Appendix A: Character Shaping Behavior

Characters from the Arabic and Indic 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 such characters to be different from their nominal glyphs (such as those shown in the Unicode charts).

In the Indic scripts, the presence of certain characters causes a change in the display order of other characters. This reordering is not commonly seen in non-Indic scripts. It also occurs independently of any bidirectional character reordering that might be in effect.

The following section describes minimal requirements for legibly rendering Arabic, Devanagari, and Tamil scripts as part of a plain text sequence. It describes the mapping between Unicode values and the glyphs in a font used to represent these characters. This section also describes the combining and ordering of those glyphs. The reader should be familiar with the General Principles of the Unicode Standard (Volume 1, Chapter 2), and especially with the distinction between characters and glyphs.

The algorithms provide minimal requirements for legibly rendering interchanged Unicode text with characters from these scripts. As with any script, a more complex algorithm can add rendering characteristics, depending on the font and application.

It is important to emphasize that in a font that is capable of rendering Arabic and Indic scripts, the set of glyphs is much greater than the number of Unicode characters. If a particular form is not found in the Unicode charts, implementors should be aware that it may be a different form of a Unicode character, or the result of a ligature of two or more Unicode characters.

For brevity, the Indic scripts are represented in some detail with Devanagari and Tamil, since they illustrate the range of types of behavior found in the other Indic scripts. The behavior of the other scripts can be derived from a close examination of these examples.

Arabic Character Shaping

Arabic is always written in "cursive" or "handwritten" form, where the characters link together as they are written. All cursive scripts require linking rules for proper rendering. The linking rules for Arabic may be easier to understand if compared to Latin cursive handwriting.

There are also many optional ligatures which may be found in different fonts, as well as special forms of characters such as the *snake-kaf* or *extended noon* which may be used for justification or æsthetic reasons.

The minimal shaping algorithm is described in terms of the display (visual) order. In other words, the position of characters in the included examples is presented as it would appear on the screen *after* the bidirectional algorithm has determined character order in a line of text. Wherever the rules do not specify a change in form, characters are presented with their nominal forms, as they are presented in the code charts.

Linking Classes

For brevity, the words ARABIC LETTER have been omitted from the names of all Arabic characters. For the purposes of describing Arabic linking, Unicode characters fall into the following classes. A complete list follows at the end of this section.

Dual-linking:	such as BAA, TAA, THAA, JEEM...
Right-linking:	ALEF, DAL, THAL, RA, ZAIN...
Link-causing:	ZERO WIDTH JOINER, TATWEEL (*kashida*)
Non-linking:	ZERO WIDTH NON-JOINER, and all spacing characters (other than those above) are non-linking, including:HAMZAH, HIGH HAMZA, spaces, digits, punctuation, non-Arabic letters, and so on.
Transparent:	All combining marks and format marks are transparent, including: FATHATAN, DAMMATAN, FATHAH, DAMMAH, KASRAH, SHADDAH, SUKUN, ALEF ABOVE, RIGHT-LEFT MARK and so on.

For convenience in the linking rules, two supersets of some of the preceding classes are defined as follows: A *right link-causing character* is either a right-linking, dual-linking or link-causing character. A *left link-causing character* is either a left-linking, dual-linking or link-causing character.

Notation

If X refers to a character, then various glyph types representing that character are referred to by adding a subscript:

X_n	Nominal form of a character, as it appears in the code charts (subject to stylistic variation). Unless otherwise stated, characters will adopt a nominal form.
X_r	Form of the character which joins to the right (right-linking and dual-linking characters may have this form).
X_l	Form of the character which joins to the left (left-linking and dual-linking characters may have this form).
X_m	Form of the character which joins to both the right and the left (dual-linking characters may have this form).
$X.Y_n$	Ligature form representing a combination of a form of X_n and a form of Y_n.

Cursive Linking Rules

J1. *Transparent glyphs do not affect the linking behavior of base (spacing) characters. For example:*

$$\text{MEEM}_n + \text{SHADDAH}_n + \text{LAM}_n \rightarrow \text{MEEM}_r + \text{SHADDA}_n + \text{LAM}_l$$

م + ◌ّ + ل → مـ + ◌ّ + ـل → لّم

J2. *A right-linking glyph X_n which has a right link-causing glyph on the right will adopt the form X_r. For example:*

$$\text{ALEF}_n + \text{TATWEEL}_n \rightarrow \text{ALEF}_r + \text{TATWEEL}_n$$

ا + ـ → ـا + ـ → ـا

J3. *A left-linking glyph X_n which has a left link-causing glyph on the left will adopt the form X_l.*

J4. *A dual-linking glyph X_n which has a right link-causing glyph on the right and a left link-causing glyph on the left will adopt the form X_m. For example:*

$$\text{TATWEEL}_n + \text{MEEM}_n + \text{TATWEEL}_n \rightarrow \text{TATWEEL}_n + \text{MEEM}_m + \text{TATWEEL}_n$$

ـ + م + ـ → ـ + مـ + ـ → ـمـ

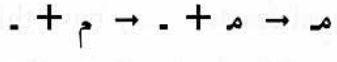
J5. *A dual-linking glyph X_n which has a right link-causing glyph on the right and no left link-causing glyph on the left will adopt the form X_r. For example:*

$$\text{MEEM}_n + \text{TATWEEL}_n \rightarrow \text{MEEM}_r + \text{TATWEEL}_n$$

م + ـ → مـ + ـ → مـ

J6. *A dual-linking glyph which has a left link-causing glyph on the left and no right-link causing glyph on the right will adopt the form X_l. For example:*

$$\text{TATWEEL}_n + \text{MEEM}_n \rightarrow \text{TATWEEL}_n + \text{MEEM}_l$$



As noted in Section 3.1 of Volume 1, "Arabic," the ZERO WIDTH NON-JOINER can be used to prevent linking, as in the Persian plural suffix or Ottoman Turkish vowels.

