The Unicode Standard
Version 6.2 – Core Specification

To learn about the latest version of the Unicode Standard, see http://www.unicode.org/versions/latest/.

The Unicode Standard / the Unicode Consortium ; edited by Julie D. Allen ... [et al.]. — Version 6.2.
Includes bibliographical references and index.
ISBN 978-1-936213-07-8) (http://www.unicode.org/versions/Unicode6.2.0/)
1. Unicode (Computer character set) I. Allen, Julie D. II. Unicode Consortium.
QA268.U545 2012

ISBN 978-1-936213-07-8
Published in Mountain View, CA
September 2012

Chapter 8

Middle Eastern Scripts

The scripts in this chapter have a common origin in the ancient Phoenician alphabet. They include:

Hebrew	***Samaritan***
Arabic	***Thaana***
Syriac	

The Hebrew script is used in Israel and for languages of the Diaspora. The Arabic script is used to write many languages throughout the Middle East, North Africa, and certain parts of Asia. The Syriac script is used to write a number of Middle Eastern languages. These three also function as major liturgical scripts, used worldwide by various religious groups. The Samaritan script is used in small communities in Israel and the Palestinian Territories to write the Samaritan Hebrew and Samaritan Aramaic languages. The Thaana script is used to write Dhivehi, the language of the Republic of Maldives, an island nation in the middle of the Indian Ocean.

The Middle Eastern scripts are mostly abjads, with small character sets. Words are demarcated by spaces. Except for Thaana, these scripts include a number of distinctive punctuation marks. In addition, the Arabic script includes traditional forms for digits, called "Arabic-Indic digits" in the Unicode Standard.

Text in these scripts is written from right to left. Implementations of these scripts must conform to the Unicode Bidirectional Algorithm (see Unicode Standard Annex #9, "Unicode Bidirectional Algorithm"). For more information about writing direction, see *Section 2.10, Writing Direction*. There are also special security considerations that apply to bidirectional scripts, especially with regard to their use in identifiers. For more information about these issues, see Unicode Technical Report #36, "Unicode Security Considerations."

Arabic and Syriac are cursive scripts even when typeset, unlike Hebrew, Samaritan, and Thaana, where letters are unconnected. Most letters in Arabic and Syriac assume different forms depending on their position in a word. Shaping rules for the rendering of text are specified in *Section 8.2, Arabic*, and *Section 8.3, Syriac*. Shaping rules are not required for Hebrew because only five letters have position-dependent final forms, and these forms are separately encoded.

Historically, Middle Eastern scripts did not write short vowels. Nowadays, short vowels are represented by marks positioned above or below a consonantal letter. Vowels and other marks of pronunciation ("vocalization") are encoded as combining characters, so support for vocalized text necessitates use of composed character sequences. Yiddish, Syriac, and Thaana are normally written with vocalization; Hebrew, Samaritan, and Arabic are usually written unvocalized.

8.1 Hebrew

Hebrew: U+0590–U+05FF

The Hebrew script is used for writing the Hebrew language as well as Yiddish, Judezmo (Ladino), and a number of other languages. Vowels and various other marks are written as *points,* which are applied to consonantal base letters; these marks are usually omitted in Hebrew, except for liturgical texts and other special applications. Five Hebrew letters assume a different graphic form when they occur last in a word.

Directionality. The Hebrew script is written from right to left. Conformant implementations of Hebrew script must use the Unicode Bidirectional Algorithm (see Unicode Standard Annex #9, "Unicode Bidirectional Algorithm").

Cursive. The Unicode Standard uses the term *cursive* to refer to writing where the letters of a word are connected. A handwritten form of Hebrew is known as cursive, but its rounded letters are generally unconnected, so the Unicode definition does not apply. Fonts based on cursive Hebrew exist. They are used not only to show examples of Hebrew handwriting, but also for display purposes.

Standards. ISO/IEC 8859-8—Part 8. *Latin/Hebrew Alphabet*. The Unicode Standard encodes the Hebrew alphabetic characters in the same relative positions as in ISO/IEC 8859-8; however, there are no points or Hebrew punctuation characters in that ISO standard.

Vowels and Other Marks of Pronunciation. These combining marks, generically called *points* in the context of Hebrew, indicate vowels or other modifications of consonantal letters. General rules for applying combining marks are given in *Section 2.11, Combining Characters*, and *Section 3.6, Combination*. Additional Hebrew-specific behavior is described below.

Hebrew points can be separated into four classes: *dagesh*, *shin dot* and *sin dot*, vowels, and other marks of punctuation.

Dagesh, U+05BC HEBREW POINT DAGESH OR MAPIQ, has the form of a dot that appears inside the letter that it affects. It is not a vowel but rather a diacritic that affects the pronunciation of a consonant. The same base consonant can also have a vowel and/or other diacritics. *Dagesh* is the only element that goes inside a letter.

The dotted Hebrew consonant *shin* is explicitly encoded as the sequence U+05E9 HEBREW LETTER SHIN followed by U+05C1 HEBREW POINT SHIN DOT. The *shin dot* is positioned on the upper-right side of the undotted base letter. Similarly, the dotted consonant *sin* is explicitly encoded as the sequence U+05E9 HEBREW LETTER SHIN followed by U+05C2 HEBREW POINT SIN DOT. The *sin dot* is positioned on the upper-left side of the base letter. The two dots are mutually exclusive. The base letter *shin* can also have a *dagesh*, a vowel, and other diacritics. The two dots are not used with any other base character.

Vowels all appear below the base character that they affect, except for *holam*, U+05B9 HEBREW POINT HOLAM, which appears above left. The following points represent vowels: U+05B0..U+05BB, and U+05C7.

The remaining three points are *marks of pronunciation*: U+05BD HEBREW POINT METEG, U+05BF HEBREW POINT RAFE, and U+FB1E HEBREW POINT JUDEO-SPANISH VARIKA. *Meteg*, also known as *siluq*, goes below the base character; *rafe* and *varika* go above it. The varika, used in Judezmo, is a glyphic variant of *rafe*.

Shin and Sin. Separate characters for the dotted letters *shin* and *sin* are not included in this block. When it is necessary to distinguish between the two forms, they should be encoded as U+05E9 HEBREW LETTER SHIN followed by the appropriate dot, either U+05C1 HEBREW POINT SHIN DOT or U+05C2 HEBREW POINT SIN DOT. (See preceding discussion.) This practice is consistent with Israeli standard encoding.

Final (Contextual Variant) Letterforms. Variant forms of five Hebrew letters are encoded as separate characters in this block, as in Hebrew standards including ISO/IEC 8859-8. These variant forms are generally used in place of the nominal letterforms at the end of words. Certain words, however, are spelled with nominal rather than final forms, particularly names and foreign borrowings in Hebrew and some words in Yiddish. Because final form usage is a matter of spelling convention, software should not automatically substitute final forms for nominal forms at the end of words. The positional variants should be coded directly and rendered one-to-one via their own glyphs—that is, without contextual analysis.

Yiddish Digraphs. The digraphs are considered to be independent characters in Yiddish. The Unicode Standard has included them as separate characters so as to distinguish certain letter combinations in Yiddish text—for example, to distinguish the digraph *double vav* from an occurrence of a consonantal *vav* followed by a vocalic *vav*. The use of digraphs is consistent with standard Yiddish orthography. Other letters of the Yiddish alphabet, such as *pasekh alef,* can be composed from other characters, although alphabetic presentation forms are also encoded.

Punctuation. Most punctuation marks used with the Hebrew script are not given independent codes (that is, they are unified with Latin punctuation) except for the few cases where the mark has a unique form in Hebrew—namely, U+05BE HEBREW PUNCTUATION MAQAF, U+05C0 HEBREW PUNCTUATION PASEQ (also known as *legarmeh*), U+05C3 HEBREW PUNCTUATION SOF PASUQ, U+05F3 HEBREW PUNCTUATION GERESH, and U+05F4 HEBREW PUNCTUATION GERSHAYIM. For paired punctuation such as parentheses, the glyphs chosen to represent U+0028 LEFT PARENTHESIS and U+0029 RIGHT PARENTHESIS will depend on the direction of the rendered text. See *Section 4.7, Bidi Mirrored*, for more information. For additional punctuation to be used with the Hebrew script, see *Section 6.2, General Punctuation*.

Cantillation Marks. Cantillation marks are used in publishing liturgical texts, including the Bible. There are various historical schools of cantillation marking; the set of marks included in the Unicode Standard follows the Israeli standard SI 1311.2.

Positioning. Marks may combine with vowels and other points, and complex typographic rules dictate how to position these combinations.

The vertical placement (meaning above, below, or inside) of points and marks is very well defined. The horizontal placement (meaning left, right, or center) of points is also very well defined. The horizontal placement of marks, by contrast, is not well defined, and convention allows for the different placement of marks relative to their base character.

When points and marks are located below the same base letter, the point always comes first (on the right) and the mark after it (on the left), except for the marks *yetiv*, U+059A HEBREW ACCENT YETIV, and *dehi*, U+05AD HEBREW ACCENT DEHI. These two marks come first (on the right) and are followed (on the left) by the point.

These rules are followed when points and marks are located above the same base letter:

- If the point is *holam*, all cantillation marks precede it (on the right) except *pashta*, U+0599 HEBREW ACCENT PASHTA.
- *Pashta* always follows (goes to the left of) points.

- *Holam* on a *sin* consonant (*shin* base + *sin dot*) follows (goes to the left of) the *sin dot*. However, the two combining marks are sometimes rendered as a single assimilated dot.
- *Shin dot* and *sin dot* are generally represented closer vertically to the base letter than other points and marks that go above it.

Meteg. *Meteg*, U+05BD HEBREW POINT METEG, frequently co-occurs with vowel points below the consonant. Typically, *meteg* is placed to the left of the vowel, although in some manuscripts and printed texts it is positioned to the right of the vowel. The difference in positioning is not known to have any semantic significance; nevertheless, some authors wish to retain the positioning found in source documents.

The alternate *vowel-meteg* ordering can be represented in terms of alternate ordering of characters in encoded representation. However, because of the fixed-position canonical combining classes to which *meteg* and vowel points are assigned, differences in ordering of such characters are not preserved under normalization. The *combining grapheme joiner* can be used within a *vowel-meteg* sequence to preserve an ordering distinction under normalization. For more information, see the description of U+034F COMBINING GRAPHEME JOINER in *Section 16.2, Layout Controls*.

