L2/02-268

3. *Standardization*

3.1 Revision of Unicode Standard-3.0 for Devanagari Script

Unicode Standards are widely being used by the Industry for the development of Multilingual Softwares. Indian scripts are also included in the Unicode Standards. Unicode consortium had taken the basic inputs for standardisation of Indian scripts from ISCII-1988 document. There are some deficiencies in the present Unicode Standards for Indian Scripts and need to be removed for proper representation of the Indian Scripts.

MIT is the voting member of the Unicode Consortium. The ministry has collected inputs for each of the Indian scripts in order to make a single unified presentation of Indian scripts to the Unicode Consortium. A number of meetings starting from November 2000 have been organized with the concerned State Governments/ Organizations/ Experts and Industry on this subject. Based on the discussions / feedback draft Code charts for each of the script have been prepared along-with code details. The first draft proposal of the proposed changes in the existing Unicode Standards for Indian scripts was published in the May 2001 issue of the *TDIL Newsletter*- VishwaBharat@tdil to get the feedback from experts/ industry / users working in the area of Indian Language Software Development. The news letter was also sent to the members of the Unicode Consortium for their comments and initial response. The draft proposal was discussed in the Unicode Technical Committee (UTC) meeting held in USA in November 2001. The minutes of the UTC Meeting are available in document No. L2/01-430R and L2/01-431R of Unicode Consortium. The URL for viewing these documents are as given below:
http://tdil.mit.gov.in/newsletter1.htm
http://www.unicode.org/L2/L2001/01304-feedback.pdf

After getting the feedback from state government / linguists / experts / industry the second draft proposal is also being prepared. The final draft of the Devanagari script is published here for your reference. The whole write-up is divided into three parts:

1. Code chart - For quick reference of the characters and their code value.

2. Code Details - For reference of the character names and annotations.

3. Devanagari:A brief review of the script – This write-up covers various aspects of the script.

Devanagari Code Chart

	090	091	092	093	094	095	096	097
0		ऐ	ठ	र	ी	ॐ	ॠ	०
1	ँ	आ	ड	ऱ	ु	॑	ॡ	ॱ
2	ं	ओ	ढ	ल	ू	॒	ॢ	क्ष
3	ः	ओ	ण	ळ	ृ	॓	ॣ	ज्ञ
4		औ	त	ऴ	ॄ	॔	।	श्र
5	अ	क	थ	व	ॅ	ॕ	॥	ऽ
6	आ	ख	द	श	ॆ	ॖ	०	ॺ
7	इ	ग	ध	ष	े	ॗ	१	ॻ
8	ई	घ	न	स	ै	क़	२	ॼ
9	उ	ङ	ऩ	ह	ॉ	ख़	३	ॽ
A	ऊ	च	प	़	ॊ	ग़	४	ॾ
B	ऋ	छ	फ		ो	ज़	५	ॿ
C	ऌ	ज	ब	़	ौ	ड़	६	ঀ
D	ऍ	झ	भ	ऽ	्	ढ़	७	
E	ऎ	ञ	म	ा		फ़	८	
F	ए	ट	य	ि		य़	९	

Devanagari Code Chart Details

Code Point	Character	Description
0901	ँ	DEVANAGARI SIGN CANDRABINDU = anunasika .0310 combining candrabindu
0902	ं	DEVANAGARI SIGN ANUSVARA = bindu
0903	ः	DEVANAGARI SIGN VISARGA

Independent vowels

0905	अ	DEVANAGARI LETTER A
0906	आ	DEVANAGARI LETTER AA
0907	इ	DEVANAGARI LETTER I
0908	ई	DEVANAGARI LETTER II
0909	उ	DEVANAGARI LETTER U
090A	ऊ	DEVANAGARI LETTER UU
090B	ऋ	DEVANAGARI LETTER VOCALIC R
090C	ऌ	DEVANAGARI LETTER VOCALIC L
090D	ऍ	DEVANAGARI LETTER CANDRA E
090E	ऎ	DEVANAGARI LETTER SHORT E • for transcribing Dravidian short e
090F	ए	DEVANAGARI LETTER E
0910	ऐ	DEVANAGARI LETTER AI
0911	ऑ	DEVANAGARI LETTER CANDRA O
0912	ओ	DEVANAGARI LETTER SHORT O • for transcribing Dravidian short o
0913	ओ	DEVANAGARI LETTER O
0914	औ	DEVANAGARI LETTER AU