Ligature Classes

A certain class of ligatures are obligatory regardless of font design: they must be used in well-rendered Arabic text. There are many other ligatures that are optional, depending on font design. Since they are optional, those ligatures are not covered by the Unicode standard.

For the purpose of describing the obligatory Arabic ligatures, Unicode characters fall into the following classes (for a complete list, see the end of this section):

Alef-types	MADDAH ON ALEF, HAMZAH ON ALEF...
Lam-types	LAM, LAM WITH SMALL V, LAM WITH DOT ABOVE...

Obligatory Ligature Rules

The following rules describe the formation of ligatures. They are applied after the preceding linking rules.

L1. *Transparent glyphs do not affect the ligating behavior of base glyph. For example:*

$$\text{ALEF}_r + \text{FATHAH}_n + \text{LAM}_l \rightarrow \text{LAM.ALEF}_n + \text{FATHAH}_n$$

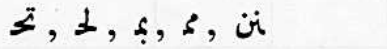
(However, placement of a combining mark on a resulting ligature will depend on which of the original base characters the mark modifies).

L2. *Any sequence with ALEF_r on the left and LAM_m on the right will form the ligature LAM.ALEF_l.*

L3. *Any sequence with ALEF_r on the left and LAM_l on the right will form the ligature LAM.ALEF_n.*

Optional Features

There are many other ligatures and contextual forms that are optional—depending on the font and application—such as the following:



In addition, the context-sensitive placement of non-spacing vowels such as FATHA can greatly improve the appearance of the text.

Character Types

The following is a detailed list of the Arabic characters that are either right-linking or dual-linking. All other Arabic characters (aside from TATWEEL) are non-linking, including ARABIC LETTER AE. For brevity in the names, the words TWO, THREE and FOUR are abbreviated.

Most of the extended Arabic characters are merely variations on the basic Arabic shapes, with additional or different marks. For compatibility, many precomposed forms are included.

In some cases there are characters that only occur at the end of words in correct spelling, called *trailing characters.* Examples include TEH MARBUTA, ALEF MAQSURAH and DAMMATAN. When trailing characters are joining (such as TEH MARBUTA), they are classed as right-joining (even when similarly-shaped characters are dual-joining). When trailing characters do not join or cause joining (such as DAMMATAN), they are classed as transparent, which treats all combining marks similarly.

Note: In the case of HA, HA *i* is also shown in the code chart box. This is often done to reduce the chance of misidentifying HA as ARABIC INDIC DIGIT 5, which has a very similar shape. The nominal form of HA is the isolate form, which looks like ARABIC LETTER AE. In the case of HA GOAL, the nominal form is not even listed, and also resembles ARABIC LETTER AE.

The characters in the following charts are grouped by shape, and not by standard Arabic alphabetical order.

Dual-linking

Group	CHAR *n*	CHAR *r*	CHAR *m*	CHAR *l*	Characters with Similar Shaping Behavior
BAA	ب	ب	بـ	بـ	TAA, THAA, TAA WITH SMALL TAH, TAA WITH 2 DOTS VERTICAL ABOVE, BAA WITH 2 DOTS VERTICAL BELOW, TAA WITH RING, TAA WITH 3 DOTS ABOVE DOWNWARD, TAA WITH 3 DOTS BELOW, TAA WITH 4 DOTS ABOVE, BAA WITH 4 DOTS BELOW
HAA	ح	ح	حـ	حـ	JEEM, KHAA, HAMZAH ON HAA, HAA WITH 2 DOTS VERTICAL ABOVE, HAA WITH MIDDLE 2 DOTS, HAA WITH MIDDLE 2 DOTS VERTICA. HAA WITH 3 DOTS ABOVE, HAA WITH MIDDLE 3 DOTS DOWNWARD, HAA WITH MIDDLE 4 DOTS
SEEN	س	س	سـ	سـ	SHEEN, SEEN WITH DOT BELOW AND DOT ABOVE, SEEN WITH 3 DOTS BELOW, SEEN WITH 3 DOTS BELOW AND 3 DOTS ABOVE
SAD	ص	ص	صـ	صـ	DAD, SAD WITH 2 DOTS BELOW, SAD WITH 3 DOTS ABOVE
TAH	ط	ط	طـ	طـ	DHAH, TAH WITH 3 DOTS ABOVE
AIN	ع	ع	عـ	عـ	GHAIN, AIN WITH 3 DOTS ABOVE
FA	ف	ف	فـ	فـ	DOTLESS FA, FA WITH DOT MOVED BELOW, FA WITH DOT BELOW, FA WITH 3 DOTS ABOVE, FA WITH 3 DOTS BELOW, FA WITH 4 DOTS ABOVE
QAF	ق	ق	قـ	قـ	QAF WITH DOT ABOVE, QAF WITH 3 DOTS ABOVE
CAF	ك	ك	كـ	كـ	CAF WITH DOT ABOVE, CAF WITH 3 DOTS ABOVE, CAF WITH 3 DOTS BELOW
LAM	ل	ل	لـ	لـ	LAM WITH SMALL V, LAM WITH DOT ABOVE, LAM WITH 3 DOTS ABOVE
MEEM	م	م	مـ	مـ	
NOON	ن	ن	نـ	نـ	DOTLESS NOON, DOTLESS NOON WITH SMALL TAH, NOON WITH RING, NOON WITH 3 DOTS ABOVE
HA	ه	ه	هـ|ھـ	هـ	

