
A	10	LF	SUB	*	:	J	Z	j	z			ऋ	ज	ब	ा	।	
B	11	VT	ESC	+	;	K	[	k	{			ऐ	झ	भ	ि		
C	12	FF	FS	'	<	L	\	l				ए	ञ	म	ी		
D	13	CR	GS	-	=	M	]	m	}			ऐ	ट	य	ु		
E	14	SO	RS	.	>	N	^	n	~			ँ	ठ	य	ू		
F	15	SI	US	/	?	O	_	o	DEL			ओ	ड	र	ृ		

### Examples:

1. Symbol	ISCII Representation (Hex)
A	41
a	61
/	2F
2	32
क	B3
ि	DB
अ	A4
2	F3

2. Ram in ISCII will be stored as 52 61 6D and राम as CF DA CC.

### Unicode:

Two individual and independent project ISO 10646 project of ISO (International Organization for Standardization), and Unicode project by manufacturers of multi-lingual software were launched in late 80s in order to create a single unified character set. But from 1991, they united in favor of world need and published their own tables. But they made them compatible and agreed to perform any future extension with coordination. Unicode 1.1 corresponded to ISO 10646-1:1993, Unicode 3.0 corresponded to ISO 10646-1:2000, Unicode 3.2 added ISO 10646-2:2001, and Unicode 4.0 corresponds to ISO 10646:2003. All Unicode versions since 2.0 are compatible, only new characters will be added; no existing characters will be removed or renamed in the future.

Computers store everything by assigning a number to it because it can store two bit 0 & 1. Unicode provides a unique number for every character, no matter what the platform, program or language.

Unicode is the universal character encoding, maintained by the Unicode Consortium. Unicode provides a unique number to every character of all the scripts of world. This encoding standard provides the basis for processing, storage and interchange of text data in any language in all modern software and information technology protocols.

Unicode covers all the modern and ancient characters, technical symbols, punctuations, and many other

characters used in writing text for all the writing systems of the world.

Unicode enables a single software product or a single website to be targeted across multiple platforms, languages and countries without re-engineering. It allows data to be transported through many different systems without corruption.

### Scripts covered:

Unicode covers almost all scripts (writing systems) in current use today.

### That includes:

- Arabic
- Armenian
- Bengali
- Braille embossing patterns
- Canadian Aboriginal Syllabics .
- Cherokee
- Coptic
- Cyrillic
- Devanagari
- Ethiopic
- Georgian
- Greek
- Gujarati
- Gurmukhi
- Hangul (Korean)
- Han (Kanji, Hanja, Hanzi)
- Japanese (Kanji, Hiragana and Katakana)
- Hebrew
- Hiragana
- Katakana
- Khmer (Cambodian)
- Kannada
- Lao
- Latin
- Malayalam
- Mongolian
- Myanmar (Burmese)
- Oriya
- Syriac
- Tamil
- Thai
- Tibetan
- Yi
- Zhuyin (Bopomofo) and more

### Unicode Standard for Indic Scripts:

The Unicode Indic ranges are based on the Indian standard ISCII (Indian Standard Code for Information Interchange, 1988). Later updates have been made. Latest version is UNICODE version 4.1.0.

Following table gives a list of scripts, Unicode range, languages which use these scripts.

Script	Unicode Range	Major Languages
Devanagari	U+0900 to U+097F	Hindi, Marathi, Sanskrit, Kankari, Sindhi
Bengali	U+0980 to U+09FF	Bengali, Assamese, Manipuri
Gurumukhi	U+0A00 to U+0A7F	Punjabi
Gujarati	U+0A80 to U+0AFF	Gujarati
Oriya	U+0B00 to U+0B7F	Oriya
Tamil	U+0B80 to U+0BFF	Tamil
Telugu	U+0C00 to U+0C7F	Telugu
Kannada	U+0C80 to U+0CFF	Kannada
Malayalam	U+0D00 to U+0D7F	Malayalam

### Devanagari Script: A Brief

#### Introduction:

Around the 11th century AD, Devanagari alphabet derived from the Brahmi script. It was developed for writing Sanskrit but later adapted for writing many other languages. The Devanagari block of the Unicode Standard is based on ISCII-1988. Unicode Standard encodes Devanagari characters in the same relative position as those coded in position A0-FF416 in the ISCII-1988 standard.