For example, to display *meteg* to the left of (after, for a right-to-left script) the vowel point *sheva*, U+05B0 HEBREW POINT SHEVA, the sequence of *meteg* following *sheva* can be used:

<sheva, meteg>

Because these marks are canonically ordered, this sequence is preserved under normalization. Then, to display *meteg* to the right of the *sheva*, the sequence with *meteg* preceding *sheva* with an intervening CGJ can be used:

<meteg, CGJ, sheva>

A further complication arises for combinations of *meteg* with *hataf* vowels: U+05B1 HEBREW POINT HATAF SEGOL, U+05B2 HEBREW POINT HATAF PATAH, and U+05B3 HEBREW POINT HATAF QAMATS. These vowel points have two side-by-side components. *Meteg* can be placed to the left or the right of a *hataf* vowel, but it also is often placed between the two components of the *hataf* vowel. A three-way positioning distinction is needed for such cases.

The *combining grapheme joiner* can be used to preserve an ordering that places *meteg* to the right of a *hataf* vowel, as described for combinations of *meteg* with non-*hataf* vowels, such as *sheva*.

Placement of *meteg* between the components of a *hataf* vowel can be conceptualized as a ligature of the *hataf* vowel and a nominally positioned *meteg*. With this in mind, the ligation-control functionality of U+200D ZERO WIDTH JOINER and U+200C ZERO WIDTH NON-JOINER can be used as a mechanism to control the visual distinction between a nominally positioned *meteg* to the left of a *hataf* vowel versus the medially positioned *meteg* within the *hataf* vowel. That is, *zero width joiner* can be used to request explicitly a medially positioned *meteg*, and *zero width non-joiner* can be used to request explicitly a left-positioned *meteg*. Just as different font implementations may or may not display an "fi" ligature by default, different font implementations may or may not display *meteg* in a medial position when combined with *hataf* vowels by default. As a result, authors who want to ensure left-position versus medial-position display of *meteg* with *hataf* vowels across all font implementations may use joiner characters to distinguish these cases.

Thus the following encoded representations can be used for different positioning of *meteg* with a *hataf* vowel, such as *hataf patah*:

left-positioned *meteg*: <hataf patah, ZWNJ, meteg>

medially positioned *meteg*: <hataf patah, ZWJ, meteg>

right-positioned *meteg*: <meteg, CGJ, hataf patah>

In no case is use of ZWNJ, ZWJ, or CGJ *required* for representation of *meteg*. These recommendations are simply provided for interoperability in those instances where authors wish to preserve specific positional information regarding the layout of a *meteg* in text.

Atnah Hafukh and Qamats Qatan. In some older versions of Biblical text, a distinction is made between the accents U+05A2 HEBREW ACCENT ATNAH HAFUKH and U+05AA HEBREW ACCENT YERAH BEN YOMO. Many editions from the last few centuries do not retain this distinction, using only *yerah ben yomo*, but some users in recent decades have begun to reintroduce this distinction. Similarly, a number of publishers of Biblical or other religious texts have introduced a typographic distinction for the vowel point *qamats* corresponding to two different readings. The original letterform used for one reading is referred to as *qamats* or *qamats gadol*; the new letterform for the other reading is *qamats qatan*. Not all users of Biblical Hebrew use *atnah hafukh* and *qamats qatan*. If the distinction between accents *atnah hafukh* and *yerah ben yomo* is not made, then only U+05AA HEBREW ACCENT YERAH BEN YOMO is used. If the distinction between vowels *qamats gadol* and *qamats qatan* is not made, then only U+05B8 HEBREW POINT QAMATS is used. Implementations that support Hebrew accents and vowel points may not necessarily support the special-usage characters U+05A2 HEBREW ACCENT ATNAH HAFUKH and U+05C7 HEBREW POINT QAMATS QATAN.

Holam Male and Holam Haser. The vowel point *holam* represents the vowel phoneme /o/. The consonant letter *vav* represents the consonant phoneme /w/, but in some words is used to represent a vowel, /o/. When the point *holam* is used on *vav*, the combination usually represents the vowel /o/, but in a very small number of cases represents the consonant-vowel combination /wo/. A typographic distinction is made between these two in many versions of Biblical text. In most cases, in which *vav* + *holam* together represents the vowel /o/, the point *holam* is centered above the *vav* and referred to as *holam male*. In the less frequent cases, in which the *vav* represents the consonant /w/, some versions show the point *holam* positioned above left. This is referred to as *holam haser*. The character U+05BA HEBREW POINT HOLAM HASER FOR VAV is intended for use as *holam haser* only in those cases where a distinction is needed. When the distinction is made, the character U+05B9 HEBREW POINT HOLAM is used to represent the point *holam male on vav*. U+05BA HEBREW POINT HOLAM HASER FOR VAV is intended for use only on *vav*; results of combining this character with other base characters are not defined. Not all users distinguish between the two forms of *holam*, and not all implementations can be assumed to support U+05BA HEBREW POINT HOLAM HASER FOR VAV.

Puncta Extraordinaria. In the Hebrew Bible, dots are written in various places above or below the base letters that are distinct from the vowel points and accents. These dots are referred to by scholars as *puncta extraordinaria*, and there are two kinds. The *upper punctum*, the more common of the two, has been encoded since Unicode 2.0 as U+05C4 HEBREW MARK UPPER DOT. The *lower punctum* is used in only one verse of the Bible, Psalm 27:13, and is encoded as U+05C5 HEBREW MARK LOWER DOT. The *puncta* generally differ in appearance from dots that occur above letters used to represent numbers; the number dots should be represented using U+0307 COMBINING DOT ABOVE and U+0308 COMBINING DIAERESIS.

Nun Hafukha. The *nun hafukha* is a special symbol that appears to have been used for scribal annotations, although its exact functions are uncertain. It is used a total of nine times in the Hebrew Bible, although not all versions include it, and there are variations in the exact locations in which it is used. There is also variation in the glyph used: it often has the appearance of a rotated or reversed *nun* and is very often called *inverted nun*; it may also appear similar to a *half tet* or have some other form.

Currency Symbol. The NEW SHEQEL SIGN (U+20AA) is encoded in the currency block.

Alphabetic Presentation Forms: U+FB1D–U+FB4F

The Hebrew characters in this block are chiefly of two types: variants of letters and marks encoded in the main Hebrew block, and precomposed combinations of a Hebrew letter or digraph with one or more vowels or pronunciation marks. This block contains all of the vocalized letters of the Yiddish alphabet. The *alef lamed* ligature and a Hebrew variant of the plus sign are included as well. The Hebrew plus sign variant, U+FB29 HEBREW LETTER ALTERNATIVE PLUS SIGN, is used more often in handwriting than in print, but it does occur in school textbooks. It is used by those who wish to avoid cross symbols, which can have religious and historical connotations.

U+FB20 HEBREW LETTER ALTERNATIVE AYIN is an alternative form of *ayin* that may replace the basic form U+05E2 HEBREW LETTER AYIN when there is a diacritical mark below it. The basic form of *ayin* is often designed with a descender, which can interfere with a mark below the letter. U+FB20 is encoded for compatibility with implementations that substitute the alternative form in the character data, as opposed to using a substitute glyph at rendering time.

Use of Wide Letters. Wide letterforms are used in handwriting and in print to achieve even margins. The wide-form letters in the Unicode Standard are those that are most commonly "stretched" in justification. If Hebrew text is to be rendered with even margins, justification should be left to the text-formatting software.

These alphabetic presentation forms are included for compatibility purposes. For the preferred encoding, see the Hebrew presentation forms, U+FB1D..U+FB4F.

For letterlike symbols, see U+2135..U+2138.

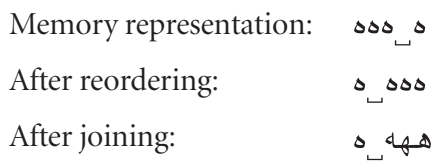
8.2 Arabic

Arabic: U+0600–U+06FF

The Arabic script is used for writing the Arabic language and has been extended to represent a number of other languages, such as Persian, Urdu, Pashto, Sindhi, and Kurdish, as well as many African languages. Urdu is often written with the ornate Nastaliq script variety. Some languages, such as Indonesian/Malay, Turkish, and Ingush, formerly used the Arabic script but now employ the Latin or Cyrillic scripts.

The Arabic script is cursive, even in its printed form (see *Figure 8-1*). As a result, the same letter may be written in different forms depending on how it joins with its neighbors. Vowels and various other marks may be written as combining marks called *harakat*, which are applied to consonantal base letters. In normal writing, however, these harakat are omitted.

Figure 8-1. Directionality and Cursive Connection



Directionality. The Arabic script is written from right to left. Conformant implementations of Arabic script must use the Unicode Bidirectional Algorithm to reorder the memory representation for display (see Unicode Standard Annex #9, "Unicode Bidirectional Algorithm").

Standards. ISO/IEC 8859-6—Part 6. *Latin/Arabic Alphabet*. The Unicode Standard encodes the basic Arabic characters in the same relative positions as in ISO/IEC 8859-6. ISO/IEC 8859-6, in turn, is based on ECMA-114, which was based on ASMO 449.

Encoding Principles. The basic set of Arabic letters is well defined. Each letter receives only one Unicode character value in the basic Arabic block, no matter how many different contextual appearances it may exhibit in text. Each Arabic letter in the Unicode Standard may be said to represent the inherent semantic identity of the letter. A word is spelled as a sequence of these letters. The representative glyph shown in the Unicode character chart for an Arabic letter is usually the form of the letter when standing by itself. It is simply used to distinguish and identify the character in the code charts and does not restrict the glyphs used to represent it. See "Arabic Cursive Joining," "Arabic Ligatures," and "Arabic Joining Groups" in the following text for an extensive discussion of how cursive joining and positional variants of Arabic letters are handled by the Unicode Standard.

The following principles guide the encoding of the various types of marks which are applied to the basic Arabic letter skeletons:

1. **Ijam**: Diacritic marks applied to basic letter forms to derive new (usually consonant) letters for extended Arabic alphabets are not separately encoded as combining marks. Instead, each letter plus *ijam* combination is encoded as a separate, atomic character. These letter plus *ijam* characters are never given decompositions in the standard. *Ijam* generally take the form of one-, two-, three- or four-dot markings above or below the basic letter skeleton, although other diacritic forms occur in extensions of the Arabic script in Central and South Asia and in Africa. In discussions of Arabic in Unicode, *ijam* are often also referred to as *nukta*, because of their functional similarity to the nukta diacritic marks which occur in many Indic scripts.

2. **Tashkil**: Marks functioning to indicate vocalization of text, as well as other types of phonetic guides to correct pronunciation, are separately encoded as combining marks. These include several subtypes: *harakat* (short vowel marks), *tanwin* (postnasalized or long vowel marks), and *shaddah* (consonant gemination mark). A basic Arabic letter plus any of these types of marks is never encoded as a separate, precomposed character, but must always be represented as a sequence of letter plus combining mark. Additional marks invented to indicate non-Arabic vowels, used in extensions of the Arabic script, are also encoded as separate combining marks.