Dual-linking (continued)

YA	ي	ي	ۓ	ۑ	HIGH HAMZAH YA, HAMZAH ON YA, DOTLESS YEH, YA WITH SMALL V, YA WITH 2 DOTS VERTICAL BELOW, YA WITH 3 DOTS BELOW
SWASH CAF	ڪ	ڪ	ڪ	ڪ	
GAF	گ	گ	گ	گ	OPEN CAF, CAF WITH RING, GAF WITH RING, GAF WITH 2 DOTS ABOVE, GAF WITH 2 DOTS BELOW, GAF WITH 2 DOTS VERTICAL BELOW, GAF WITH 3 DOTS ABOVE
KNOTTED HA	ھ	ھ	ھ	ھ	
HA GOAL	ہ	ہ	ۂ	ۃ	HAMZAH ON HA GOAL

Right-linking

Group	*CHAR$_n$*	*CHAR$_r$*	*Characters with Similar Shaping Behavior*
ALEF	ا	ـا	ALEF, HAMZAH ON ALEF, MADDAH ON ALF, HAMZAH UNDER ALEF, WAVY HAMZAH ON ALEF, HIGH HAMZAH ALEF
DAL	د	ـد	THAL, DAL WITH SMALL TAH, DALL WITH RING, DALL WITH DOT BELOW, DAL WITH DOT BELOW AND SMALL TAH, DAL WITH 2 DOTS, DAL WITH 2 DOTS BELOW, DALL WITH 3 DOTS ABOVE, DALL WITH 3 DOTS ABOVE DOWNWARD, DAL WITH 4 DOTS ABOVE
RA	ر	ـر	ZAIN, RA WITH SMALL TAH, RA WITH SMALL V, RA WITH RING, RA WITH DOT BELOW, RA WITH SAMM V BELOW, RA WITH DOT BELOW AND DOT ABOVE, RA WITH 2 DOTS ABOVE, RA WITH 3 DOTS ABOVE, RA WITH 4 DOTS ABOVE
WAW	و	ـو	HAMZAH ON WAW, HIGH HAMZAH WAW, HIGH HANZAH WAW WITH DAMMAH, WAW WITH RING, WAW WITH BAR, WAW WITH SMALL V, WAW WITH DAMMAH, WAW WITH ALEF ABOVE, WAW WITH INVERTED SMALL V, WAW WITH 2 DOTS ABOVE, WAW WITH 3 DOTS ABOVE
ALEF MAQSURAH	ى	ـى	YA WITH TAIL
TAA MARBUTAH	ة	ـة	
TAA MARBUTAH GOAL	ۃ	ـۃ	HAMZAH ON HA
YA BAREE	ے	ـے	HAMZAH ON YA BARREE

Other Classes

Link-causing	ZERO WIDTH JOINER, TATWEEL
Non-linking	ZERO WIDTH NON-JOINER, and all spacing characters (other than those mentioned above) are non-linking, including: HAMZAH, HIGH HAMZAH, HAMZAT WASL ON ALEF, spaces, digits, punctuation, non-Arabic letters and so on.
Transparent	All combining marks and format marks are transparent, including: FATHATAN, DAMMATAN, FATHAH, DAMMAH, KASRAH, SHADDAH, SUKUN, ALEF ABOVE, RIGHT-LEFT MARK and so on.

Index by Unicode Value

Unic.	*Link*	*Link Group*	*Name*
0622	R	ALEF	MADDAH ON ALEF
0623	R	ALEF	HAMZAH ON ALEF
0624	R	WAW	HAMZAH ON WAW
0625	R	ALEF	HAMZAH UNDER ALEF
0626	D	YA	HAMZAH ON YA
0627	R	ALEF	ALEF
0628	D	BAA	BAA
0629	R	TAA MARBUTAH	TAA MARBUTAH
062A	D	BAA	TAA
062B	D	BAA	THAA
062C	D	HAA	JEEM
062D	D	HAA	HAA
062E	D	HAA	KHAA
062F	R	DAL	DAL
0630	R	DAL	THAL
0631	R	RA	RA
0632	R	RA	ZAIN
0633	D	SEEN	SEEN
0634	D	SEEN	SHEEN
0635	D	SAD	SAD
0636	D	SAD	DAD
0637	D	TAH	TAH
0638	D	TAH	DHAH
0639	D	AIN	AIN
063A	D	AIN	GHAIN
0640	C	<no shaping>	TATWEEL
0641	D	FA	FA
0642	D	QAF	QAF
0643	D	CAF	CAF
0644	D	LAM	LAM
0645	D	MEEM	MEEM
0646	D	NOON	NOON
0647	D	HA	HA
0648	R	WAW	WAW
0649	R	ALEF MAQSURAH	ALEF MAQSURAH
064A	D	YA	YA
0671	U	<no shaping>	HAMZAT WASL ON ALEF
0672	R	ALEF	WAVY HAMZAH ON ALEF
0673	R	ALEF	WAVY HAMZAH UNDER ALEF
0674	U	<no shaping>	HIGH HAMZAH
0675	R	ALEF	HIGH HAMZAH ALEF
0676	R	WAW	HIGH HAMZAH WAW
0677	R	WAW	HIGH HAMZAH WAW WITH DAMMAH
0678	D	YA	HIGH HAMZAH YA
0679	D	BAA	TAA WITH SMALL TAH
067A	D	BAA	TAA WITH 2 DOTS VERTICAL ABOVE
067B	D	BAA	BAA WITH 2 DOTS VERTICAL BELOW