Consonants

0915	क	DEVANAGARI LETTER KA
0916	ख	DEVANAGARI LETTER KHA
0917	ग	DEVANAGARI LETTER GA
0918	घ	DEVANAGARI LETTER GHA
0919	ङ	DEVANAGARI LETTER NGA
091A	च	DEVANAGARI LETTER CA
091B	छ	DEVANAGARI LETTER CHA
091C	ज	DEVANAGARI LETTER JA
091D	झ	DEVANAGARI LETTER JHA
091E	ञ	DEVANAGARI LETTER NYA
091F	ट	DEVANAGARI LETTER TTA
0920	ठ	DEVANAGARI LETTER TTHA
0921	ड	DEVANAGARI LETTER DDA
0922	ढ	DEVANAGARI LETTER DDHA
0923	ण	DEVANAGARI LETTER NNA
0924	त	DEVANAGARI LETTER TA
0925	थ	DEVANAGARI LETTER THA
0926	द	DEVANAGARI LETTER DA
0927	ध	DEVANAGARILETTERDHA
0928	न	DEVANAGARI LETTER NA
0929	ऩ	DEVANAGARI LETTER NNNA • for transcribing Dravidian alveolar n ≡ 0928 न 093C ़
092A	प	DEVANAGARI LETTER PA
092B	फ	DEVANAGARI LETTER PHA
092C	ब	DEVANAGARI LETTER BA
092D	भ	DEVANAGARI LETTER BHA
092E	म	DEVANAGARI LETTER MA
092F	य	DEVANAGARI LETTER YA
0930	र	DEVANAGARI LETTER RA
0931	ऱ	DEVANAGARI LETTER RRA • for transcribing Dravidian alveolar r
0932	ल	DEVANAGARI LETTER LA
0933	ळ	DEVANAGARI LETTER LLA
0934	ऴ	DEVANAGARI LETTER LLLA • for transcribing Dravidian l ≡ 0933 ळ 093C ़
0935	व	DEVANAGARI LETTER VA
0936	श	DEVANAGARI LETTER SHA

0937	ष	DEVANAGARI LETTER SSA
0938	स	DEVANAGARI LETTER SA
0939	ह	DEVANAGARI LETTER HA
093A	ऺ	DEVANAGARI INVISIBLE LETTER

Various signs

093C	़	DEVANAGARI SIGN NUKTA • for extending the alphabet to new letters
093D	ऽ	DEVANAGARI SIGN AVAGRAHA
093E	ा	DEVANAGARI VOWEL SIGN AA

Dependent vowel signs

093F	ि	DEVANAGARI VOWEL SIGN I • stands to the left of the consonant
0940	ी	DEVANAGARI VOWEL SIGN II
0941	ु	DEVANAGARI VOWEL SIGN U
0942	ू	DEVANAGARI VOWEL SIGN UU
0943	ृ	DEVANAGARI VOWEL SIGN VOCALIC R
0944	ॄ	DEVANAGARI VOWEL SIGN VOCALIC RR
0945	ॅ	DEVANAGARI VOWEL SIGN CANDRA E = candra
0946	े	DEVANAGARI VOWEL SIGN SHORT E • for transcribing Dravidian vowels
0947	े	DEVANAGARI VOWEL SIGN E
0948	ै	DEVANAGARI VOWEL SIGN AI
0949	ॉ	DEVANAGARI VOWEL SIGN CANDRA O
094A	ॊ	DEVANAGARI SHORT O • for transcribing Dravidian vowels
094B	ो	DEVANAGARI VOWEL SIGN O
094C	ौ	DEVANAGARI VOWEL SIGN AU

Various signs

094D	्	DEVANAGARI SIGN HAL • suppresses inherent vowel
094E		<reserved>
094F		<reserved>
0950	ॐ	DEVANAGARI OM
0951	॑	DEVANAGARI STRESS SIGN UDATTA
0952	॒	DEVANAGARI STRESS SIGN ANUDATTA
0953	॓	DEVANAGARI GRAVE ACCENT
0954	॔	DEVANAGARI ACUTE ACCENT
0955	◌	DEVANAGARI ANUSWARA • Used in Sanskrit Yajurveda
0956	◌	DEVANAGARI SIGN YAJURVEDIC ANUSWARA • Used in Sanskrit
0957	◌	DEVANAGARI JIVHAMULIYA • Used in Sanskrit

Additional consonants (Their use should be avoided)

0958	क़	DEVANAGARI LETTER QA ≡0915 क 093C ़
0959	ख़	DEVANAGARI LETTER KHHA ≡0916 ख 093C ़
095A	ग़	DEVANAGARI LETTER GHHA ≡0917 ग 093C ़
095B	ज़	DEVANAGARI LETTER ZA ≡091C ज 093C ़
095C	ड़	DEVANAGARI LETTER DDDHA

		≡0921 ड 093C ़
095D	ढ़	DEVANAGARI LETTER RHA ≡0922 ढ 093C ़
095E	फ़	DEVANAGARI LETTER FA ≡092B फ 093C ़
095F	य़	DEVANAGARI LETTER YYA ≡092F य 093C ़