#### Important points:

- Consonant character carry with it an inherent vowel which can be changed or muted by diacritics or *matra* as eg.

क + ि → कि [ki]

- Vowels can be written as independent letters also.

अ, आ, इ, ई, उ, ऊ, ऋ, ए, ऐ, ऌ, ओ, औ, ऑ

- Devanagari conjuncts the consonants that lose the vowel typically lose their characteristic vertical bar such glyphs are referred to as *half-forms*. For example,

स् [s] + म [ma] → स्म [sma].

### Devanagari alphabet:

#### Independent Vowels Letters:

The independent vowels in Devanagari are letters that stand on their own. The writing system treats independent vowels as orthographic Consonant-Vowel syllables in which the consonant is null. The independent vowel letters are used to write syllables that start with a vowel.

#### Dependent Vowel Sign (Matras):

Dependent vowels used for writing noninherent vowels and are generally referred to as vowel signs, or as Matras. Dependent vowels do not stand alone. Explicit appearance of a dependent vowel in a syllable overrides the inherent vowel of a single consonant

Primary Independent vowels						
	Short		Long		Diphthongs	
	Initial	Diacritic	Initial	Diacritic	Initial	Diacritic
Unrounded low central	अ a	प pa	आ aa	पा paa		
Unrounded high front	इ i	पि pi	ई ee	पि pee		
Rounded high back	उ u	पु pu	ऊ oo	पू poo		
Syllabic variant	ऋ ri	पृ pri	ऌ re	प्रे pre		
Secondary Independent vowels						
Unrounded front			ए e	पे pe	ऐ ai	पै pai
Rounded back			ओ o	पो po	औ au	पौ pau

letter. A one-to-one correspondence exists between the independent vowels and the dependent vowel signs.

Dependent Vowel sign	ा	ि	ी	ु	ू	ृ	ै	ौ	ँ	ं	ो	ौ	ों
----------------------	---	---	---	---	---	---	---	---	---	---	---	---	----

#### Consonants Letters:

Each consonant letter represents a single consonantal sound but also having an inherent vowel, generally the short vowel /a/ in Devanagari. Thus Devanagari Letter represents not just /k/ but also /ka/. In the presence of a dependent vowel the inherent vowel associated with a consonant letter is overridden by the dependent vowel.

	Voiceless plosives		Voiced plosives		Nasals
	unaspirated	aspirated	unaspirated	aspirated	
Velar	क ka	ख kha	ग ga	घ gha	ङ ṅa
Palatal	च ca	छ cha	ज ja	झ jha	ञ ña
Retroflex	ट ṭa	ठ ṭha	ड ḍa	ढ ḍha	ण ṇa
Dental	त ta	थ tha	द da	ध dha	न na
Labial	प pa	फ pha	ब ba	भ bha	म ma

### Occlusives

	Palatal	Retroflex	Dental	Labial
Sonorants	य ya	र ra	ल la	व va
Sibilants	श ṣa	ष ṣa	स sa	

### Sonorants and fricatives

#### Other Letters

ह ha      ळ ḷa

अं (an) anusvara-nasalises vowel

अँ (am) anunasika/candrabindu-nasalises vowel

अः (ah) visarga-adds voiceless breath after vowel

प् (p) virama-mutes vowel

#### Numerals:

० १ २ ३ ४ ५ ६ ७ ८ ९ १०  
0 1 2 3 4 5 6 7 8 9 10

#### Halant Sign:

A hal sign is used to show a consonant without its inherent vowel. Hal sign has been wrongly annotated as Virama in Unicode.

TA<sub>n</sub> + HAL<sub>n</sub> → TA<sub>d</sub>

त + ् → त्

#### Consonant Conjuncts:

When two consonant combined, then first consonant appear without its inherent vowel where as second

one with it. This *shyuktakshar* is called as consonant conjuncts.