067C	D	BAA	TAA WITH RING
067D	D	BAA	TAA WITH 3 DOTS ABOVE DOWNWARD
067E	D	BAA	TAA WITH 3 DOTS BELOW
067F	D	BAA	TAA WITH 4 DOTS ABOVE
0680	D	BAA	BAA WITH 4 DOTS BELOW
0681	D	HAA	HAMZAH ON HAA
0682	D	HAA	HAA WITH 2 DOTS VERTICAL ABOVE
0683	D	HAA	HAA WITH MIDDLE 2 DOTS
0684	D	HAA	HAA WITH MIDDLE 2 DOTS VERTICAL
0685	D	HAA	HAA WITH 3 DOTS ABOVE
0686	D	HAA	HAA WITH MIDDLE 3 DOTS DOWNWARD
0687	D	HAA	HAA WITH MIDDLE 4 DOTS
0688	R	DAL	DAL WITH SMALL TAH
0689	R	DAL	DAL WITH RING
068A	R	DAL	DAL WITH DOT BELOW
068B	R	DAL	DAL WITH DOT BELOW AND SMALL TAH
068C	R	DAL	DAL WITH 2 DOTS ABOVE
068D	R	DAL	DAL WITH 2 DOTS BELOW
068E	R	DAL	DAL WITH 3 DOTS ABOVE
068F	R	DAL	DAL WITH 3 DOTS ABOVE DOWNWARD
0690	R	DAL	DAL WITH 4 DOTS ABOVE
0691	R	RA	RA WITH SMALL TAH
0692	R	RA	RA WITH SMALL V
0693	R	RA	RA WITH RING
0694	R	RA	RA WITH DOT BELOW
0695	R	RA	RA WITH SMALL V BELOW
0696	R	RA	RA WITH DOT BELOW AND DOT ABOVE
0697	R	RA	RA WITH 2 DOTS ABOVE
0698	R	RA	RA WITH 3 DOTS ABOVE
0699	R	RA	RA WITH 4 DOTS ABOVE
069A	D	SEEN	SEEN WITH DOT BELOW AND DOT ABOVE
069B	D	SEEN	SEEN WITH 3 DOTS BELOW
069C	D	SEEN	SEEN WITH 3 DOTS BELOW AND 3 DOTS ABOVE
069D	D	SAD	SAD WITH 2 DOTS BELOW
069E	D	SAD	SAD WITH 3 DOTS ABOVE
069F	D	TAH	TAH WITH 3 DOTS ABOVE
06A0	D	AIN	AIN WITH 3 DOTS ABOVE
06A1	D	FA	DOTLESS FA
06A2	D	FA	FA WITH DOT MOVED BELOW
06A3	D	FA	FA WITH DOT BELOW
06A4	D	FA	FA WITH 3 DOTS ABOVE
06A5	D	FA	FA WITH 3 DOTS BELOW
06A6	D	FA	FA WITH 4 DOTS ABOVE
06A7	D	QAF	QAF WITH DOT ABOVE
06A8	D	QAF	QAF WITH 3 DOTS ABOVE
06A9	D	GAF	OPEN CAF
06AA	D	SWASH CAF	SWASH CAF
06AB	D	GAF	CAF WITH RING
06AC	D	CAF	CAF WITH DOT ABOVE
06AD	D	CAF	CAF WITH 3 DOTS ABOVE