Generic additions

0960	ॠ	DEVANAGARI LETTER VOCALIC RR
0961	ॡ	DEVANAGARI LETTER VOCALIC LL
0962	ॢ	DEVANAGARI VOWEL SIGN VOCALIC L
0963	ॣ	DEVANAGARI VOWEL SIGN VOCALIC LL
0964	।	DEVANAGARI PURNA VIRAMA = phrase separator
0965	॥	DEVANAGARI DEERGH VIRAM

Digits

0966	०	DEVANAGARI DIGIT ZERO
0967	१	DEVANAGARI DIGIT ONE • Shape १ is also used in Hindi
0968	२	DEVANAGARI DIGIT TWO
0969	३	DEVANAGARI DIGIT THREE
096A	४	DEVANAGARI DIGIT FOUR
096B	५	DEVANAGARI DIGIT FIVE • Shape ५ is also used in Hindi
096C	६	DEVANAGARI DIGIT SIX
096D	७	DEVANAGARI DIGIT SEVEN
096E	८	DEVANAGARI DIGIT EIGHT • Shape ८ is also used in Hindi
096F	९	DEVANAGARI DIGIT NINE • Shape ९ is also used in Hindi

Devanagari-specific additions

0970	॰	DEVANAGARI ABBREVIATION SIGN
0971	ॱ	DEVANAGARI CURRENCY SIGN
0972	क्ष	DEVANAGARI LETTER KSHA
0973	ज्ञ	DEVANAGARI LETTER GNYA
0974	श्र	DEVANAGARI SIGN SHRA
0975	◌ॎ	DEVANAGARI SIGN REPH
0976	श	LETTER SHA Used in Marathi • श (0936) ≡ श (0976)
0977	ल	DEVANAGARI LETTER LA Used in Marathi • ल (0932) ≡ ल (0977)
0978	ऱ	DEVANAGARI SIGN/ SOFT RA/ •Used in MARATHI as consonant modifier
0979	ऩ	DEVANAGARI CONSONANT • Used for SINDHI implosive placed just below the consonant
097A	ऺ	DEVANAGARI CONSONANT • Used for SINDHI implosive placed just below the consonant
097B	ऻ	DEVANAGARI CONSONANT •Used for SINDHI implosive placed just below the consonant
097C	ऽ	DEVANAGARI CONSONANT • Used for SINDHI implosive placed just below the consonant

Explanations for Revised Devanagari Code Chart

Devanagari : U +0900-U +097F

The Devanagari script is used for writing classical Sanskrit and its modern historical derivative, Hindi. Extensions to Devanagari are used to write other related languages of India (such as Marathi, Konkani, Sindhi and Sanskrit) and of Nepal (Nepali). In addition, the Devanagari script is used to write the dialects of Hindi and various other regional & tribal languages.

All other Indic scripts including Nandi Nagari, as well as the Sinhala script of Sri Lanka, the Tibetan script, and the Southeast Asian scripts (Thai, Lao, Khmer, and Myanmar), are historically connected with the Devanagari script as descendants of the ancient Brahmi script. The entire family of scripts shares a large number of structural features.

The principles of the Indic scripts are covered in some detail in this introduction to the Devanagari script. The remaining introductions to the Indic scripts are abbreviated but highlight any differences from Devanagari where appropriate.

Standards : The Devanagari block of the Unicode Standard is based on ISCII-1988 (Indian Standard Code for Information Interchange). The ISCII standard of 1988 differs from and is an update of earlier ISCII standards issued in 1983 and 1986.

The Unicode Standard encodes Devanagari characters in the same relative position as those coded in positions A0-F4$_{16}$ in the ISCII-1988 standard. The same character code layout is followed for eight other Indic scripts in the Unicode Standard: Bengali, Gurmukhi, Gujarati, Oriya, Tamil, Telugu, Kannada, and Malayalam. This parallel code layout emphasizes the structural similarities of the Brahmi script and follows the stated intention of the Indian coding standards to enable one-to-one mappings between analogous coding positions in different scripts in the family. Sinhala, Thai, Lao, Khmer, and Myanmar depart to a greater extent from the Devanagari structural pattern, so the Unicode Standard does not attempt to provide any direct mappings for these scripts to the Devanagari order.

In November 1991, at the time The Unicode Standard, Version 1.0, was published, the Bureau of Indian Standards published a new version of ISCII in Indian Standard (IS)13194:1991. This new version partially modified the layout and repertoire of the ISCII-1988 standard. Because of these events, the Unicode Standard does not precisely follow the layout of the current version of ISCII. Nevertheless, the Unicode Standard remains a superset of the ISCII-1991 repertoire except for a number of new Vedic extension characters defined in IS 13194:1991 Annex G - Extended Character Set for Vedic. Modern, non-Vedic texts encoded with ISCII-1991 may be automatically converted to Unicode code values and back to their original encoding without loss of information.