3. **Maddah**: The *maddah* is a particular case of a *harakat* mark which has exceptional treatment in the standard. It occurs only above *alef*, and in that combination represents the sound /ʔaa/. For historical reasons, the precomposed combination U+0622 ARABIC LETTER ALEF WITH MADDA ABOVE is encoded, but the combining mark U+0653 ARABIC MADDAH ABOVE is also encoded. U+0622 is given a canonical decomposition to the sequence of *alef* followed by the *combining maddah*.

4. **Hamza**: The *hamza* may occur above or below other letters. Its treatment in the Unicode Standard is also exceptional and rather complex. The general principle is that when such a *hamza* is used to indicate an actual glottal stop in text, it should be represented with a separate combining mark, either U+0654 ARABIC HAMZA ABOVE or U+0655 ARABIC HAMZA BELOW. However, when the *hamza*

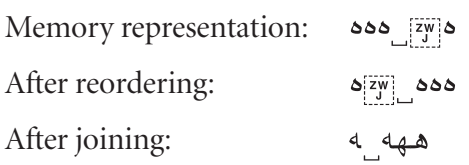
mark is used as a diacritic to derive a separate letter as an extension of the Arabic script, then the basic letter skeleton plus the *hamza* mark is represented by a single, precomposed character. See "Combining Hamza Above" later in this section for discussion of the complications for particular characters.

5. **Annotation Marks**: Koranic annotation marks are always encoded as separate combining marks.

Punctuation. Most punctuation marks used with the Arabic script are not given independent codes (that is, they are unified with Latin punctuation), except for the few cases where the mark has a significantly different appearance in Arabic—namely, U+060C ARABIC COMMA, U+061B ARABIC SEMICOLON, U+061E ARABIC TRIPLE DOT PUNCTUATION MARK, U+061F ARABIC QUESTION MARK, and U+066A ARABIC PERCENT SIGN. For paired punctuation such as parentheses, the glyphs chosen to represent U+0028 LEFT PARENTHESIS and U+0029 RIGHT PARENTHESIS will depend on the direction of the rendered text.

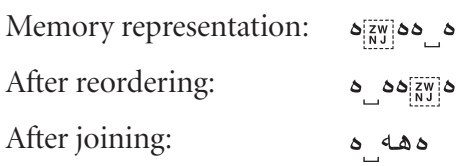
The Non-joiner and the Joiner. The Unicode Standard provides two user-selectable formatting codes: U+200C ZERO WIDTH NON-JOINER and U+200D ZERO WIDTH JOINER. The use of a joiner adjacent to a suitable letter permits that letter to form a cursive connection without a visible neighbor. This provides a simple way to encode some special cases, such as exhibiting a connecting form in isolation, as shown in *Figure 8-2*.

Figure 8-2. Using a Joiner



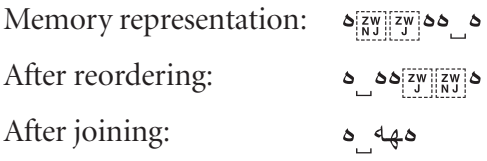
The use of a non-joiner between two letters prevents those letters from forming a cursive connection with each other when rendered, as shown in *Figure 8-3*. Examples include the Persian plural suffix, some Persian proper names, and Ottoman Turkish vowels.

Figure 8-3. Using a Non-joiner



Joiners and non-joiners may also occur in combinations. The effects of such combinations are shown in *Figure 8-4*. For further discussion of joiners and non-joiners, see *Section 16.2, Layout Controls*.

Figure 8-4. Combinations of Joiners and Non-joiners



Harakat (Vowel) Nonspacing Marks. *Harakat* are marks that indicate vowels or other modifications of consonant letters. The code charts depict a character in the harakat range in relation to a dashed circle, indicating that this character is intended to be applied via some process *to the character that precedes it* in the text stream (that is, the base character). General rules for applying nonspacing marks are given in *Section 7.9, Combining Marks*. The few marks that are placed after (to the left of) the base character are treated as ordinary spacing characters in the Unicode Standard. The Unicode Standard does not specify a sequence order in case of multiple harakat applied to the same Arabic base character, as there is no possible ambiguity of interpretation. For more information about the canonical ordering of nonspacing marks, see *Section 2.11, Combining Characters,* and *Section 3.11, Normalization Forms*.

The placement and rendering of vowel and other marks in Arabic strongly depends on the typographical environment or even the typographical style. For example, in the Unicode code charts, the default position of U+0651 ◌ّ ARABIC SHADDA is with the glyph placed above the base character, whereas for U+064D ◌ٍ ARABIC KASRATAN the glyph is placed below the base character, as shown in the first example in *Figure 8-5*. However, computer fonts often follow an approach that originated in metal typesetting and combine the *kasratan* with *shadda* in a ligature placed above the text, as shown in the second example in *Figure 8-5*.

Figure 8-5. Placement of Harakat

Arabic-Indic Digits. The names for the forms of decimal digits vary widely across different languages. The decimal numbering system originated in India (Devanagari ०१२३…) and was subsequently adopted in the Arabic world with a different appearance (Arabic ٠١٢٣…). The Europeans adopted decimal numbers from the Arabic world, although once again the forms of the digits changed greatly (European 0123…). The European forms were later adopted widely around the world and are used even in many Arabic-speaking countries in North Africa. In each case, the interpretation of decimal numbers remained the same. However, the forms of the digits changed to such a degree that they are no longer recognizably the same characters. Because of the origin of these characters, the European decimal numbers are widely known as "Arabic numerals" or "Hindi-Arabic numerals," whereas the decimal numbers in use in the Arabic world are widely known there as "Hindi numbers."

The Unicode Standard includes *Indic* digits (including forms used with different Indic scripts), *Arabic* digits (with forms used in most of the Arabic world), and *European* digits (now used internationally). Because of this decision, the traditional names could not be retained without confusion. In addition, there are two main variants of the Arabic digits: those used in Iran, Pakistan, and Afghanistan (here called *Eastern Arabic-Indic*) and those used in other parts of the Arabic world. In summary, the Unicode Standard uses the names shown in *Table 8-1*. A different set of digits, called Rumi, was used in historical materials from Egypt to Spain, and is discussed in the subsection on "Rumi Numeral Forms" in *Section 15.3, Numerals*.

There is substantial variation in usage of glyphs for the Eastern Arabic-Indic digits, especially for the digits four, five, six, and seven. *Table 8-2* illustrates this variation with some example glyphs for digits in languages of Iran, Pakistan, and India. While some usage of the

Table 8-1. Arabic Digit Names

Name	Code Points	Forms
European	U+0030..U+0039	0123456789
Arabic-Indic	U+0660..U+0669	٠١٢٣٤٥٦٧٨٩
Eastern Arabic-Indic	U+06F0..U+06F9	۰۱۲۳۴۵۶۷۸۹
Indic (Devanagari)	U+0966..U+096F	०१२३४५६७८९

Persian glyph for U+06F7 EXTENDED ARABIC-INDIC DIGIT SEVEN can be documented for Sindhi, the form shown in *Table 8-2* is predominant.

Table 8-2. Glyph Variation in Eastern Arabic-Indic Digits

Code Point	Digit	Persian	Sindhi	Urdu
U+06F4	4	۴	۴	۴
U+06F5	5	۵	۵	۵
U+06F6	6	۶	٦	٦
U+06F7	7	۷	۷	۷

The Unicode Standard provides a single, complete sequence of digits for Persian, Sindhi, and Urdu to account for the differences in appearance and directional treatment when rendering them. (For a complete discussion of directional formatting of numbers in the Unicode Standard, see Unicode Standard Annex #9, "Unicode Bidirectional Algorithm.")

Extended Arabic Letters. Arabic script is used to write major languages, such as Persian and Urdu, but it has also been used to transcribe some lesser-used languages, such as Baluchi and Lahnda, which have little tradition of printed typography. As a result, the Unicode Standard encodes multiple forms of some Extended Arabic letters because the character forms and usages are not well documented for a number of languages. For additional extended Arabic letters, see the Arabic Supplement block, U+0750..U+077F and the Arabic Extended-A block, U+08A0 ..U+08FF.

Koranic Annotation Signs. These characters are used in the Koran to mark pronunciation and other annotation. The enclosing mark U+06DE is used to enclose a digit. When rendered, the digit appears in a smaller size. Several additional Koranic annotation signs are encoded in the Arabic Extended-A block, U+08A0..U+08FF.

Additional Vowel Marks. When the Arabic script is adopted as the writing system for a language other than Arabic, it is often necessary to represent vowel sounds or distinctions not made in Arabic. In some cases, conventions such as the addition of small dots above and/or below the standard Arabic *fatha*, *damma*, and *kasra* signs have been used.

Classical Arabic has only three canonical vowels (/a/, /i/, /u/), whereas languages such as Urdu and Persian include other contrasting vowels such as /o/ and /e/. For this reason, it is imperative that speakers of these languages be able to show the difference between /e/ and /i/ (U+0656 ARABIC SUBSCRIPT ALEF), and between /o/ and /u/ (U+0657 ARABIC INVERTED DAMMA). At the same time, the use of these two diacritics in Arabic is redundant, merely emphasizing that the underlying vowel is long.

U+065F ARABIC WAVY HAMZA BELOW is an additional vowel mark used in Kashmiri. It can appear in combination with many characters. The particular combination of an *alef* with this vowel mark should be written with the sequence <U+0627 ARABIC LETTER ALEF, U+065F ARABIC WAVY HAMZA BELOW>, rather than with the character U+0673 ARABIC LETTER ALEF WITH WAVY HAMZA BELOW, which has been deprecated and which is not

canonically equivalent. However, implementations should be aware that there may be existing legacy Kashmiri data in which U+0673 occurs.

Honorifics. Marks known as honorifics represent phrases expressing the status of a person and are in widespread use in the Arabic-script world. Most have a specifically religious meaning. In effect, these marks are combining characters at the word level, rather than being associated with a single base character. Depending on the letter shapes present in the name and the calligraphic style in use, the honorific mark may be applied to a letter somewhere in the middle of the name. The normalization algorithm does not move such word-level combining characters to the end of the word.

Arabic Mathematical Symbols. A few Arabic mathematical symbols are encoded in this block. The Arabic mathematical radix signs, U+0606 ARABIC-INDIC CUBE ROOT and U+0607 ARABIC-INDIC FOURTH ROOT, differ from simple mirrored versions of U+221B CUBE ROOT and U+221C FOURTH ROOT, in that the digit portions of the symbols are written with Arabic-Indic digits and are not mirrored. U+0608 ARABIC RAY is a letterlike symbol used in Arabic mathematics.

Date Separator. U+060D ARABIC DATE SEPARATOR is used in Pakistan and India between the numeric date and the month name when writing out a date. This sign is distinct from U+002F SOLIDUS, which is used, for example, as a separator in currency amounts.