06AE	D	CAF	CAF WITH 3 DOTS BELOW
06AF	D	GAF	GAF
06B0	D	GAF	GAF WITH RING
06B1	D	GAF	GAF WITH 2 DOTS ABOVE
06B2	D	GAF	GAF WITH 2 DOTS BELOW
06B3	D	GAF	GAF WITH 2 DOTS VERTICAL BELOW
06B4	D	GAF	GAF WITH 3 DOTS ABOVE
06B5	D	LAM	LAM WITH SMALL V
06B6	D	LAM	LAM WITH DOT ABOVE
06B7	D	LAM	LAM WITH 3 DOTS ABOVE
06BA	D	NOON	DOTLESS NOON
06BB	D	NOON	DOTLESS NOON WITH SMALL TAH
06BC	D	NOON	NOON WITH RING
06BD	D	NOON	NOON WITH 3 DOTS ABOVE
06BE	D	KNOTTED HA	KNOTTED HA
06C0	R	TAA MARBUTAH	HAMZAH ON HA
06C1	D	HA GOAL	HA GOAL
06C2	R	HAMZAH ON HA GOAL	HAMZAH ON HA GOAL
06C3	R	HAMZAH ON HA GOAL	TAA MARBUTAH GOAL
06C4	R	WAW	WAW WITH RING
06C5	R	WAW	WAW WITH BAR
06C6	R	WAW	WAW WITH SMALL V
06C7	R	WAW	WAW WITH DAMMAH
06C8	R	WAW	WAW WITH ALEF ABOVE
06C9	R	WAW	WAW WITH INVERTED SMALL V
06CA	R	WAW	WAW WITH 2 DOTS ABOVE
06CB	R	WAW	WAW WITH 3 DOTS ABOVE
06CC	D	YA	DOTLESS YA
06CD	R	ALEF MAQSURAH	YA WITH TAIL
06CE	D	YA	YA WITH SMALL V
06D0	D	YA	YA WITH 2 DOTS VERTICAL BELOW
06D1	D	YA	YA WITH 3 DOTS BELOW
06D2	R	YA BARREE	YA BARREE
06D3	R	YA BARREE	HAMZAH ON YA BARREE
06D5	U	<no shaping>	AE

Devanagari Character Shaping

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 be different from their nominal glyphs (such as those shown in the Unicode 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.

The following algorithm describes minimal rendering of Devanagari as part of a plain text sequence. It describes the mapping between Unicode values and the glyphs in a Devanagari font. It also describes the combining and ordering of those glyphs.

This algorithm provides minimal requirements for legibly rendering interchanged Unicode Devanagari text. As with any script, a more complex algorithm 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 Unicode Devanagari characters.

Notation

For brevity, the words "DEVANAGARI LETTER" and "DEVANAGARI VOWEL SIGN" are omitted from the names of Unicode characters in this section where such omission is not ambiguous. If X refers to a character, then various glyph types representing that character are referred to by adding a subscript.

X_m	Non-spacing mark form.
X_l	Left side form.
X_n	Nominal form of a character, as it appears in the code charts (subject to stylistic variation). Unless otherwise stated, characters will adopt a nominal form.
X_d	Form of a consonant representing a "dead" form, where the implicit vowel is not pronounced.
X_h	Form of a dead consonant where the *virama* is absorbed, and a different shape is taken (see below). This is also known as a "half consonant."
$X.Y_n$	Ligature of a dead-consonant XA_n with a consonant Y_n. Such ligatures are also called *conjunct consonants.*

Vowel signs are indicated by superscripting with^{vs}.

X^{vs}	Vowel sign (not to be confused with the independent vowels). AA is not the same as AA^{vs}.

Character Classes

The principal classes of Devanagari characters are the following (a complete list follows at the end of this section):

Consonants: KA, KHA, GA, GHA...

Vowels: A, AA...

Vowel Signs: AA^{vs}, I^{vs}...

Background

Independent vowels can be followed by various members of the class of non-spacing marks. The marks are applied to the vowel in a manner similar to marks on Latin letters. Independent vowels generally have no other interaction with other characters.

Other well-formed Devanagari is organized as a sequence of syllable clusters. Phonetically, a syllable cluster is a sequence of consonants (usually one or two, occasionally three, rarely four), followed by a vowel, with perhaps some modification of the vowel (such as nasalization).

Other than the application of non-spacing marks, digits and free-standing signs have no interaction with other characters.

The *virama* indicates a null vowel. As such, it is used to join together the sequence of consonants in a syllable cluster. Additionally, it may be used at the end of the cluster if the cluster's vowel is null.

Dead Consonants

A dead consonant is defined as a sequence consisting of a consonant followed by a VIRAMA$_n$.

The minimal rendering for a dead consonant is to position the *virama* as a non-spacing mark bound to the character. (See, however, Optional Forms in the following section.)

For example:

TA$_n$ + VIRAMA$_n$ → TA$_d$

त + ◌् → त्

Consonant Clusters

A *consonant cluster* is defined as a consonant or independent vowel, optionally preceded by a sequence of one or more dead consonants, and optionally followed by a vowel sign, as in the following rules.

(Note that for the purposes of rendering, the independent vowels behave very much like consonants, even though they are rarely used this way. For example, ॠ or ઍ can be produced by an application of rules discussed below. See "Independent Vowels" for more information.)

For example:

त त + ◌् + क त + ◌् + क + ◌् + ष

C1. *When RA_n occurs to the right of a $VIRAMA_n$, then it combines with the $VIRAMA_n$ to form a low non-spacing mark (RA_m), which attaches to the consonant on the left.*

$$VIRAMA_n + RA_n \rightarrow RA_m$$

◌् + र → ◌्र

For example:

$$THA_n + VIRAMA_n + RA_n \rightarrow TTA_n + RA_m$$

ठ + ◌् + र → ठ्र

C2. *If the non-spacing RA_m is attached to a consonant having a vertical bar on its right, then it loses its right side, and is positioned attached to the vertical bar:*

$$PHA_n + RA_l \rightarrow PHA_n + RA_l$$

फ + ◌् → फ्र

C3. *The half consonant form of RA_n behaves as a non-spacing mark when followed by a consonant. The vowel sign is then reordered (as in rule O1 in a following section). The ordering of these rules allows the production of ligatures such as श्र्य.*

$$RA_n + VIRAMA_n \rightarrow RA_h$$

र + ◌् → ◌ं

C4. *There are three forms that produce significantly different shapes when ligating. These forms are mandatory ligatures:*