Encoding Principles : The writing systems that employ Devanagari and other Indic scripts constitute a cross between syllabic writing systems and phonemic writing systems (alphabets). The effective unit of these writing systems is the orthographic syllable, consisting of a consonant and vowel (CV) core and, optionally, one or more preceding consonants, with a canonical structure of ((C) C) CV. The orthographic syllable need not correspond exactly with a phonological syllable, especially when a consonant cluster is involved, but the writing system is built on phonological principles and tends to correspond quite closely to pronunciation.

The orthographic syllable is built up of alphabetic pieces, the actual letters of the Devanagari script. These pieces consist of three distinct character types: consonant letters with inherent vowel /a/, pure consonant, independent vowels, and dependent vowel signs. In a text sequence, these characters are stored in logical (phonetic) order.

Principles of the Script

Rendering Devanagari Characters : Devanagari characters, like characters from many other scripts, can combine or change shape depending on their context. A character's appearance is affected by its ordering with respect to other characters, the font used to render the character, and the application or system environment. These variables can cause the appearance of Devanagari characters to differ from their nominal glyphs (used in the code charts).

Additionally, a few Devanagari characters cause a change in the order of the displayed characters. This reordering is not commonly seen in non-Indic scripts

and occurs independently of any bidirectional character reordering that might be required.

Consonant Letters : Each consonant letter represents a single consonantal sound but also has the peculiarity of having an inherent vowel, generally the short vowel /a/ in Devanagari and the other Indic scripts. Thus U+0915 DEVANAGARI LETTER KA represents not just /k/ but also /ka/. In the presence of a dependent vowel, however, the inherent vowel associated with a consonant letter is overridden by the dependent vowel.

Consonant letters may also be rendered as half-forms, which are presentation forms used to depict the initial consonant in consonant clusters. These half-forms do not have an inherent vowel. Their rendered forms in Devanagari often resemble the full consonant but are missing the vertical stem, which marks a syllabic core. (The stem glyph is graphically and historically related to the sign denoting the inherent /a/ vowel.)

Some Devanagari consonant letters have alternative presentation forms whose choice depends upon neighboring consonants. This variability is especially notable for U+0930 DEVANAGARI LETTER RA, which has numerous different forms, both as the initial element and as the final element of a consonant cluster. Only the nominal forms, rather than the contextual alternatives, are depicted in the code chart.

The traditional Sanskrit / Devanagari alphabetic encoding order for consonants follows articulatory phonetic principles, starting with pre velar consonants and moving forward to bilabial consonants, followed by liquids, semi vowels and then fricatives, sibilents (Ushma). ISCII and the Unicode standard both observe this traditional order.

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

Dependent Vowel Signs (Matras) : The dependent vowels serve as the common manner of writing noninherent vowels and are generally referred to as vowel signs, or as Matras in Sanskrit. The dependent vowels do not stand alone; rather, they are visibly depicted in combination with a base letterform. A single consonant, or a consonant cluster, may have a dependent vowel applied to it to indicate the vowel quality of the syllable, when it is different from the inherent vowel. Explicit appearance of a dependent vowel in a syllable overrides the inherent vowel of a single consonant letter.

The greatest variation among different Indic scripts is found in the way that the dependent vowels are applied to base letterforms. Devanagari has a collection of nonspacing dependent vowel signs that may appear above or below a consonant letter, as well as spacing dependent vowel signs that may occur to the right or to the left of a consonant letter or consonant cluster. Other Indic scripts generally have one or more of these forms, but what is a nonspacing mark in one script may be a spacing mark in another. Also, some of the Indic scripts have single dependent vowels that are indicated by two or more glyph components and those glyph components may surround a consonant letter both to the left and right or may occur both above and below it.

The Devanagari script has only one character denoting a left-side dependent vowel sign: U+093F DEVANAGARI VOWEL SIGN I. Other Indic scripts either have no such vowel (Telugu and Kannada) or include as many as three of these signs (Bengali, Tamil, Malayalam).

A one-to-one correspondence exists between the independent vowels and the dependent vowel signs. Independent vowels are sometimes represented by a sequence consisting of the independent form of the vowel /a/ followed by a dependent vowel sign. For example Figure 9.1 illustrates this relationship (see the notation formally described in the “Rules for Rendering” later in this section).