Full Stop. U+061E ARABIC TRIPLE DOT PUNCTUATION MARK is encoded for traditional orthographic practice using the Arabic script to write African languages such as Hausa, Wolof, Fulani, and Mandinka. These languages use ARABIC TRIPLE DOT PUNCTUATION MARK as a full stop.

Currency Symbols. U+060B AFGHANI SIGN is a currency symbol used in Afghanistan. The symbol is derived from an abbreviation of the name of the currency, which has become a symbol in its own right. U+FDFC RIAL SIGN is a currency symbol used in Iran. Unlike the AFGHANI SIGN, U+FDFC RIAL SIGN is considered a compatability character, encoded for compatibility with Iranian standards. Ordinarily in Persian "rial" is simply spelled out as the sequence of letters, <0631, 06CC, 0627, 0644>.

End of Ayah. U+06DD ARABIC END OF AYAH graphically encloses a sequence of zero or more digits (of General Category Nd) that follow it in the data stream. The enclosure terminates with any non-digit. For behavior of a similar prefixed formatting control, see the discussion of U+070F SYRIAC ABBREVIATION MARK in *Section 8.3, Syriac*.

Other Signs Spanning Numbers. Several other special signs are written in association with numbers in the Arabic script. U+0600 ARABIC NUMBER SIGN signals the beginning of a number; it is written below the digits of the number.

U+0601 ARABIC SIGN SANAH indicates a year (that is, as part of a date). This sign is rendered below the digits of the number it precedes. Its appearance is a vestigial form of the Arabic word for *year*, /sanatu/ (*seen noon teh-marbuta*), but it is now a sign in its own right and is widely used to mark a numeric year even in non-Arabic languages where the Arabic word would not be known. The use of the year sign is illustrated in *Figure 8-6*.

Figure 8-6. Arabic Year Sign

U+0602 ARABIC FOOTNOTE MARKER is another of these signs; it is used in the Arabic script in conjunction with the footnote number itself. It also precedes the digits in logical order and is written to extend underneath them.

Finally, U+0603 ARABIC SIGN SAFHA functions as a page sign, preceding and extending under a sequence of digits for a page number.

Like U+06DD ARABIC END OF AYAH, all of these signs can span multiple-digit numbers, rather than just a single digit. They are not formally considered *combining marks* in the sense used by the Unicode Standard, although they clearly interact graphically with the sequence of digits that follows them. They *precede* the sequence of digits that they span, rather than following a base character, as would be the case for a combining mark. Their General Category value is Cf (format control character). Unlike most other format control characters, however, they should be rendered with a visible glyph, even in circumstances where no suitable digit or sequence of digits follows them in logical order.

Poetic Verse Sign. U+060E ARABIC POETIC VERSE SIGN is a special symbol often used to mark the beginning of a poetic verse. Although it is similar to U+0602 ARABIC FOOTNOTE MARKER in appearance, the poetic sign is simply a symbol. In contrast, the footnote marker is a format control character that has complex rendering in conjunction with following digits. U+060F ARABIC SIGN MISRA is another symbol used in poetry.

Arabic Cursive Joining

Minimum Rendering Requirements. A rendering or display process must convert between the logical order in which characters are placed in the backing store and the visual (or physical) order required by the display device. See Unicode Standard Annex #9, "Unicode Bidirectional Algorithm," for a description of the conversion between logical and visual orders.

The cursive nature of the Arabic script imposes special requirements on display or rendering processes that are not typically found in Latin script-based systems. At a minimum, a display process must select an appropriate glyph to depict each Arabic letter according to its immediate *joining* context; furthermore, it must substitute certain ligature glyphs for sequences of Arabic characters. The remainder of this section specifies a minimum set of rules that provide legible Arabic joining and ligature substitution behavior.

Joining Types. Each Arabic letter must be depicted by one of a number of possible contextual glyph forms. The appropriate form is determined on the basis of the cursive joining behavior of that character as it interacts with the cursive joining behavior of adjacent characters. In the Unicode Standard, such cursive joining behavior is formally described in terms of values of a character property called Joining_Type. Each Arabic character falls into one of the types shown in *Table 8-3*. (See ArabicShaping.txt in the Unicode Character Database for a complete list.) In this table, *right* and *left* refer to visual order. The characters of the right-joining type are exemplified in more detail in *Table 8-9*, and those of the dual-joining type are shown in *Table 8-8*. When characters do not join or cause joining (such as DAMMATAN), they are classified as transparent.

Table 8-3. Primary Arabic Joining Types

Description	Joining_Type	Examples and Comments
Right-joining	R	ALEF, DAL, THAL, REH, ZAIN …
Left-joining	L	None
Dual-joining	D	BEH, TEH, THEH, JEEM …
Join-causing	C	U+200D ZERO WIDTH JOINER and TATWEEL (0640). These characters are distinguished from the dual-joining characters in that they do not change shape themselves.
Non-joining	U	U+200C ZERO WIDTH NON-JOINER and all spacing characters, except those explicitly mentioned as being one of the other joining types, are non-joining. These include HAMZA (0621), HIGH HAMZA (0674), spaces, digits, punctuation, non-Arabic letters, and so on. Also, U+0600 ARABIC NUMBER SIGN..U+0603 ARABIC SIGN SAFHA and U+06DD ARABIC END OF AYAH.
Transparent	T	All nonspacing marks (General Category Mn or Me) and most format control characters (General Category Cf) are transparent to cursive joining. These include FATHATAN (064B) and other Arabic *harakat*, HAMZA BELOW (0655), SUPERSCRIPT ALEF (0670), combining Koranic annotation signs, and nonspacing marks from other scripts. Also U+070F SYRIAC ABBREVIATION MARK.

Table 8-4 defines derived superclasses of the primary Arabic joining types; those derived types are used in the cursive joining rules. In this table, *right* and *left* refer to visual order.

Table 8-4. Derived Arabic Joining Types

Description	Derivation
Right join-causing	Superset of dual-joining, left-joining, and join-causing
Left join-causing	Superset of dual-joining, right-joining, and join-causing

Joining Rules. The following rules describe the joining behavior of Arabic letters in terms of their display (visual) order. In other words, the positions of letterforms in the included examples are presented as they would appear on the screen *after* the Bidirectional Algorithm has reordered the characters of a line of text.

An implementation may choose to restate the following rules according to logical order so as to apply them *before* the Bidirectional Algorithm's reordering phase. In this case, the words *right* and *left* as used in this section would become *preceding* and *following*.

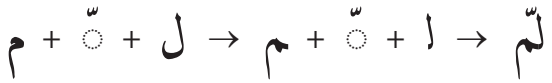
In the following rules, if X refers to a character, then various glyph types representing that character are referred to as shown in *Table 8-5*.

Table 8-5. Arabic Glyph Types

Glyph Type	Description
X_n	Nominal glyph form as it appears in the code charts
X_r	Right-joining glyph form (both right-joining and dual-joining characters may employ this form)
X_l	Left-joining glyph form (both left-joining and dual-joining characters may employ this form)
X_m	Dual-joining (medial) glyph form that joins on both left and right (only dual-joining characters employ this form)

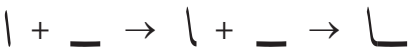
R1 ***Transparent characters do not affect the joining behavior of base (spacing) characters. For example:***

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



R2 ***A right-joining character X that has a right join-causing character on the right will adopt the form X_r. For example:***

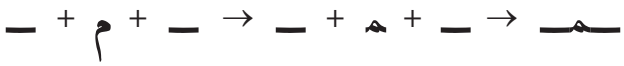
$$\text{ALEF}_n + \text{TATWEEL}_n \rightarrow \text{ALEF}_r + \text{TATWEEL}_n$$



R3 ***A left-joining character X that has a left join-causing character on the left will adopt the form X_l.***

R4 ***A dual-joining character X that has a right join-causing character on the right and a left join-causing character on the left will adopt the form X_m. For example:***

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

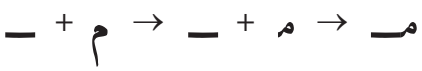


R5 ***A dual-joining character X that has a right join-causing character on the right and no left join-causing character on the left will adopt the form X_r. For example:***

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

R6 ***A dual-joining character X that has a left join-causing character on the left and no right join-causing character on the right will adopt the form X_l. For example:***

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



R7 ***If none of the preceding rules applies to a character X, then it will adopt the nominal form X_n.***

The cursive joining behavior described here for the Arabic script is also generally applicable to other cursive scripts such as Syriac. Specific circumstances may modify the application of the rules just described.

As noted earlier in this section, the ZERO WIDTH NON-JOINER may be used to prevent joining, as in the Persian plural suffix or Ottoman Turkish vowels.

Arabic Ligatures

Ligature Classes. Certain types of ligatures are obligatory in Arabic script regardless of font design. Many other optional ligatures are possible, depending on font design. Because they are optional, those ligatures are not covered in this discussion.

For the purpose of describing the obligatory Arabic ligatures, certain characters fall into two joining groups, as shown in *Table 8-6*. The complete list is available in ArabicShaping.txt in the Unicode Character Database.

Table 8-6. Arabic Obligatory Ligature Joining Groups

Joining Group	Examples
ALEF	MADDA-ON-ALEF, HAMZA ON ALEF, ...
LAM	LAM, LAM WITH SMALL V, LAM WITH DOT ABOVE, ...

Ligature Rules. The following rules describe the formation of ligatures. They are applied after the preceding joining rules. As for the joining rules just discussed, the following rules describe ligature behavior of Arabic letters in terms of their display (visual) order.

In the ligature rules, if X and Y refer to characters, then various glyph types representing combinations of these characters are referred to as shown in *Table 8-7*.

Table 8-7. Arabic Ligature Notation

Symbol	Description
$(X\text{-}Y)_n$	Nominal ligature glyph form representing a combination of an X_r form and a Y_l form
$(X\text{-}Y)_r$	Right-joining ligature glyph form representing a combination of an X_r form and a Y_m form
$(X\text{-}Y)_l$	Left-joining ligature glyph form representing a combination of an X_m form and a Y_l form
$(X\text{-}Y)_m$	Dual-joining (medial) ligature glyph form representing a combination of an X_m form and a Y_m form

L1 ***Transparent characters do not affect the ligating behavior of base (nontransparent) characters. For example:***

$$\mathsf{ALEF_r + FATHA_n + LAM_l \rightarrow (LAM\text{-}ALEF)_n + FATHA_n}$$

L2 ***Any sequence with $ALEF_r$ on the left and LAM_m on the right will form the ligature $(LAM\text{-}ALEF)_r$. For example:***

ﻟ + ﻠ → ﻼ (not ﷲ)

L3 ***Any sequence with $ALEF_r$ on the left and LAM_l on the right will form the ligature $(LAM\text{-}ALEF)_n$. For example:***

ﻟ + ﺍ → ﻻ (not ﷲ)

Optional Features. Many other ligatures and contextual forms are optional, depending on the font and application. Some of these presentation forms are encoded in the ranges FB50..FDFB and FE70..FEFE. However, these forms should *not* be used in general interchange. Moreover, it is not expected that every Arabic font will contain all of these forms, nor that these forms will include all presentation forms used by every font.