$KA_d + SSHA_n \rightarrow K.SSHA_n$

क् + ष → क्ष

$JA_d + NYA_n \rightarrow J.NYA_n$

ज् + ञ → ज्ञ

$TA_d + RA_n \rightarrow T.RA_n$

त् + र → त्र

C5. *Non-spacing marks modifying a consonant are placed right after that consonant in the backing store, and are attached to that consonant in rendering. If a* virama *is included, it should be placed after the consonant-modifying non-spacing marks in the backing store.*

$KA_n + NUKTA_n + AA_n$

क + ़ + ा → क़ा

C6. *Non-spacing marks modifying a cluster are placed right after the final element in the cluster in the backing store, and are attached to the rightmost free-standing element of the cluster. Note that the* bindus *follow vowel signs, but precede* svaras *in the backing store. The relative placement of these non-spacing marks is horizontal rather than vertical: the horizontal rendering order may vary according to typographical concerns.*

$KA_n + AA_n + CANDRABINDU_n$

क + ा + ँ → काँ

Optional Forms

In most cases, the following ligatures are the preferred forms, although the unligated form can be used and understood. There are generally two shape changes that happen in forming ligatures of consonant clusters.

L1. *If a preceding consonant contains a vertical bar on the right, then that bar is removed, the* virama *disappears, and the characters are joined:*

$GA_d + GHA_n \rightarrow G.GHA_n$

ग् + घ → ग्घ

In implementation, this form can be rendered either by making a single composed ligature glyph to represent the cluster, or just by replacing the dead consonant by a special "half-consonant" form, such as:

$GA_d \rightarrow GA_h$

ग् → ग

$NA_d \rightarrow NA_h$

न् → न

The choice of whether to use the half-consonant forms or to use a composite glyph is implementation-dependent. Depending on the characters involved, both can produce acceptable display. The choice depends on the extent to which the two consonants undergo further shape modifications in order to link properly.

Character shaping can be explicitly controlled by using the ZERO WIDTH NON-JOINER and ZERO WIDTH JOINER. As described in Volume 1, the ZERO WIDTH NON-JOINER and ZERO WIDTH JOINER can be used to request that the *virama* be explicitly rendered to prevent any automatic formation of a ligature or half-consonant. The ZERO WIDTH JOINER can be used to request that the virama be absorbed into the half consonant form, and prevent any further automatic formation of a ligature.

For example:

$KA_n + VIRAMA_n + SSHA_n \rightarrow K.SSHA_n$

क + ◌् + ष → क्ष

$KA_n + VIRAMA_n + ZWNJ + SSHA_n \rightarrow KA_d + SSHA_n$

क + ◌् + ZWNJ + ष → क्‌ष

$KA_n + ZWJ + VIRAMA_n + SSHA_n \rightarrow KA_h + SSHA_n$

क + ZWJ + ◌् + ष → क्ष

L2. *Where the vertical bar does not exist or cannot be detached, or where the distinctive portion of the following character does not provide a linkage point, the letters are joined vertically, and then generally reduced in size.*

$TTA_d + TTHA_n \rightarrow TT.TTHA_n$

ट् + ठ → ट्ठ

L3. *In some cases, non-spacing marks will also combine with a base consonant, either attaching at a non-standard location, or changing shape. In minimal rendering there are only two cases, RA_n with $U_{vs}{}^n$ and $UU_{vs}{}^n$.*

$$RA_n + UU_n \rightarrow RUU_n$$

र + ◌ू → रू

L4. *Whenever a ligature is formed, it can then combine with other glyphs, ligatures, and half-forms in an even more complicated ligature:*

स + ◌् + त + ◌् + र →

स + ◌् + त्र →

स्त्र

Character Ordering

There are cases in Devanagari where characters are rendered in a different order from the backing store. The reordering rules affect two characters.

O1. *Cluster + I_n → I_n + Cluster*

Reordering depends upon whether consonants within the cluster ligate. For example:

$$GA_d + GHA_n + I_n \rightarrow I_n + G.GHA_n$$

ग् + घ + ◌ि → गि्घ

$$G.GHA_n + I_n \rightarrow I_n + G.GHA_n$$

ग्घ + ◌ि → गि्घ

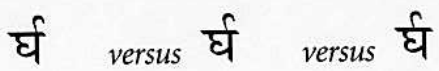
O2. *RA_h + Cluster → Cluster + RA_d*

Example:

$$RA_h + G.GHA_n \rightarrow G.GHA_n + RA_d$$

◌ं + ग्घ → र्ग्घ

Once conjunct consonants have been formed and reordering has taken place, more advanced contextual modifications of the character shapes can be used to create a better presentation. For example, the hook in the I^{vs} can be extended to intersect the last vertical bar in the following (after reordering) consonant cluster, or the RA_d can be aligned with the vertical bar.

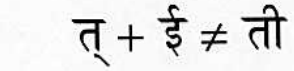


These modifications are optional, but produce a better appearance.