Figure 9.1 : Dependent Versus Independent Vowels

/a/ + Dependent Vowel				Independent Vowel
$A_n + I_{vs}$	⟶	$I_{vs} + A_n$	=	I_n
अ + ि	⟶	अि	=	इ
$A_n + U_{vs}$	⟶	$A_n + U_{vs}$	=	U_n
अ + ु	⟶	अु	=	उ

The combination of the independent form of the default vowel /a/ (in the Devanagari script, U+0905

DEVANAGARI LETTER A) with a dependent vowel sign may be viewed as an alternative spelling of the phonetic information normally represented by an isolated independent vowel form. However, these two representations should not be considered equivalent for the purposes of rendering. Higher-level text processes may choose to consider these alternative spellings equivalent in terms of information content, but such an equivalence is not stipulated by this standard.

Hal Sign : Devanagari and other Indic scripts employ a sign known as the hal sign (representing consonant), or vowel omission sign. A hal sign (for example, U+094D DEVANAGARI SIGN HAL) normally serves to cancel (or kill) the inherent vowel of the consonant to which it is applied. The hal functions as a combining character, with its shape varying from script to script. When a consonant has lost its inherent vowel by the application of hal, it is known as a dead consonant; in contrast, a live consonant is one that retains its inherent vowel or is written with an explicit dependent vowel sign. In the Unicode Standard, a dead, consonant is defined as a sequence consisting of a consonant letter followed by a hal sign. The default rendering for a dead consonant is to position the hal as a combining mark bound to the consonant letterform.

For example, if C_n denotes the nominal form of consonant C and C_d denotes the dead consonant form, then a dead consonant is encoded as shown in Figure 9.2.

Figure 9.2 : Dead Consonants

$TA_n + HAL_n \longrightarrow TA_d$

त + ् ⟶ त्

Consonant Conjuncts : The Indic scripts are noted for a large number of consonant conjunct forms that serve as orthographic abbreviations (ligatures) of two or more adjacent letterforms. This abbreviation takes place only in the context of a consonant cluster. An orthographic consonant cluster is defined as a sequence of characters that represents one or more dead consonants (denoted C_d) followed by a normal, live consonant letter (denoted C_l) or an independent vowel having an inherent vowel or a vowel sign representing other vowel.

Under normal circumstances, a consonant cluster is depicted with a conjunct glyph if such a glyph is available in the current font(s). In the absence of a conjunct glyph, the one or more dead consonants that form part of the cluster are depicted using half-form glyphs. In the absence of half-form glyphs, the dead consonants are depicted using the nominal consonant forms combined with visible hal signs (see Figure 9.3).

Figure 9.3 : Conjunct Formations

(1) $GA_d + DHA_l \longrightarrow GA_h + DHA_n$

ग् + ध ⟶ ग्ध

(2) $KA_d + KA_l \longrightarrow K.KA_n$

क् + क ⟶ क्क

A number of types of conjunct formations appear in these examples: (1) a half-form of GA in its combination with the full form of DHA; (2) a vertical conjunct K.KA.

A well-designed Indic script font may contain hundreds of conjunct glyphs, but they are not encoded as Unicode characters because they are the result of ligation of distinct letters. Indic script rendering software must be able to map appropriate combinations of characters in context to the appropriate conjunct glyphs in fonts.

Explicit Hal : Normally a hal sign serves to create dead consonants that are, in turn, combined with subsequent consonants to form conjuncts. This behavior usually results in a hal sign not being depicted visually. Occasionally, 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. To accomplish this goal, the 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.

$KA_d + ZWNJ + SSHA_l \longrightarrow KA_d + SSHA_n$

क् + ZWNJ + ष ⟶ क्ष

Explicit Half-Consonants : When a dead consonant participates in forming a conjunct, the dead

consonant form is often absorbed into the conjunct form, such that it is no longer distinctly visible. In other contexts, however, the dead consonant may remain visible as a half-consonant form. In general, a half-consonant form is distinguished from the nominal consonant form by the loss of its inherent vowel stem, a vertical stem appearing to the right side of the consonant form. In other cases, the vertical stem remains but some part of its right-side geometry is missing.

In certain cases, it is desirable to prevent a dead consonant from assuming full conjunct formation yet still not appear with an explicit hal. In these cases, the half-form of the consonant is used. 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. The ZERO WIDTH JOINER denotes a nonvisible letter that presents linking or cursive joining behavior on either side (that is, to the previous or following letter). Therefore, in the present context, the ZERO WIDTH JOINER may be considered to present a context to which a preceding dead consonant may join so as to create the half-form of the consonant.

For example, if C_h denotes the half-form glyph of consonant C, then a half-consonant form is encoded as shown in Figure 9.5.