More sophisticated rendering systems will use additional shaping and placement. For example, contextual placement of the nonspacing vowels such as *fatha* will provide better appearance. The justification of Arabic tends to stretch words instead of adding width to spaces. Basic stretching can be done by inserting *tatweel* between characters shaped by rules R2, R4, R5, R6, L2, and L3; the best places for inserting tatweel will depend on the font and rendering software. More powerful systems will choose different shapes for characters such as *kaf* to fill the space in justification.

Arabic Joining Groups

The Arabic characters with the property values Joining_Type=Dual_Joining and Joining_Type=Right_Joining can each be subdivided into shaping groups, based on the

behavior of their letter skeletons when shaped in context. The Unicode character property that specifies these groups is called Joining_Group.

The Joining_Type and Joining_Group values for all Arabic characters are explicitly specified in ArabicShaping.txt in the Unicode Character Database. For convenience in reference, the Joining_Type values are extracted and listed in DerivedJoiningType.txt and the Joining_Group values are extracted and listed in DerivedJoiningGroup.txt.

Dual-Joining. *Table 8-8* exemplifies dual-joining Arabic characters and illustrates the forms taken by the letter skeletons and their diacritical marks in context. Dual-joining characters have four distinct forms, for isolated, final, medial, and initial contexts, respectively. The name for each joining group is based on the name of a representative letter that is used to illustrate the shaping behavior. All other Arabic characters are merely variations on these basic shapes, with diacritics added, removed, moved, or replaced. For instance, the BEH joining group applies not only to U+0628 ARABIC LETTER BEH, which has a single dot below the skeleton, but also to U+062A ARABIC LETTER TEH, which has two dots above the skeleton, and to U+062B ARABIC LETTER THEH, which has three dots above the skeleton, as well as to the Persian and Urdu letter U+067E ARABIC LETTER PEH, which has three dots below the skeleton. The joining groups in the table are organized by shape and not by standard Arabic alphabetical order. Note that characters in some joining groups have dots in some contextual forms, but not others. These joining groups include NYA, FARSI YEH, and BURUSHASKI YEH BARREE.

Table 8-8. Dual-Joining Arabic Characters

Joining Group	X_n	X_r	X_m	X_l	Notes
BEH	ب	ب	بـ	بـ	Includes TEH and THEH.
NOON	ن	ن	نـ	نـ	
NYA	ڽ	ڽ	ڽـ	ڽـ	Jawi NYA.
YEH	ي	ي	يـ	يـ	Includes ALEF MAKSURA.
FARSI YEH	ی	ی	یـ	یـ	
BURUSHASKI YEH BARREE	ے	ے	ۓـ	ۓـ	Left-connecting form of YEH BARREE
HAH	ح	ح	حـ	حـ	Includes KHAH and JEEM.
SEEN	س	س	سـ	سـ	Includes SHEEN.
SAD	ص	ص	صـ	صـ	Includes DAD.
TAH	ط	ط	طـ	طـ	Includes ZAH.
AIN	ع	ع	عـ	عـ	Includes GHAIN.
FEH	ف	ف	فـ	فـ	

Table 8-8. Dual-Joining Arabic Characters (Continued)

Joining Group	X_n	X_r	X_m	X_l	Notes
QAF	ق	ق	قـ	قـ	
MEEM	م	م	مـ	مـ	
HEH	ه	ه	هـ	هـ	
KNOTTED HEH	ھ	ھ	ھـ	ھـ	
HEH GOAL	ہ	ہ	ہـ	ہـ	Includes HAMZA ON HEH GOAL.
KAF	ك	ك	كـ	كـ	
SWASH KAF	ک	ک	کـ	کـ	
GAF	گ	گ	گـ	گـ	
LAM	ل	ل	لـ	لـ	

Right-Joining. *Table 8-9* exemplifies right-joining Arabic characters, illustrating the forms they take in context. Right-joining characters have only two distinct forms, for isolated and final contexts, respectively.

Table 8-9. Right-Joining Arabic Characters

Joining Group	X_n	X_r	Notes
ALEF	ا	ـا	
WAW	و	ـو	
DAL	د	ـد	Includes THAL.
REH	ر	ـر	Includes ZAIN.
TEH MARBUTA	ة	ـة	Includes HAMZA ON HEH.
TEH MARBUTA GOAL	ۃ	ـۃ	
YEH WITH TAIL	ی	ـی	
YEH BARREE	ے	ـے	
ROHINGYA YEH		ـۄ	Isolated form does not occur.

In some cases, characters occur only at the end of words in correct spelling; they are called *trailing characters*. Examples include TEH MARBUTA and DAMMATAN. When trailing characters are joining (such as TEH MARBUTA), they are classified as right-joining, even when similarly shaped characters are dual-joining.

***Letter* heh.** In the case of U+0647 ARABIC LETTER HEH, the glyph ﻫ is shown in the code charts. This form is often used to reduce the chance of misidentifying *heh* as U+0665 ARABIC-INDIC DIGIT FIVE, which has a very similar shape. The isolate forms of U+0647 ARABIC LETTER HEH and U+06C1 ARABIC LETTER HEH GOAL both look like U+06D5 ARABIC LETTER AE.

***Letter* yeh.** There are many complications in the shaping of the Arabic letter *yeh*. These complications have led to the encoding of several different characters for *yeh* in the Unicode Standard, as well as the definition of several different joining groups involving *yeh*. The relationships between those characters and joining groups for *yeh* are explained here.

U+06CC ARABIC LETTER FARSI YEH is used in Persian, Urdu, Pashto, Azerbaijani, Kurdish, and various minority languages written in the Arabic script, and also Koranic Arabic. It behaves differently from most Arabic letters, in a way surprising to native Arabic language speakers. The letter has two horizontal dots below the skeleton in initial and medial forms, but no dots in final and isolated forms. Compared to the two Arabic language *yeh* forms, FARSI YEH is exactly like U+0649 ARABIC LETTER ALEF MAKSURA in final and isolated forms, but exactly like U+064A ARABIC LETTER YEH in initial and medial forms, as shown in *Table 8-10*.

Table 8-10. Forms of the Arabic Letter *yeh*

Character	Joining Group	X_n	X_r	X_m	X_l
U+0649 ALEF MAKSURA	YEH	ى	ى	ـى	ـى
U+064A YEH	YEH	ي	ي	ـيـ	ـيـ
U+06CC FARSI YEH	FARSI YEH	ی	ی	ـیـ	ـیـ
U+0777 YEH WITH DIGIT FOUR BELOW	YEH	ۍ	ۍ	ـۍـ	ـۍ
U+0620 KASHMIRI YEH	YEH	ؠ	ؠ	ـؠـ	ـؠ
U+06D2 YEH BARREE	YEH BARREE	ے	ے		
U+077A YEH BARREE WITH DIGIT TWO ABOVE	BURUSHASKI YEH BARREE	ۓ	ۓ	ـۓـ	ـۓ
U+08AC ROHINGYA YEH	ROHINGYA YEH		ۼ		

Other characters of the joining group FARSI YEH follow the same pattern. These YEH forms appear with two dots aligned horizontally below them in initial and medial forms, but with no dots below them in final and isolated forms. Characters with the joining group YEH behave in a different manner. Just as U+064A ARABIC LETTER YEH retains two dots below in all contextual forms, other characters in the joining group YEH retain whatever mark appears below their isolated form in all other contexts. For example, U+0777 ARABIC LETTER FARSI YEH WITH EXTENDED ARABIC-INDIC DIGIT FOUR BELOW carries an Urdu-style

digit four as a diacritic below the *yeh* skeleton, and retains that diacritic in all positions, as shown in the fourth row of *Table 8-10*. Note that the joining group cannot always be derived from the character name alone. The complete list of characters with the joining group YEH or FARSI YEH is available in ArabicShaping.txt in the Unicode Character Database.

In the orthographies of Arabic and Persian, the *yeh barree* has always been treated as a stylistic variant of *yeh* in final and isolated positions. When the Perso-Arabic writing system was adapted and extended for use with the Urdu language, *yeh barree* was adopted as a distinct letter to accommodate the richer vowel repertoire of Urdu. South Asian languages such as Urdu and Kashmiri use *yeh barree* to represent the /e/ vowel. This contrasts with the /i/ vowel, which is usually represented in those languages by U+06CC ARABIC LETTER FARSI YEH. The encoded character U+06D2 ARABIC LETTER YEH BARREE is classified as a right-joining character, as shown in *Table 8-10*. On that basis, when the /e/ vowel needs to be represented in initial or medial positions with a *yeh* shape in such languages, one should use U+06CC ARABIC LETTER FARSI YEH. In the unusual circumstances where one wishes to distinctly represent the /e/ vowel in word-initial or word-medial positions, a higher level protocol should be used.

For the Burushaski language, two characters that take the form of *yeh barree* with a diacritic, U+077A ARABIC LETTER YEH BARREE WITH EXTENDED ARABIC-INDIC DIGIT TWO ABOVE and U+077B ARABIC LETTER YEH BARREE WITH EXTENDED ARABIC-INDIC DIGIT THREE ABOVE, are classified as dual-joining. These characters have a separate joining group called BURUSHASKI YEH BARREE, as shown for U+077A in the last row of *Table 8-10*.

U+0620 ARABIC LETTER KASHMIRI YEH is used in Kashmiri text to indicate that the preceding consonantal sound is palatalized. The letter has the form of a *yeh* with a diacritic small circle below. It has the YEH joining group, with the shapes as shown in the fifth row of *Table 8-10*. However, when Kashmiri is written in Nastaliq style, the final and isolated forms of *kashmiri yeh* usually appear as truncated *yeh* shapes (ﯼ) without the diacritic ring.

U+08AC ARABIC LETTER ROHINGYA YEH is used in the Arabic orthography for the Rohingya language of Myanmar. It represents a *medial ya*, corresponding to the use of U+103B MYANMAR CONSONANT SIGN MEDIAL YA in the Myanmar script. It is a right-joining letter, but never occurs in isolated form. It only occurs after certain consonants, forming a conjunct letter with those consonants.

Combining Hamza Above. U+0654 ARABIC HAMZA ABOVE is intended both for the representation of *hamza* semantics in combination with certain Arabic letters, and as a diacritic mark occasionally used in combinations to derive extended Arabic letters. There are a number of complications regarding its use, which interact with the rules for the rendering of Arabic letter *yeh* and which result from the need to keep Unicode normalization stable.