#### Example: Conjunct Formations

(1) GA<sub>d</sub> + DHA<sub>l</sub> → GA<sub>h</sub> + DHA<sub>n</sub>

ग् + ध → ग्ध

(2) KA<sub>d</sub> + KA<sub>l</sub> → K.KA<sub>n</sub>

क् + क → क्क (a vartical conjunct)

#### Explicit Halant

A hal sign used to create dead consonant, that is combined with subsequent consonants to form conjuncts. This behavior of hal sign results hal sign not being show visually. However, this default behavior is not desired when a dead consonant should be excluded from conjunct formation, in which case the hal sign is visibly rendered. For this Unicode Standard adopts the convention of placing the character U+200C ZERO WIDTH NON-JOINER immediately after the encoded dead consonant that is to be excluded from conjunct formation. In this case, the hal sign is always depicted as appropriate for the consonant to which it is attached.

#### Example:

क् + ष → क्ष

KA<sub>d</sub> + SSHA<sub>l</sub> → KA<sub>d</sub> + SSHA<sub>l</sub>



KA<sub>d</sub> + ZWNJ + SSHA<sub>I</sub> → KA<sub>d</sub> + SSHA  
 क् + ZWNJ + ष → क्ष

### Explicit Half-Consonants:

Dead consonant form diminished into the conjunct form in forming a conjunct, only remain visible as a half-consonant form. A half-consonant form loss its inherent vowel stem.

To explicitly encode a half-consonant form, the Unicode Standard adopts the convention of placing the character U+200D ZERO WIDTH JOINER immediately after the encoded dead consonant.

### Example:

KA<sub>d</sub> + ZWJ + SSHA<sub>I</sub> → KA<sub>h</sub> + SSHA<sub>n</sub>  
 क् + ZWJ + ष → क्ष

This encoding of half-consonant forms also applies in the absence of a base letterform.

### Example:

GA<sub>d</sub> + ZWJ → GA<sub>h</sub>  
 ग् + ZWJ → र

### Example:

प्रतीक्षा = प र त ी क ष ा  
 092A 094D 0930 0924 0940 0958 094D 0937 093E

## Bengali Script: A Brief

### Introduction:

Bangla is fourth most popular language in the world. Bengali alphabet is derived from the **Brahmi alphabet** and closely related to the **Devanagari** alphabet. Bengali has two literary styles, Sadhubhasa (elegant language) and Chaltibhasa (current language). Differences between Sadhubhasa and Chaltibhasa are in pronouns and verb conjugations.

### Important points:

- Special conjunct letters are used, when consonants occur together in clusters. The inherent vowel only applies to the final consonant.

- The Bengali alphabet is a syllabic alphabet in which consonants all have an inherent vowel which has two different pronunciations.
- Vowels can be written as independent letters also.
- Bengali script used to write Bengali, Assamese, Manipuri, Garo, Mundari languages.

### Bengali alphabet:

#### Vowels and vowel diacritics:

There are 11 vowels in Bangla. In Unicode independent vowels are placed in the range U+0985 to U+0994. Some of the places are reserved (098D, 098E, 0991, 0992).

অ	আ	ই	ঈ	উ	ঊ	ঋ	এ	ঐ	ও	ঔ
a	ā	i	ī	u	ū	r̥	e	ai	o	au
[ɔ, ɒ]	[æ]	[i, ɪ]	[i]	[u, ʊ]	[u]	[r]	[e, æ]	[aj]	[o]	[ow]
ক	কা	কি	কী	কু	কূ	ক্	কে	কৈ	কো	কৌ
ka	kā	ki	kī	ku	kū	k̥	ke	kai	ko	kau

#### Consonants:

There are 39 consonant in modern Bangla alphabet. In Unicode consonants are placed in the range U+0995 to U+09B9 and again from U+09DC to U+09DE.