Figure 9.5 : Half-Consonants

$KA_d + ZWJ + SSHA_l \longrightarrow KA_h + SSHA_n$

क् + ZWJ + ष ⟶ क्ष

This encoding of half-consonant forms also applies in the absence of a base letterform. That is, this technique may also be used to encode independent half-forms, as shown in Figure 9-6.

Figure 9.6 : Independent Half-Forms

$GA_d + ZWJ \longrightarrow GA_h$

ग् + ZWJ ⟶ ग

Consonant Forms. In summary, each consonant may be encoded such that it denotes a live consonant, a dead consonant that may be absorbed into a conjunct, or the half-form of a dead consonant (see Figure 9.7).

Figure 9.7 : Consonant Forms

क ⟶ क KA_l

क + ◌् ⟶ क् KA_d

क + ◌् + ZWJ ⟶ क KA_h

Rendering

Rules for Rendering : The following provides more formal and detailed rules for minimal rendering of Devanagari as part of a plain text sequence. It describes the mapping between Unicode characters and the glyphs in a Devanagari font. It also describes the combining and ordering of those glyphs.

These rules provide minimal requirements for legibly rendering interchanged Devanagari text. As with any script, a more complex procedure can add rendering characteristics, depending on the font and application.

It is important to emphasize that in a font that is capable of rendering Devanagari, the set of glyphs is greater than the number of Devanagari Unicode characters.

Notation : In the next set of rules, the following notation applies:

C_n Nominal glyph form of consonant C as it appears in the code charts.

C_l A live consonant, depicted identically to C_n.

C_d Glyph depicting the dead consonant form of consonant C.

C_h Glyph depicting the half-consonant form of consonant C.

L_n Nominal glyph form of a conjunct ligature consisting of two or more component consonants. A conjunct ligature composed of two consonants X and Y is also denoted $X.Y_n$.

RA_{sub} A nonspacing combining mark glyph form of the U+0930 DEVANAGARI LETTER RA positioned below or attached to the lower part of a base glyph form.

V_{vs} Glyph depicting the dependent vowel sign form of a vowel V.

HAL The nominal glyph form nonspacing

combining mark depicting U+094D DEVANAGARI SIGN HAL.

- A HAL character is not always depicted; when it is depicted, it adopts this nonspacing mark form.

Dead Consonant Rule : The following rule logically precedes the application of any other rule to form a dead consonant. Once formed, a dead consonant may be subject to other rules described next.

Rl When a consonant C_n precedes a Hal_n it is considered to be a dead consonant C_d. A consonant C_n that does not precede Hal_n is considered to be a live consonant C_l.

$TA_n + Hal_n \longrightarrow TA_d$

त + ् ⟶ त्

Consonant RA Rules : The character U+0930 DEVANAGARI LETTER RA takes one of a number of visual forms depending on its context in a consonant cluster. By default, this letter is depicted with its nominal glyph form (as shown in the code charts). In two contexts, it is depicted using a nonspacing glyph form that combines with a base letterform.

R1 Except for the dead consonant RA_d, when a dead consonant C_d precedes the live consonant RA_l, then C_d is replaced with its nominal form C_n and RA is replaced by the subscript nonspacing mark RA_{sub}, which is positioned so that it applies to C_n.

$THA_d + RA_l \longrightarrow THA_n + RA_{sub}$ Displayed Output

ठ् + र ⟶ ठ + ् ⟶ ठ्र

R2 For certain consonants, the mark RA_{sub} may graphically combine with the consonant to form a conjunct ligature form. These combinations, such as the one shown here, are further addressed by the ligature rules described shortly.

$PHA_d + RA_l \longrightarrow PHA_n + RA_{sub}$ Displayed Output

फ् + र ⟶ फ + ् ⟶ फ्र

R3 If a dead consonant (other than RA_d) precedes RA_d then subsitution of RA for RA_{sub} is performed as described above; however, the Hal that formed RA_d remains so as to form a dead consonant conjunct form.

$TA_d + RA_d \longrightarrow TA_n + RA_{sub} + HAL_n \longrightarrow T.RA_d$

त् + र् ⟶ त + ् + ् ⟶ त्र्

A dead consonant conjunct form that contains an absorbed RA_d may subsequently combine to form a multipart conjunct form.

$T.RA_d + YA_l \longrightarrow T.R.YA_n$

त्र् + य ⟶ त्र्य

Modifier Mark Rules : In addition to vowel signs, three other types of combining marks may be applied to a component of an orthographic (visual) syllable or to the syllable as a whole: nukta, bindus, and svaras.

R4 The nukta sign, which modifies a consonant form, is placed immediately after the consonant in the memory representation and is attached to that consonant in rendering. If the consonant represents a dead consonant, then NUKTA should precede Hal in the memory representation.