U+0654 ARABIC HAMZA ABOVE should not be used with U+0649 ARABIC LETTER ALEF MAKSURA. Instead, the precomposed U+0626 ARABIC LETTER YEH WITH HAMZA ABOVE should be used to represent a *yeh*-shaped base with no dots in any positional form, and with a *hamza* above. Because U+0626 is canonically equivalent to the sequence <U+064A ARABIC LETTER YEH, U+0654 ARABIC HAMZA ABOVE>, when U+0654 is applied to U+064A ARABIC LETTER YEH, the *yeh* should lose its dots in all positional forms, even though *yeh* retains its dots when combined with other marks.

A separate, non-decomposable character, U+08A8 ARABIC LETTER YEH WITH TWO DOTS BELOW AND HAMZA ABOVE, is used to represent a *yeh*-shaped base with a *hamza* above, but with retention of dots in all positions. This letter is used in the Fulfulde language in Cameroun, to represent a palatal implosive.

In most other cases when a *hamza* is needed as a mark above for an extended Arabic letter, U+0654 ARABIC HAMZA ABOVE can be freely used in combination with basic Arabic letters.

Two exceptions are the extended Arabic letters U+0681 ARABIC LETTER HAH WITH HAMZA ABOVE and U+076C ARABIC LETTER REH WITH HAMZA ABOVE, where the *hamza* mark is functioning as an *ijam* (diacritic), rather than as a normal *hamza*. In those two cases, the extended Arabic letters have no canonical decompositions; consequently, the preference is to use those two precomposed forms, rather than applying U+0654 ARABIC HAMZA ABOVE to *hah* or to *reh*, respectively.

These interactions between various letters and the *hamza* are summarized in *Table 8-11*.

Table 8-11. Arabic Letters With Hamza Above

Code Point	Name	Decomposition
0623	alef with hamza above	0627 0654
0624	waw with hamza above	0648 0654
0626	yeh with hamza above	064A 0654
06C2	heh goal with hamza above	06C1 0654
06D3	yeh barree with hamza above	06D2 0654
0681	hah with hamza above	None
076C	reh with hamza above	None
08A8	yeh with 2 dots below and hamza above	None

The first five entries in *Table 8-11* show the cases where the *hamza above* can be freely used, and where there is a canonical equivalence to the precomposed characters. The last three entries show the exceptions, where use of the *hamza above* is inappropriate, and where only the precomposed characters should be used.

Jawi. U+06BD ARABIC LETTER NOON WITH THREE DOTS ABOVE is used for Jawi, which is Malay written using the Arabic script. Malay users know the character as *Jawi Nya*. Contrary to what is suggested by its Unicode character name, U+06BD displays with the three dots *below* the letter pointing downward when it is in the initial or medial position, making it look exactly like the initial and medial forms of U+067E ARABIC LETTER PEH. This is done to avoid confusion with U+062B ARABIC LETTER THEH, which appears in words of Arabic origin, and which has the same base letter shapes in initial or medial position, but with three dots above in all positions.

Kurdish. The Kurdish language is written in several different orthographies, which use either the Latin, Cyrillic, or Arabic scripts. When written using the Arabic script, Kurdish uses a number of extended Arabic letters, for an alphabet known as Soraní. Some of those extensions are shared with Persian, Urdu, or other languages: for example, U+06C6 ARABIC LETTER OE, which represents the Kurdish vowel [o]. Soraní also makes other unusual adaptations of the Arabic script, including the use of a digraph *waw+waw* to represent the long Kurdish vowel [u:]. That digraph is represented by a sequence of two characters, <U+0648 ARABIC LETTER WAW, U+0648 ARABIC LETTER WAW>.

Among the extended Arabic characters used exclusively for Soraní are U+0695 ARABIC LETTER REH WITH SMALL V BELOW (for the Kurdish *flap r*) and U+06B5 ARABIC LETTER LAM WITH SMALL V (for the Kurdish *velarized l*).

The Unicode Standard also includes several extended Arabic characters whose origin was to represent dialectal or other poorly attested alternative forms of the Soraní alphabet extensions. U+0692 ARABIC LETTER REH WITH SMALL V is a dialectal variant of U+0695 which places the *small v* diacritic above the letter rather than below it. U+0694 is another variant of U+0695. U+06B6 and U+06B7 are poorly attested variants of U+06B5, and U+06CA is a poorly attested variant of U+06C6. None of these alternative forms is required (or desired) for a regular implementation of the Kurdish Soraní orthography.

Arabic Supplement: U+0750–U+077F

The Arabic Supplement block contains additional extended Arabic letters for the languages used in Northern and Western Africa, such as Fulfulde, Hausa, Songhoy, and Wolof. In the second half of the twentieth century, the use of the Arabic script was actively promoted for these languages. This block also contains a number of letters used for the Khowar, Torwali, and Burushaski languages, spoken primarily in Pakistan. Characters used for other languages are annotated in the character names list. Additional vowel marks used with these languages are found in the main Arabic block.

Marwari. U+076A ARABIC LETTER LAM WITH BAR is used to represent a flapped retroflexed lateral in the Marwari language in southern Pakistan. It has also been suggested for use in the Gawri language of northern Pakistan but it is unclear how widely it has been adopted there. Contextual shaping for this character is similar to that of U+0644 ARABIC LETTER LAM, including the requirement to form ligatures with ALEF and related characters.

Arabic Extended-A: U+08A0–U+08FF

The Arabic Extended-A block contains additional Arabic letters and vowel signs for use by a number of African languages from Chad, Senegal, Guinea, and Cameroon, and for languages of the Philippines. It also contains extended letters, vowel signs, and tone marks used by the Rohingya Fonna writing system for the Rohingya language in Myanmar, as well as several additional Koranic annotation signs.

Arabic Presentation Forms-A: U+FB50–U+FDFF

This block contains a list of presentation forms (glyphs) encoded as characters for compatibility. As with most other compatibility encodings, these characters have a preferred encoding that makes use of noncompatibility characters.

The presentation forms in this block consist of contextual (positional) variants of Extended Arabic letters, contextual variants of Arabic letter ligatures, spacing forms of Arabic diacritic combinations, contextual variants of certain Arabic letter/diacritic combinations, and Arabic phrase ligatures. The ligatures include a large set of presentation forms. However, the set of ligatures appropriate for any given Arabic font will generally not match this set precisely. Fonts will often include only a subset of these glyphs, and they may also include glyphs outside of this set. These glyphs are generally not accessible as characters and are used only by rendering engines.

Ornate Parentheses. The alternative, ornate forms of parentheses (U+FD3E ORNATE LEFT PARENTHESIS and U+FD3F ORNATE RIGHT PARENTHESIS) for use with the Arabic script are considered traditional Arabic punctuation, rather than compatibility characters. These ornate parentheses are exceptional in rendering in bidirectional text; for legacy reasons, they do not have the Bidi_Mirrored property. Thus, unlike other parentheses, they do not automatically mirror when rendered in a bidirectional context.

Nuktas. Various patterns of single or multiple dots or other small marks are used diacritically to extend the core Arabic set of letters to represent additional sounds in other languages written with the Arabic script. Such dot patterns are known as *nuktas*. In the Unicode Standard, extended Arabic characters with nuktas are simply encoded as fully-formed base characters. However, there is an occasional need in pedagogical materials about the Arabic script to exhibit the various nuktas in isolation. The range of characters U+FBB2..U+FBC1 provides a set of symbols for this purpose. These are ordinary, spacing symbols with right-to-left directionality. They are *not* combining marks, and are not intended for the construction of new Arabic letters by use in combining character sequences. Any use in juxtaposition with an Arabic letter skeleton is undefined.

The Arabic nukta symbols do not partake of any Arabic shaping behavior. For clarity in display, those with the names including the word "above" should have glyphs that render high above the baseline, and those with names including "below" should be at or below the baseline.

Arabic Presentation Forms-B: U+FE70–U+FEFF

This block contains additional Arabic presentation forms consisting of spacing or *tatweel* forms of Arabic diacritics, contextual variants of primary Arabic letters, and the obligatory LAM-ALEF ligature. They are included here for compatibility with preexisting standards and legacy implementations that use these forms as characters. They can be replaced by letters from the Arabic block (U+0600..U+06FF). Implementations can handle contextual glyph shaping by rendering rules when accessing glyphs from fonts, rather than by encoding contextual shapes as characters.

Spacing and Tatweel Forms of Arabic Diacritics. For compatibility with certain implementations, a set of spacing forms of the Arabic diacritics is provided here. The tatweel forms are combinations of the joining connector tatweel and a diacritic.

Zero Width No-Break Space. This character (U+FEFF), which is not an Arabic presentation form, is described in *Section 16.8, Specials*.

8.3 Syriac

Syriac: U+0700–U+074F

Syriac Language. The Syriac language belongs to the Aramaic branch of the Semitic family of languages. The earliest datable Syriac writing dates from the year 6 CE. Syriac is the active liturgical language of many communities in the Middle East (Syrian Orthodox, Assyrian, Maronite, Syrian Catholic, and Chaldaean) and Southeast India (Syro-Malabar and Syro-Malankara). It is also the native language of a considerable population in these communities.

Syriac is divided into two dialects. West Syriac is used by the Syrian Orthodox, Maronites, and Syrian Catholics. East Syriac is used by the Assyrians (that is, Ancient Church of the East) and Chaldaeans. The two dialects are very similar and have almost no differences in grammar and vocabulary. They differ in pronunciation and use different dialectal forms of the Syriac script.

Languages Using the Syriac Script. A number of modern languages and dialects employ the Syriac script in one form or another. They include the following:

1. *Literary Syriac*. The primary usage of Syriac script.
2. *Neo-Aramaic dialects*. The Syriac script is widely used for modern Aramaic languages, next to Hebrew, Cyrillic, and Latin. A number of Eastern Modern Aramaic dialects known as *Swadaya* (also called vernacular Syriac, modern Syriac, modern Assyrian, and so on, and spoken mostly by the Assyrians and Chaldaeans of Iraq, Turkey, and Iran) and the Central Aramaic dialect, *Turoyo* (spoken mostly by the Syrian Orthodox of the Tur Abdin region in southeast Turkey), belong to this category of languages.
3. *Garshuni* (Arabic written in the Syriac script). It is currently used for writing Arabic liturgical texts by Syriac-speaking Christians. Garshuni employs the Arabic set of vowels and overstrike marks.

4. *Christian Palestinian Aramaic* (also known as Palestinian Syriac). This dialect is no longer spoken.
5. *Other languages*. The Syriac script was used in various historical periods for writing Armenian and some Persian dialects. Syriac speakers employed it for writing Arabic, Ottoman Turkish, and Malayalam. Six special characters used for Persian and Sogdian were added in Version 4.0 of the Unicode Standard.

Shaping. The Syriac script is cursive and has shaping rules that are similar to those for Arabic. The Unicode Standard does not include any presentation form characters for Syriac.

Directionality. The Syriac script is written from right to left. Conformant implementations of Syriac script must use the Unicode Bidirectional Algorithm (see Unicode Standard Annex #9, "Unicode Bidirectional Algorithm").