ক	খ	গ	ঘ	ঙ	চ	ছ	জ	ঝ	ঞ
ka	kha	ga	gha	ṅa	ca	cha	ja	jha	ña
[kɔ]	[kʰɔ]	[gɔ]	[gʱɔ]	[ŋɔ]	[tɔ]	[tʰɔ]	[dɔ]	[dʱɔ]	[ɲɔ]
ট	ঠ	ড	ঢ	ণ	ত	থ	দ	ধ	ন
ṭa	ṭha	ḍa	ḍha	ṇa	ta	tha	da	dha	na
[ʈɔ]	[ʈʰɔ]	[ʈɔ]	[ʈʱɔ]	[ɳɔ]	[tɔ]	[tʰɔ]	[dɔ]	[dʱɔ]	[nɔ]
প	ফ	ব	ভ	ম	য	র	ল	শ	স
pa	pha	ba	bha	ma	ya	ra	la	sa	sa
[pɔ]	[pʰɔ]	[bɔ]	[bʱɔ]	[mɔ]	[jɔ]	[rɔ]	[lɔ]	[ʃɔ]	[sɔ]

ক	ক্কা	ক্ট	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা
ক	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা	ক্কা		

### Additional symbols:

— anunāsika/candrabindu - nasalises the vowel its attached to  
 ৳ anusvāra - adds an 'ng' [ŋ] sound to a consonant, e.g. বং = bam [baŋ]  
 ৲ hasanta - mutes the inherent vowel of a consonant, e.g. ব = ba [ba] ৲ = b [b]

### Numerals:

০	১	২	৩	৪	৫	৬	৭	৮	৯	১০
sunna	ek	dui	tin	cār	pānc	chay	sāt	āt	nay	daś (Bengali)
0	1	2	3	4	5	6	7	8	9	10

### Sample text in Bengali:

সমস্ত মানুষ স্বাধীনভাবে সমান মর্যাদা এবং অধিকার নিয়ে জন্মগ্রহণ করে  
 তাঁদের বিবেক এবং বুদ্ধি আছে; সুতরাং সকলেরই একে অপরের প্রতি  
 আন্তরিক মনোভাব নিয়ে আচরণ করা উচিত।

### Transliteration

Shômosto manush shadhinbhave shôman môrjada  
 ebong odhikar nie jônmogrohon kôre. Tâder bibek  
 ebong buddhi achhe; shutorang shôkoleri êke ôperer  
 proti bhratrittoshulobh mônobhab nie achoron kôra  
 uchit.

### Translation

All human beings are born free and equal in dignity  
 and rights. They are endowed with reason and  
 conscience and should act towards one another in a  
 spirit of brotherhood.

### Malayalam Script: A Brief

#### Introduction:

Malayalam script developed from *vattezhuthu* (round writing) script, a descendant of the *Brahmi* script. Consonant and diacritics written as separately for printing purpose, because it was not easy to print the Malayalam script.

#### Important points:

- This is a syllabic alphabet in which all consonants have an inherent vowel. Diacritics are used to change the inherent vowel.
- When they appear the beginning of a syllable, vowels are written as independent letters.
- When certain consonants occur together, special

conjunct symbols are used which combine the essential parts of each letter.

### Malayalam alphabet:

#### Independent Vowels (*svaram*) Letters:

Independent vowels in Malayalam are signs that stand on their own. These are used to write syllables, which start with a vowel.

അ	ആ	ഇ	ഈ	ഉ	ഊ	എ	ഐ	ഒ	ഓ	ഔ
a	ā	i	ī	u	ū	e	ē	o	ō	au
[a]	[ā]	[i]	[ī]	[u]	[ū]	[e]	[ē]	[o]	[ō]	[au]

#### Dependent vowel signs:

These occur only in combination with a base consisting of a sign for a single consonant or a consonant cluster. Explicit appearance of a dependent vowel in a syllable overrides the inherent vowel of the consonant/consonant cluster.

Example: Vowel diacritics with ka

ക	കാ	കി	കീ	ക/കു	കു	കു/കു	കെ	കേ
ka	kā	ki	kī	ku	kū	ke	ke	kē
കൈ	കൊ	കോ	കൗ	കം	കഃ	ക്ക	ക്ക	
kai	ko	kō	kau	kam	kaḥ	kk	kk	

### Consonants (*vyantjanam*) Letters:

Each consonant letter represents a single consonant sound followed by the inherent vowel /a/.