$KA_n + NUKTA_n + Hal \longrightarrow QA_d$

क + ़ + ् ⟶ क़्

R5 The other modifying marks, bindus and svaras, apply to the orthographic syllable as a whole and should follow (in the memory representation) all other characters that constitute the syllable. In particular, the bindus should follow any vowel signs, and the svaras should come last. The relative placement of these marks is horizontal rather than vertical; the horizontal rendering order may vary according to typographic concerns.

$KA_n + AA_{vs} + CANDRABINDU_n$

क + ा + ँ ⟶ काँ

Ligature Rules : Subsequent to the application of the rules just described, a set of rules governing ligature formation apply. The precise application of these rules depends on the availability of glyphs in the current font(s) being used to display the text.

R6 If a dead consonant immediately precedes another dead consonant or a live consonant, then the first dead consonant may join the subsequent element to form a two-part conjunct ligature form.

JA_d	+	NYA_l	⟶	$J.NYA_n$		
ज्	+	ञ	⟶	ज्ञ		
TTA_d	+	$TTHA_l$	+	$TT.TTHA_n$		
ट्	+	ठ	⟶	ट्ठ		

R7 A conjunct ligature form can itself behave as a dead consonant and enter into further, more complex ligatures.

$SA_d + TA_d + RA_n \longrightarrow SA_d + T.R.A_n + S.T.RA_n$

स् + त् + र ⟶ स् + त्र + स्त्र

A conjunct ligature form can also produce a half-form.

$T.R.A_d$	+	YA_l	⟶	$T.R.A_h$	+	YA_n
त्र्	+	य	⟶	त्र्य		

R8 If a nominal consonant or conjunct ligature form precedes RA_{sub}, then the consonant or ligature form may join with RA_{sub} to form a multipart conjunct ligature.

KA_n	+	RA_{sub}	⟶	$K.RA_n$
क	+	◌्र	⟶	क्र
PHA_n	+	RA_{sub}	⟶	$PH.RA_n$
फ	+	◌्र	⟶	फ्र

R9 In some cases, other combining marks will also combine with a base consonant, either attaching at a nonstandard location or changing shape. In minimal rendering there are only two cases, RA_l with U_{vs} or UU_{vs} .

RA_l	+	U_{vs}	RU_n
र	+	◌ु	रु
RA_l	+	UU_{vs}	RUU_n
र	+	◌ू	रू

Memory Representation and Rendering Order

The order for storage of plain text in Devanagari and all other Indic scripts generally follows phonetic order; that is, a CV syllable with a dependent vowel is always encoded as a consonant letter C followed by a vowel sign V in the memory representation. This order is employed by the ISCII standard and corresponds with both the phonetic and keying order of textual data.

Rendering Order

Character Order				Glyphh Order
KA_n +	I_{vs}	⟶	I_{vs} +	KA_n
क +	◌ि	⟶	कि	

Because Devanagari and other lndic scripts have some dependent vowels that must be depicted to the left side of their consonant letter, the software that renders the Indic scripts must be able to reorder elements in mapping from the logical (character) store to the presentational (glyph) rendering. For example, if C_n denotes the nominal form of consonant C, and V_{vs} denotes a left-side dependent vowel sign form of vowel V, then a reordering of glyphs with respect to encoded characters occurs as just shown.

R10 When the dependent vowel I_{vs} is used to override the inherent vowel of a syllable, it is always written to the extreme left of the orthographic syllable. If the orthographic syllable contains a consonant cluster, then this vowel is always depicted to the left of that cluster. For example:

$TA_d + RA_d + I_{vs} \longrightarrow T.RA_n + I_{vs} \longrightarrow I_{vs} + T.RA_d$

त् + र् + ◌ि ⟶ त्र + ◌ि ⟶ त्रि

Sample Half-Forms : Dev.1 shows examples of half-consonant forms that are commonly used with the Devanagari script. These forms are glyphs, not characters. They may be encoded explicitly using ZERO WIDTH JOINER as shown; in normal conjunct formation, they may be used spontaneously to depict a dead consonant in combination with subsequent consonant forms.

Dev.1 : Sample Half Forms

क	◌्	ZWJ	क्
ख	◌्	ZWJ	ख्
ग	◌्	ZWJ	ग्

च	्	ZWJ	च
ज	्	ZWJ	ज
झ	्	ZWJ	झ
ञ	्	ZWJ	ञ
ण	्	ZWJ	ण
त	्	ZWJ	त
थ	्	ZWJ	थ
ध	्	ZWJ	ध
न	्	ZWJ	न
प	्	ZWJ	प
फ	्	ZWJ	फ
ब	्	ZWJ	ब
भ	्	ZWJ	भ
म	्	ZWJ	म
य	्	ZWJ	य
ल	्	ZWJ	ल
व	्	ZWJ	व
स	्	ZWJ	स
ष	्	ZWJ	ष
स	्	ZWJ	स
क्ष	्	ZWJ	क्ष
ज्ञ	्	ZWJ	ज्ञ
श्र	्	ZWJ	श्र

Sample Ligatures : Dev.2 shows examples of conjunct ligature forms that are commonly used with the Devanagari script. These forms are glyphs, not characters. Not every writing system that employs this script uses all of these forms; in particular, many of these forms are used only in writing Sanskrit texts. Furthermore, individual fonts may provide fewer or more ligature forms than are depicted here.