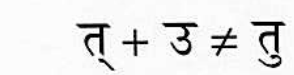
Independent Vowels

There is a one-to-one correspondence between the independent vowels and the dependent vowel signs. Except for A, independent vowels are sometimes represented by a form of the dependent vowels. This representation is sometimes found in schoolbooks. Independent forms can be represented by a sequence of the A followed by the dependent vowel. *This does not imply that the corresponding independent and dependent vowels are equivalent backing stores: they are two separate (but related) backing stores and renderings.*

TA + VIRAMA + II≠ TA + II

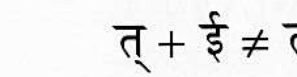


TA + VIRAMA + U≠ TA + U

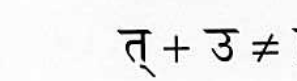


The dependent vowels should not be represented by a combination of *virama* plus independent vowel in interchange:

TA + VIRAMA + II TA + II



TA + VIRAMA + U≠ TA + U



Character Classes

The following character types are found in written Devanagari:

Digits

0966 DIGIT ZERO
...
096F DIGIT NINE

Various Free-Standing Signs

0903 SIGN VISARGA
093D SIGN AVAGRAHA
0950 OM
0964 DANDA
0965 DOUBLE DANDA
0970 ABBREVIATION SIGN

Independent Vowels

0905 A
...
0914 AU
0960 VOCALIC RR
0961 VOCALIC LL

Cluster-Modifying Non-Spacing Marks

Bindus

0901 SIGN CANDRABINDU
0902 SIGN ANUSVARA

Svaras

0951 STRESS SIGN UDATTA
0952 STRESS SIGN ANUDATTA
0953 GRAVE ACCENT
0954 ACUTE ACCENT

Consonant-Modifying Non-Spacing Marks

093C SIGN NUKTA

Diacritical Vowel Signs

0941 VOWEL SIGN U
...
0948 VOWEL SIGN AI
0962 VOWEL SIGN VOCALIC L
0963 VOWEL SIGN VOCALIC LL

Free-Standing Vowel Signs

093E VOWEL SIGN AA
0940 VOWEL SIGN II

094A VOWEL SIGN SHORT O
094B VOWEL SIGN O
094C VOWEL SIGN AU

Preceding Type

093F VOWEL SIGN I

Virama

094D VIRAMA

Sample Ligatures

The following charts show examples of ligatures that can be used in Devanagari. These ligatures are glyphs, not characters. (See L4 in the previous section.)

क	्	क	क्क
क	्	त	क्त
क	्	र	क्र
क	्	ष	क्ष
ङ	्	क	ङ्क
ङ	्	ख	ङ्ख
ङ	्	ग	ङ्ग
ङ	्	घ	ङ्घ
ञ	्	ज	ञ्ज
ज	्	ञ	ज्ञ
द	्	घ	द्ध
द	्	द	द्द
द	्	ध	द्ध
द	्	ब	द्ब
द	्	भ	द्भ
द	्	म	द्म
द	्	य	द्य
द	्	व	द्व
ट	्	ट	ट्ट

ट	्	ठ	ट्ठ
ठ	्	ठ	ठ्ठ
ड	्	ग	ड्ग
ड	्	ड	ड्ड
ड	्	ढ	ड्ढ
त	्	त	त्त
त	्	र	त्र
न	्	न	न्न
फ	्	र	फ्र
श	्	र	श्र
ह	्	म	ह्म
ह	्	य	ह्य
ह	्	ल	ह्ल
ह	्	व	ह्व
ह		ृ	ह्र
र		ु	रु
र		ू	रू
स	्	त्र	स्त्र

Sample Half-Forms

क	्	क्	न	्	न्
ख	्	ख्	प	्	प्
ग	्	ग्	फ	्	फ्
घ	्	घ्	ब	्	ब्
च	्	च्	भ	्	भ्
ज	्	ज्	म	्	म्
झ	्	झ्	य	्	य्
ञ	्	ञ्	ल	्	ल्
ण	्	ण्	व	्	व्
त	्	त्	श	्	श्
थ	्	थ्	ष	्	ष्
ध	्	ध्	स	्	स्

As explained in the text, ligatures may also have half-forms. The following shows examples of this:

Ligature Half-Forms

क	्	ष	्	क्ष
ज	्	ञ	्	ज्ञ
त	्	त	्	त्त
त	्	र	्	त्र
श	्	र	्	श्र

Tamil Shaping Behavior

The South Indic scripts function in much the same way as Devanagari, with the additional feature of two-part vowels. The example of Tamil will be used to illustrate this feature. The following is a summary of the Tamil letters. As in the Devanagari example, the words "TAMIL LETTER" and "TAMIL VOWEL SIGN" will be omitted where this does not cause ambiguity. The latter are indicated by superscripting with *vs*.

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

க KA	ங NGA	ச CA	ஜ JA	ஞ NYA	ட TTA	ண NNA	த TA	ந NA	ன NNNA	ப PA	
ம MA	ய YA	ர RA	ற RRA	ல LA	ள LLA	ழ LLLA	வ VA	ஷ SSA	ஸ SA	ஹ HA	
அ A	ஆ AA	இ I	ஈ II	உ U	ஊ UU	எ E	ஏ EE	ஐ AI	ஒ O	ஓ OO	ஔ AU
A	◌ா AA	◌ி I	◌ீ II	◌ு U	◌ூ UU	◌ெ E	◌ே EE	◌ை AI	◌ொ O	◌ோ OO	◌ௌ AU
◌் VIRAMA	◌ௗ AU LENGTH										

Independent versus Dependent Vowels

As with Devanagari, the dependent vowel signs are not equivalent to a sequence of *virama* + independent vowel. For example:

$$\text{ன} + \text{ி} \neq \text{ன} + \text{◌்} + \text{இ}$$

As in the case of Devanagari, a consonant cluster is any sequence of one or more consonants separated by *viramas*, possibly terminated with a *virama*.