ക	ka [ka]	ഖ	kha [kʰa]	ഗ	ga [ga]	ഘ	gha [gʱa]	ങ	ṅa [ŋa]
ച	ca [tʃa]	ഛ	cha [tʃʰa]	ജ	ja [dʒa]	ഝ	jha [dʒʱa]	ഞ	ña [ɲa]
ട	ta [ta]	ത	tha [tʰa]	ഡ	da [da]	ഢ	dha [dʱa]	ണ	ṇa [ɳa]
ത	ta [ta]	ഥ	tha [tʰa]	ദ	da [da]	ധ	dha [dʱa]	ന	na [na]
പ	pa [pa]	ഫ	pha [pʰa]	ബ	ba [ba]	ഭ	bha [bʱa]	മ	ma [ma]
യ	ya [ja]	ര	ra [ra]	ല	la [la]	വ	va [va]		
ശ	śa [ʃa]	ഷ	ṣa [ʃʰa]	സ	sa [sa]	ഹ	ha [ha]		
ള	ḷa [la]	ഴ	ḷa [ʒa]	റ	ra [ra]				

#### A selection of conjunct consonants

സ്ക	യ്ക	ച്ക	ജ്ക	ട്ക	ഡ്ക	ബ്ക	പ്ക	യ്ക	സ്ക	പ്ക	സ്ക	പ്ക
sk	ykk	ck	jk	tk	dk	bdk	pk	ykk	sk	pk	sk	pk
ക്ക	ക്ക	ക്ക	ക്ക	ക്ക	ക്ക	ക്ക	ക്ക	ക്ക	ക്ക	ക്ക	ക്ക	ക്ക
kk	kt	nk	rk	rk	rk	rk	rk	rk	rk	rk	rk	rk

Consonant diacritics  
 ന്യ പ്ര/പ ശ  
 ny pr sv

## Numerals

൧	൨	൩	൪	൫	൬	൭	൮	൯	൧൦
ഒന്നു	രണ്ടു	മൂന്നു	നാലു	അഞ്ചു	ആറു	ഏഴു	എട്ടു	ഒമ്പതു	പത്തു
onnu	ranu	munn	nal	aricu	aru	il	ettu	onpatu	patu
1	2	3	4	5	6	7	8	9	10

## Sample text:

### Traditional orthography

മനുഷ്യരെല്ലാവരും തുല്യാവകാശങ്ങളോടും അന്തസ്സോടും സ്വാതന്ത്ര്യത്തോടുംകൂടി ജനിച്ചവരാണ്. മനുഷ്യനാ വിവേകബുദ്ധിയും മനസ്സാക്ഷിയും സിദ്ധമായിരിക്കുന്നത്.

### Reformed orthography

മനുഷ്യരെല്ലാവരും തുല്യാവകാശങ്ങളോടും അന്തസ്സോടും സ്വാതന്ത്ര്യത്തോടുംകൂടി ജനിച്ചവരാണ്. മനുഷ്യനാ വിവേകബുദ്ധിയും മനസ്സാക്ഷിയും സിദ്ധമായിരിക്കുന്നത്.

## Transliteration

Syarellavarum tulyavakasannalotum antassotum  
svatantryattotumkuti janiccavaran marusyannu  
vivekabuddhiyum manassaksiyum  
siddhamayirikkunnat.

## Translation

All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.

## UTF:

A Universal transformation format (UTF) is an algorithmic mapping from every Unicode code point (except surrogate code points) to a unique byte sequence.

Actual implementations in computer systems represent integers in specific code units of particular size (8 bit, 16 bit, or 32 bit). Encoding forms specify how each integer (code point) for a Unicode character is to be expressed as a sequence of one or more code units.

**UTF-16:** UTF-16 is a variable-width encoding form. In the UTF-16 encoding form, code points in the range U+0000...U+FFFF are represented as a single

16-bit code unit; code points in the supplementary planes, in the range U+10000..U+10FFFF, are instead represented as pairs of 16-bit code units. These pairs of special code unit are known as surrogate pairs.