Syriac Type Styles. Syriac texts employ several type styles. Because all type styles use the same Syriac characters, even though their shapes vary to some extent, the Unicode Standard encodes only a single Syriac script.

1. *Estrangela type style*. Estrangela (a word derived from Greek *strongulos,* meaning "rounded") is the oldest type style. Ancient manuscripts use this writing style exclusively. Estrangela is used today in West and East Syriac texts for writing headers, titles, and subtitles. It is the current standard in writing Syriac texts in Western scholarship.
2. *Serto or West Syriac type style*. This type style is the most cursive of all Syriac type styles. It emerged around the eighth century and is used today in West Syriac texts, Turoyo (Central Neo-Aramaic), and Garshuni.
3. *East Syriac type style*. Its early features appear as early as the sixth century; it developed into its own type style by the twelfth or thirteenth century. This type style is used today for writing East Syriac texts as well as Swadaya (Eastern Neo-Aramaic). It is also used today in West Syriac texts for headers, titles, and subtitles alongside the Estrangela type style.
4. *Christian Palestinian Aramaic*. Manuscripts of this dialect employ a script that is akin to Estrangela. It can be considered a subcategory of Estrangela.

The Unicode Standard provides for usage of the type styles mentioned above. It also accommodates letters and diacritics used in Neo-Aramaic, Christian Palestinian Aramaic, Garshuni, Persian, and Sogdian languages. *Examples are supplied in the Serto type style, except where otherwise noted.*

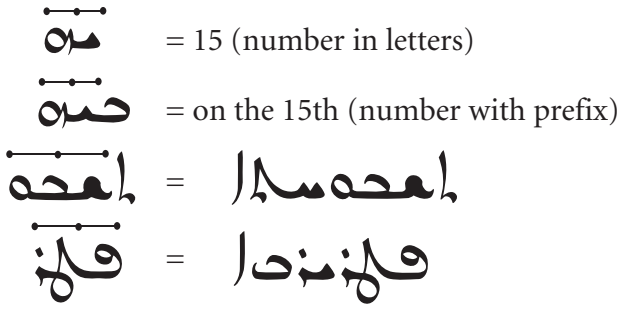
Character Names. Character names follow the East Syriac convention for naming the letters of the alphabet. Diacritical points use a descriptive naming—for example, SYRIAC DOT ABOVE.

Syriac Abbreviation Mark. U+070F SYRIAC ABBREVIATION MARK (SAM) is a zero-width formatting code that has no effect on the shaping process of Syriac characters. The SAM specifies the beginning point of a *Syriac abbreviation*, which is a line drawn horizontally above one or more characters, at the end of a word or of a group of characters followed by a character other than a Syriac letter or diacritic mark. A Syriac abbreviation may contain Syriac diacritics.

Ideally, the Syriac abbreviation is rendered by a line that has a dot at each end and the center, as shown in the examples. While not preferable, it has become acceptable for computers to render the Syriac abbreviation as a line without the dots. The line is acceptable for the presentation of Syriac in plain text, but the presence of dots is recommended in liturgical texts.

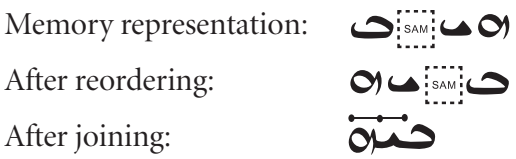
The Syriac abbreviation is used for letter numbers and contractions. A Syriac abbreviation generally extends from the last tall character in the word until the end of the word. A common exception to this rule is found with letter numbers that are preceded by a preposition character, as seen in *Figure 8-7*.

Figure 8-7. Syriac Abbreviation



A SAM is placed before the character where the abbreviation begins. The Syriac abbreviation begins over the character following the SAM and continues until the end of the word. Use of the SAM is demonstrated in *Figure 8-8*.

Figure 8-8. Use of SAM



Note: Modern East Syriac texts employ a punctuation mark for contractions of this sort.

Ligatures and Combining Characters. Only one ligature is included in the Syriac block: U+071E SYRIAC LETTER YUDH HE. This combination is used as a unique character in the same manner as an "æ" ligature. A number of combining diacritics unique to Syriac are encoded, but combining characters from other blocks are also used, especially from the Arabic block.

Diacritic Marks and Vowels. The function of the diacritic marks varies. They indicate vowels (as in Arabic and Hebrew), mark grammatical attributes (for example, verb versus noun, interjection), or guide the reader in the pronunciation and/or reading of the given text.

> "The reader of the average Syriac manuscript or book is confronted with a bewildering profusion of points. They are large, of medium size and small, arranged singly or in twos and threes, placed above the word, below it, or upon the line."

There are two vocalization systems. The first, attributed to Jacob of Edessa (633–708 CE), utilizes letters derived from Greek that are placed above (or below) the characters they modify. The second is the more ancient dotted system, which employs dots in various shapes and locations to indicate vowels. East Syriac texts exclusively employ the dotted system, whereas West Syriac texts (especially later ones and in modern times) employ a mixture of the two systems.

Diacritic marks are nonspacing and are normally centered above or below the character. Exceptions to this rule follow:

1. U+0741 SYRIAC QUSHSHAYA and U+0742 SYRIAC RUKKAKHA are used only with the letters *beth*, *gamal* (in its Syriac and Garshuni forms), *dalath*, *kaph*, *pe,* and *taw*.
   - The *qushshaya* indicates that the letter is pronounced hard and unaspirated.
   - The *rukkakha* indicates that the letter is pronounced soft and aspirated. When the *rukkakha* is used in conjunction with the *dalath*, it is printed slightly to the right of the *dalath*'s dot below.
2. In Modern Syriac usage, when a word contains a *rish* and a *seyame*, the dot of the *rish* and the *seyame* are replaced by a *rish* with two dots above it.
3. The *feminine dot* is usually placed to the left of a final *taw*.

Punctuation. Most punctuation marks used with Syriac are found in the Latin-1 and Arabic blocks. The other marks are encoded in this block.

Digits. Modern Syriac employs European numerals, as does Hebrew. The ordering of digits follows the same scheme as in Hebrew.

Harklean Marks. The Harklean marks are used in the Harklean translation of the New Testament. U+070B SYRIAC HARKLEAN OBELUS and U+070D SYRIAC HARKLEAN ASTERISCUS mark the beginning of a phrase, word, or morpheme that has a marginal note. U+070C SYRIAC HARKLEAN METOBELUS marks the end of such sections.

Dalath and Rish. Prior to the development of pointing, early Syriac texts did not distinguish between a *dalath* and a *rish*, which are distinguished in later periods with a dot below the former and a dot above the latter. Unicode provides U+0716 SYRIAC LETTER DOTLESS DALATH RISH as an ambiguous character.

Semkath. Unlike other letters, the joining mechanism of *semkath* varies through the course of history from right-joining to dual-joining. It is necessary to enter a U+200C ZERO WIDTH NON-JOINER character after the *semkath* to obtain the right-joining form where required. Two common variants of this character exist: U+0723 SYRIAC LETTER SEMKATH and U+0724 SYRIAC LETTER FINAL SEMKATH. They occur interchangeably in the same document, similar to the case of Greek sigma.

Vowel Marks. The so-called Greek vowels may be used above or below letters. As West Syriac texts employ a mixture of the Greek and dotted systems, both versions are accounted for here.

Miscellaneous Diacritics. Miscellaneous general diacritics are used in Syriac text. Their usage is explained in *Table 8-12*.

Use of Characters of the Arabic Block. Syriac makes use of several characters from the Arabic block, including U+0640 ARABIC TATWEEL. Modern texts use U+060C ARABIC COMMA, U+061B ARABIC SEMICOLON, and U+061F ARABIC QUESTION MARK. The *shadda* (U+0651) is also used in the core part of literary Syriac on top of a *waw* in the word "O". Arabic *harakat* are used in Garshuni to indicate the corresponding Arabic vowels and diacritics.

Syriac Shaping

Minimum Rendering Requirements. Rendering requirements for Syriac are similar to those for Arabic. The remainder of this section specifies a minimum set of rules that provides legible Syriac joining and ligature substitution behavior.

Joining Types. Each Syriac letter must be depicted by one of a number of possible contextual glyph forms. The appropriate form is determined on the basis of the cursive joining behavior of that character as it interacts with the cursive joining behavior of adjacent char-

Table 8-12. Miscellaneous Syriac Diacritic Use

Code Points	Use
U+0303, U+0330	These are used in Swadaya to indicate letters not found in Syriac.
U+0304, U+0320	These are used for various purposes ranging from phonological to grammatical to orthographic markers.
U+0307, U+0323	These points are used for various purposes—grammatical, phonological, and otherwise. They differ typographically and semantically from the *qushshaya*, *rukkakha* points, and the dotted vowel points.
U+0308	This is the plural marker. It is also used in Garshuni for the Arabic *teh marbuta*.
U+030A, U+0325	These are two other forms for the indication of *qushshaya* and *rukkakha*. They are used interchangeably with U+0741 SYRIAC QUSHSHAYA and U+0742 SYRIAC RUKKAKHA, especially in West Syriac grammar books.
U+0324	This diacritic mark is found in ancient manuscripts. It has a grammatical and phonological function.
U+032D	This is one of the *digit markers*.
U+032E	This is a mark used in late and modern East Syriac texts as well as in Swadaya to indicate a fricative *pe*.

acters. The basic joining types are identical to those specified for the Arabic script. However, there are additional contextual rules which govern the shaping of U+0710 SYRIAC LETTER ALAPH in final position. The additional glyph types associated with final *alaph* are listed in *Table 8-13*.

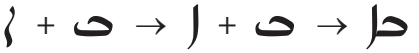
Table 8-13. Syriac Final Alaph Glyph Types

Glyph Type	Description
A_{fj}	Final joining (alaph only)
A_{fn}	Final non-joining *except* following dalath and rish (alaph only)
A_{fx}	Final non-joining following dalath and rish (alaph only)

In the following rules, *alaph* refers to U+0710 SYRIAC LETTER ALAPH, which has Joining_Group=Alaph.

These rules are intended to augment joining rules for Syriac which would otherwise parallel the joining rules specified for Arabic in *Section 8.2, Arabic*. Characters with Joining_Type=Transparent are skipped over when applying the Syriac rules for shaping of *alaph*. In other words, the Syriac parallel for Arabic joining rule R1 would take precedence over the *alaph* joining rules.

S1 ***An* alaph *that has a left-joining character to its right and a non-joining character to its left will take the form of A_{fj}.***



S2 ***An* alaph *that has a non-left-joining character to its right, except for a character with Joining_Group=Dalath_Rish, and a non-joining character to its left will take the form of A_{fn}.***

S3 ***An* alaph *that has a character with Joining_Group=Dalath_Rish to its right and a non-joining character to its left will take the form of A_{fx}.***

The example in rule S3 is shown in the East Syriac font style.