Two-Part Vowels

Certain Indic vowels consist of two discontiguous elements. As in other cases of discontiguous elements, there are two sequences of Unicode values which can be used to express equivalent spellings. This is similar to the case of letters such as "â," which can either be spelled with "a" followed by non-spacing "^", or spelled with a single Unicode character "â".

ொ (0BCA) ≈ ெ + ா (0BC6 + 0BBE)

ோ (0BCB) ≈ ே + ா (0BC7 + 0BBE)

ௌ (0BCC) ≈ ெ + ௗ (0BC7 + 0BD7)

Note that the ௗ in the third example is *not* the LLA vowel; it is the AU LENGTH MARK.

If the precomposed forms are used in the backing store instead of the separate characters, then a similar transformation occurs in the rendering process. The precomposed form on the left is transformed into the two separate forms equivalent to those on the right, which are then subject to vowel reordering, as below. Thus in rendering:

ொ	→	ெ + ா
ோ	→	ே + ா
ௌ	→	ெ + ௗ

Vowel Reordering

The following vowels are always reordered in front of the previous consonant cluster, similar to the rendering behavior of the DEVANAGARI VOWEL SIGN I.

ெ (0BC6) ே (0BC7) ை (0BC8)

For example:

Backing store			Display
க	ெ	→	கெ
க	ே	→	கே
க	ை	→	கை

The same effect occurs with the results of vowel splitting, as follows:

Backing store				Display
க	ொ		→	கொ
க	ெ	ா	→	கொ
க	ோ		→	கோ
க	ே	ா	→	கோ
க	ெள		→	கௌ
க	ெ	ள	→	கௌ

In both cases, the ordering of the elements is *unambiguous*: the consonant (cluster) occurs *first* in the backing store. The vowel ஔ also has two discontinuous parts, and can also be composed using the AU LENGTH MARK.

Ligatures

The following examples illustrate the range of ligatures available in Tamil. These changes take place after vowel reordering and vowel splitting. Unlike Devanagari, there are very few conjunct consonants; most ligatures are located between a vowel and a neighboring consonant.

1. Conjunct consonants.

க + ◌் + ஷ → க்ஷ

As with Devanagari, vowel reordering occurs around conjunct consonants. For example:

க + ◌் + ஷ + ெ + ா → க்ஷொ

2. The vowel ா optionally ligates with ண, ன, or ற on its left:

ண + ா → ணா

ன + ா → னா

ற + ா → றா

Since this process takes place after reordering and splitting, the following ligatures may also occur:

Separate Vowels	*Precomposed Vowels*
ண + ெ + ா → ெணை	ண + ொ → ொண
ண + ே + ா → ேணை	ண + ோ → ோண
ன + ெ + ா → ெனை	ன + ொ → ொன
ன + ே + ா → ேனை	ன + ோ → ோன
ற + ெ + ா → ெறை	ற + ொ → ொற
ற + ே + ா → ேறை	ற + ோ → ோற

3. The vowel signs ி and ீ form ligatures with ட on their left.

ட + ி → டி
ட + ீ → டீ

These vowels often change shape or position slightly in order to link up with the appropriate shape of the consonant on their left:

ல + ி → லி
ல + ீ → லீ

4. The vowel signs ு◌ and ூ◌ typically change form or ligate:

x	x + ு◌	x + ூ◌
க	கு	கூ
ங	ங	ஙூ
ச	சு	சூ
ஞ	ஞு	ஞூ
ட	டு	டூ
ண	ணு	ணூ
த	து	தூ
ந	நு	நூ
ன	னு	னூ

x	x + ு◌	x + ூ◌
ப	பு	பூ
ம	மு	மூ
ய	யு	யூ
ர	ரு	ரூ
ற	று	றூ
ல	லு	லூ
ள	ளு	ளூ
ழ	ழு	ழூ
வ	வு	வூ

To the right of ஜ, ஷ, ஸ, ஹ, or க்ஷ, these forms have a spacing form. For example:

ஜ + ◌ → ஜா

ஜ + ◌ → ஜி

5. The vowel sign ◌ெ changes to ◌ே to the left of ண, ன, ல, or ள.

◌ெ + ண	→	◌ேண
◌ெ + ன	→	◌ேன
◌ெ + ல	→	◌ேல
◌ெ + ள	→	◌ேள

Remember that this change takes place after the vowel reordering: in the first example, the vowel ◌ெ follows ண in the backing store. After vowel reordering, it is on the left of ண, and thus changes form. The complete process is:

ண + ◌ெ → ◌ெ + ண → ◌ேண

6. The consonant ர changes shape to ற.

This occurs when the ற would not be confused with the vowel ற. That is, when ர is combined with ◌், ◌ி, or ◌ீ.

ர + ◌் → ற + ◌்

ர + ◌ி → ற + ◌ி

ர + ◌ீ → ற + ◌ீ