**UTF-16** optimizes the representation of characters in the Basic Multilingual Plane (BMP) that is, the range U+0000..U+FFFF. BMP contains the vast majority of common-use characters for all modern scripts of the world, each character requires only one 16-bit code unit, thus requiring just half the memory or storage of the UTF-32 encoding form.

**UTF-8:** To meet the requirements of byte-oriented, ASCII-based systems, UTF-8 is used. It is a variable-width encoding form that preserves ASCII transparency, making use of 8-bit code units.

Unicode code points U+0000..U+007F are converted to single bytes 0x00..0x7F in UTF-8, and are thus indistinguishable from ASCII itself. Beyond the ASCII range of Unicode, many of the non-ideographic scripts are represented by two bytes per code point in UTF-8 and U+FFFF are represented by three bytes; and supplementary code points above U+FFFF require four bytes.

**UTF-8** is typically the preferred encoding 8; all non-surrogate code points between U+ form for HTML and similar protocols, particularly for the Internet. It is already inherently byte-serialized, as for most existing 8-bit character sets; strings of UTF-8 work easily with C or other programming languages.

**UTF-32:** Each Unicode point is represented directly by a single 32-bit code unit.

It is a fixed-width character encoding form.

For ex.

U+10000 is represented as <00010000> in UTF-32.

**UTF-32** preferred where memory for character is no particular concern, but where fixed-width, single code unit access to char is desire.



### Conversion from UTF-16 to UTF-8:

Use this table for converting UTF-16 to UTF-8

U-00000000	U-0000007F:	0xxxxxxx
U-00000080	U-000007FF:	110xxxxx 10xxxxxx
U-00000800	U-0000FFFF:	1110xxxx 10xxxxxx 10xxxxxx
U-00010000	U-001FFFFF:	11110xxx 10xxxxxx 10xxxxxx 10xxxxxx
U-00200000	U-03FFFFFF:	111110xx 10xxxxxx 10xxxxxx 10xxxxxx 10xxxxxx
U-04000000	U-7FFFFFFF:	1111110x 10xxxxxx 10xxxxxx 10xxxxxx 10xxxxxx 10xxxxxx

Devanagari script lies between U+0900 - U+097F. Thus we will use 3<sup>rd</sup> row of above table. To explain the process of conversion, we will consider one word of hindi language.

**Example:** भारतवर्ष

Take an alphabet one by one from starting of word. So, 'भ' is the first character. UTF-16 code (Hexadecimal equivalent) of भ is 092D and its binary is 0000 1001 0010 1101.

Now, follow the formula of above table for changing code from UTF-16 to UTF-8. Put binary digits below the symbol x in formula from right to left and just copy the binary digit 1 and 0 as it is.

Now, make the six pairs of 4 digits of resultant bits. Write down its hexadecimal equivalent. It will be the UTF-8 code of भ.

$$\begin{array}{rcl}
 \bullet \text{ भ (in UTF-16) - 092D} & & 0000\ 1001\ 0010\ 1101 \text{ (Binary form)} \\
 \begin{array}{r}
 \text{1110xxxx} \\
 \text{11100000} \\
 \text{E} \quad 0
 \end{array} & & \begin{array}{r}
 10xxxxxx \quad 10xxxxxx \\
 \text{10100100} \quad \text{10101101} \\
 \text{A} \quad 4 \quad \text{A} \quad \text{D}
 \end{array} \\
 \text{भ (in UTF-8) - E0A4AD} & & 
 \end{array}$$