Dev.2 : Sample Ligatures

क	्	क	क्क
क	्	त	क्त
क	्	र	क्र
ङ्	्	क	ङ्क
ङ्	्	ख	ङ्ख
ङ्	्	ग	ङ्ग
ङ्	्	घ	ङ्घ
ञ्	्	ज	ञ्ज
ज्	्	ञ	ज्ञ
द्	्	घ	द्ध
द्	्	द	द्द
द्	्	ध	द्ध
द्	्	ब	द्ब
द्	्	भ	द्भ
द्	्	म	द्म
द्	्	य	द्य
द्	्	व	द्व
ट्	्	ट	ट्ट
ट्	्	ठ	ट्ठ
ठ्	्	ठ	ठ्ठ
ड्	्	ग	
ड्	्	ड	ड्ड
ड्	्	ढ	ड्ढ
त्	्	त	त्त
त्	्	र	त्र
न्	्	न	न्न
फ्	्	र	फ्र
ह्	्	म	ह्म
ह्	्	य	ह्य
ह्	्	ल	ह्ल
ह्	्	व	ह्व
ह	ृ	ह	
र	ु	रु	

र ◌ू रू

स् त्र स्त्र

Sample Half - Ligature Forms : In addition to half-form glyphs of individual consonants, half-forms are also used to depict conjunct ligature forms. A sample of such forms is shown in Dev.3. These forms are glyphs, not characters. They may be encoded explicitly using ZERO WIDTH JOINER as shown; in normal conjunct formation, they may be used spontaneously to depict a conjunct ligature in combination with subsequent consonant forms.

Dev.3 : Sample Half-Ligature Forms

त ◌् त ◌् त्र

त ◌् र ◌् त्र

Combining Marks : Devanagari and other Indic scripts have a number of combining marks that could be considered diacritic. One class of these marks, known as bindus, is represented by U+0901 DEVANAGARI SIGN CANDRABINDU and U+0902 DEVANAGARI SIGN ANUSVARA. The first mark indicates nasalization of a vowel and the second mark represent a nasal consonant occurring after a vowel or final nasal closure of a syllable. U+093C DEVANAGARI SIGN NUKTA is a true diacritic. It is used to extend the basic set of consonant letters by modifying them (with a subscript dot in Devanagari) to create new letters. U+0951..U+0957 are a set of combining marks used in transcription of Sanskrit texts.

Digits : Each Indic script has a distinct set of digits appropriate to that script. These digits may or may not be used in ordinary text in that script. The international form of Indian Digits (Hindsa) have displaced the Indic script forms in modern usage in many of the scripts. Some Indic scripts-notably Tamil-lack a distinct digit for zero.

Punctuation and Symbols : U+0964 DEVANAGARI PURNA VIRAM is similar to a full stop. Corresponding forms occur in many other Indic scripts. U+0965 DEVANAGARI DEERGH VIRAM marks the end of a verse in traditional texts.

Many modern languages written in the Devanagari script intersperse punctuation derived from the Latin script. Thus U+002C COMMA and U+00E FULL STOP are freely used in writing Hindi, and the 'PURNA VIRAM (danda) is usually restricted to more traditional texts.

Encoding Structure : The Unicode Standard organizes the nine principal Indic scripts in blocks of 128 encoding points each. The first six columns in each script are isomorphic with the ISCII-1988 encoding, except that the last 11 positions (U+0955 .. U+095F in Devanagari, for example), which are unassigned or undefined in ISCII-1988, are used in the Unicode encoding.

The seventh column in each of these scripts, along with the last 11 positions in the sixth column, represent additional character assignments in the Unicode Standard that are matched across all nine scripts. For example, positions U+xx66 ... U+xx6F and U+xxE6 ... U+xxEF code the Indic script digits for each script.

The eighth column for each script is reserved for script-specific additions that do not correspond from one Indic script to the next.

(The above revision is based on detailed discussions with National & State level Institutions/Directorates dealing with Devanagari based languages - Sanskrit, Hindi, Marathi, Nepali, Konkani & Sindhi)

Contact : mjain@mit.gov.in
tdilinfo@mit.gov.in