Follow the same process for other characters:

$$\begin{array}{rcl}
 \bullet \text{ ा (in UTF-16) - 093E} & & 0000\ 1001\ 0011\ 1110 \text{ (Binary form)} \\
 \begin{array}{r}
 \text{1110xxxx} \\
 \text{11100000} \\
 \text{E} \quad 0
 \end{array} & & \begin{array}{r}
 10xxxxxx \quad 10xxxxxx \\
 \text{10100100} \quad \text{10111110} \\
 \text{A} \quad 4 \quad \text{B} \quad \text{E}
 \end{array} \\
 \text{ा (in UTF-8) - E0A4BE} & & 
 \end{array}$$

$$\begin{array}{rcl}
 \bullet \text{ र (in UTF-16) - 0930} & & 0000\ 1001\ 0011\ 0000 \text{ (Binary form)} \\
 \begin{array}{r}
 \text{1110xxxx} \\
 \text{11100000} \\
 \text{E} \quad 0
 \end{array} & & \begin{array}{r}
 10xxxxxx \quad 10xxxxxx \\
 \text{10100100} \quad \text{10110000} \\
 \text{A} \quad 4 \quad \text{B} \quad 0
 \end{array} \\
 \text{र (in UTF-8) - E0A4B0} & & 
 \end{array}$$



- त(in UTF-16) - 0924**      0000 1001 0010 0100(Binary form)
 

➤	1 1 1 0 x x x x	1 0 x x x x x x	1 0 x x x x x x
	<u>1 1 1 0 0 0 0 0</u>	<u>1 0 1 0 0 1 0 0</u>	<u>1 0 1 0 0 1 0 0</u>
	E      0	A      4	A      4

**त(in UTF-8) - E0A4A4**
- व(in UTF-16) - 0935**      0000 1001 0011 0101(Binary form)
 

➤	1 1 1 0 x x x x	1 0 x x x x x x	1 0 x x x x x x
	<u>1 1 1 0 0 0 0 0</u>	<u>1 0 1 0 0 1 0 0</u>	<u>1 0 1 1 0 1 0 1</u>
	E      0	A      4	B      5

**व(in UTF-8) - E0A4B5**
- र(in UTF-16) - 0930**      0000 1001 0011 0000(Binary form)
 

➤	1 1 1 0 x x x x	1 0 x x x x x x	1 0 x x x x x x
	<u>1 1 1 0 0 0 0 0</u>	<u>1 0 1 0 0 1 0 0</u>	<u>1 0 1 1 0 0 0 0</u>
	E      0	A      4	B      0

**र(in UTF-8) - E0A4B0**
- ऌ(in UTF-16) - 094D**      0000 1001 0100 1101(Binary form)
 

➤	1 1 1 0 x x x x	1 0 x x x x x x	1 0 x x x x x x
	<u>1 1 1 0 0 0 0 0</u>	<u>1 0 1 0 0 1 0 1</u>	<u>1 0 0 0 1 1 0 1</u>
	E      0	A      5	8      D

**ऌ(in UTF-8) - E0A58D**
- ष(in UTF-16) - 0937**      0000 1001 0011 0111(Binary form)
 

➤	1 1 1 0 x x x x	1 0 x x x x x x	1 0 x x x x x x
	<u>1 1 1 0 0 0 0 0</u>	<u>1 0 1 0 0 1 0 0</u>	<u>1 0 1 1 0 1 1 1</u>
	E      0	A      4	B      7

**ष(in UTF-8) - E0A4B7**

### Memory Comparison:

UTF-16 taking 2 bytes for storing each alphabet, whereas UTF-8 takes 3 bytes for each alphabet of Devanagari script. In the above example takes 7\*2 bytes in UTF-16 format and 7\*3 bytes in UTF-8 format. Memory requirement for UTF-8 will be 1.5 times more than UTF-16.

### References:

- Characteristics of Indic Scripts - <http://people.w3.org/rishida/scripts/indic-overview/>
- ISCII - <http://tdil.mit.gov.in/standards.htm>
- Unicode - <http://www.unicode.org/>
- Unicode Standard For Indic Scripts - <http://tdil.mit.gov.in/uni.htm>  
<http://www.omniglot.com/writing>
- UTF - The Unicode Standard 4.0  
(Addison Wesley Publication)  
<http://www.unicode.org/versions/Unicode4.0.1>  
<http://www.unicode.org/versions/Unicode4.1.0>

### Courtesy :

Manoj Jain & Vijay Kumar  
Department of Information Technology  
New Delhi-110003  
Manish Kumar Singh  
CDAC, Noida