Syriac Character Joining Groups. Syriac characters can be subdivided into shaping groups, based on the behavior of their letter skeletons when shaped in context. The Unicode character property that specifies these groups is called Joining_Group, and is specified in ArabicShaping.txt in the Unicode Character Database. It is described in the subsection on character joining groups in *Section 8.2, Arabic*.

Table 8-14 exemplifies dual-joining Syriac characters and illustrates the forms taken by the letter skeletons in context. This table and the subsequent table use the Serto (West Syriac) font style, whereas the Unicode code charts are in the Estrangela font style.

Table 8-14. Dual-Joining Syriac Characters

Joining Group	X_n	X_r	X_m	X_l	Notes
Beth	ܒ	ܒ	ܒ	ܒ	Includes PERSIAN BHETH
Gamal	ܓ	ܓ	ܓ	ܓ	Includes GAMAL GARSHUNI and PERSIAN GHAMAL
Heth	ܚ	ܚ	ܚ	ܚ	
Teth	ܛ	ܛ	ܛ	ܛ	Includes TETH GARSHUNI
Yudh	ܝ	ܝ	ܝ	ܝ	
Kaph	ܟ	ܟ	ܟ	ܟ	
Khaph	ܟ̰	ܟ̰	ܟ̰	ܟ̰	Sogdian
Lamadh	ܠ	ܠ	ܠ	ܠ	
Mim	ܡ	ܡ	ܡ	ܡ	
Nun	ܢ	ܢ	ܢ	ܢ	
Semkath	ܣ	ܣ	ܣ	ܣ	
Final_Semkath	ܤ	ܤ	ܣ	ܣ	
E	ܥ	ܥ	ܥ	ܥ	
Pe	ܦ	ܦ	ܦ	ܦ	
Reversed_Pe	ܦ̮	ܦ̮	ܦ̮	ܦ̮	
Fe	ܦ̰	ܦ̰	ܦ̰	ܦ̰	Sogdian
Qaph	ܩ	ܩ	ܩ	ܩ	
Shin	ܫ	ܫ	ܫ	ܫ	

Table 8-15 exemplifies right-joining Syriac characters, illustrating the forms they take in context. Right-joining characters have only two distinct forms, for isolated and final contexts, respectively.

Table 8-15. Right-Joining Syriac Characters

Joining Group	X_n	X_r	Notes
Dalath_Rish	ܕ	ܪ	Includes RISH, DOTLESS DALATH RISH, and PERSIAN DHALATH
He	ܗ	ܗ	
Syriac_Waw	ܘ	ܘ	
Zain	ܙ	ܙ	
Zhain	ܟ	ܟ	Sogdian
Yudh_He	ܝܗ	ܝܗ	
Sadhe	ܨ	ܨ	
Taw	ܬ	ܬ	

U+0710 SYRIAC LETTER ALAPH has the Joining_Group=Alaph and is a right-joining character. However, as specified above in rules S1, S2, and S3, its glyph is subject to additional contextual shaping. *Table 8-16* illustrates all of the glyph forms for *alaph* in each of the three major Syriac type styles.

Table 8-16. Syriac Alaph Glyph Forms

Type Style	X_n	X_r	A_{fj}	A_{fn}	A_{fx}
Estrangela	ܐ	ܐ	ܐ	ܐ	ܐ
Serto (West Syriac)	ܐ	ܐ	ܐ	ܐ	ܐ
East Syriac	ܐ	ܐ	ܐ	ܐ	ܐ

Ligature Classes. As in other scripts, ligatures in Syriac vary depending on the font style. *Table 8-17* identifies the principal valid ligatures for each font style. When applicable, these ligatures are obligatory, unless denoted with an asterisk (*).

Table 8-17. Syriac Ligatures

Characters	Estrangela	Serto (West Syriac)	East Syriac	Sources
ALAPH LAMADH	N/A	Dual-joining	N/A	Beth Gazo
GAMAL LAMADH	N/A	Dual-joining*	N/A	Armalah
GAMAL E	N/A	Dual-joining*	N/A	Armalah
HE YUDH	N/A	N/A	Right-joining*	Qdom
YUDH TAW	N/A	Right-joining*	N/A	Armalah*
KAPH LAMADH	N/A	Dual-joining*	N/A	Shhimo
KAPH TAW	N/A	Right-joining*	N/A	Armalah
LAMADH SPACE ALAPH	N/A	Right-joining*	N/A	Nomocanon
LAMADH ALAPH	Right-joining*	Right-joining	Right-joining*	BFBS
LAMADH LAMADH	N/A	Dual-joining*	N/A	Shhimo
NUN ALAPH	N/A	Right-joining*	N/A	Shhimo
SEMAKATH TETH	N/A	Dual-joining*	N/A	Qurobo
SADHE NUN	Right-joining*	Right-joining*	Right-joining*	Mushhotho

Table 8-17. Syriac Ligatures (Continued)

Characters	Estrangela	Serto (West Syriac)	East Syriac	Sources
RISH SEYAME	Right-joining	Right-joining	Right-joining	BFBS
TAW ALAPH	Right-joining*	N/A	Right-joining*	Qdom
TAW YUDH	N/A	N/A	Right-joining*	

8.4 Samaritan

Samaritan: U+0800–U+083F

The Samaritan script is used today by small Samaritan communities in Israel and the Palestinian Territories to write the Samaritan Hebrew and Samaritan Aramaic languages, primarily for religious purposes. The Samaritan religion is related to an early form of Judaism, but the Samaritans did not leave Palestine during the Babylonian exile, so the script evolved from the linear Old Hebrew script, most likely directly descended from Phoenician (see *Section 14.10, Phoenician*). In contrast, the more recent square Hebrew script associated with Judaism derives from the Imperial Aramaic script (see *Section 14.11, Imperial Aramaic*) used widely in the region during and after the Babylonian exile, and thus well-known to educated Hebrew speakers of that time.

Like the Phoenician and Hebrew scripts, Samaritan has 22 consonant letters. The consonant letters do not form ligatures, nor do they have explicit final forms as some Hebrew consonants do.

Directionality. The Samaritan script is written from right to left. Conformant implementations of Samaritan script must use the Unicode Bidirectional Algorithm. For more information, see Unicode Standard Annex #9, "Unicode Bidirectional Algorithm."

Vowel Signs. Vowel signs are optional in Samaritan, just as points are optional in Hebrew. Combining marks are used for vowels that follow a consonant, and are rendered above and to the left of the base consonant. With the exception of *o* and *short a*, vowels may have up to three lengths (normal, long, and overlong), which are distinguished by the size of the corresponding vowel sign. *Sukun* is centered above the corresponding base consonant and indicates that no vowel follows the consonant.

Two vowels, *i* and *short a*, may occur in a word-initial position preceding any consonant. In this case, the separate spacing versions U+0828 SAMARITAN MODIFIER LETTER I and U+0824 SAMARITAN MODIFIER LETTER SHORT A should be used instead of the normal combining marks.

When U+0824 SAMARITAN MODIFIER LETTER SHORT A follows a letter used numerically, it indicates thousands, similar to the use of U+05F3 HEBREW PUNCTUATION GERESH for the same purpose in Hebrew.

Consonant Modifiers. The two marks, U+0816 SAMARITAN MARK IN and U+0817 SAMARITAN MARK IN-ALEF, are used to indicate a pharyngeal voiced fricative /ʕ/. These occur immediately following their base consonant and preceding any vowel signs, and are rendered above and to the right of the base consonant.

U+0818 SAMARITAN MARK OCCLUSION "strengthens" the consonant, for example changing /w/ to /b/. U+0819 SAMARITAN MARK DAGESH indicates consonant gemination. The *occlusion* and *dagesh* marks may both be applied to the same consonant, in which case the *occlusion* mark should precede the *dagesh* in logical order, and the *dagesh* is rendered above the *occlusion* mark. The *occlusion* mark is also used to designate personal names to distinguish them from homographs.

Epenthetic yut represents a kind of glide-vowel which interacts with another vowel. It was originally used only with the consonants *alaf*, *iy*, *it*, and *in*, in combination with a vowel sign. The combining U+081B SAMARITAN MARK EPENTHETIC YUT should be used for this purpose. When *epenthetic yut* is not fixed to one of the four consonants listed above, a new behavior evolved in which the mark for the *epenthetic yut* behaves as a spacing character, capable of bearing its own diacritical mark. U+081A SAMARITAN MODIFIER LETTER EPENTHETIC YUT should be used instead to represent the *epenthetic yut* in this context.

Punctuation. Samaritan uses a large number of punctuation characters. U+0830 SAMARITAN PUNCTUATION NEQUDAA and U+0831 SAMARITAN PUNCTUATION AFSAAQ ("interruption") are similar to the Hebrew *sof pasuq* and were originally used to separate sentences, and later to mark lesser breaks within a sentence. They have also been described respectively as "semicolon" and "pause." Samaritan also uses a smaller dot as a word separator, which can be represented by U+2E31 WORD SEPARATOR MIDDLE DOT. U+083D SAMARITAN PUNCTUATION SOF MASHFAAT is equivalent to the full stop. U+0832 SAMARITAN PUNCTUATION ANGED ("restraint") indicates a break somewhat less strong than an *afsaaq*. U+083E SAMARITAN PUNCTUATION ANNAAU ("rest") is stronger than the *afsaaq* and indicates that a longer time has passed between actions narrated in the sentences it separates.

U+0839 SAMARITAN PUNCTUATION QITSA is similar to the *annaau* but is used more frequently. The *qitsa* marks the end of a section, and may be followed by a blank line to further make the point. It has many glyph variants. One important variant, U+0837 SAMARITAN PUNCTUATION MELODIC QITSA, differs significantly from any of the others, and indicates the end of a sentence "which one should read melodically."

Many of the punctuation characters are used in combination with each other, for example: *afsaaq* + *nequdaa* or *nequdaa* + *afsaaq*, *qitsa* + *nequdaa*, and so on.

U+0836 SAMARITAN ABBREVIATION MARK follows an abbreviation. U+082D SAMARITAN MARK NEQUDAA is an editorial mark which indicates that there is a variant reading of the word.

Other Samaritan punctuation characters mark some prosodic or performative attributes of the text preceding them, as summarized in *Table 8-18*.

Table 8-18. Samaritan Performative Punctuation Marks

Code Point	Name	Description
0833	*bau*	request, prayer, humble petition
0834	*atmaau*	expression of surprise
0835	*shiyyaalaa*	question
0838	*ziqaa*	shout, cry
083A	*zaef*	outburst indicating vehemence or anger
083B	*turu*	didactic expression, a "teaching"
083C	*arkaanu*	expression of submissiveness

8.5 Thaana

Thaana: U+0780–U+07BF

The Thaana script is used to write the modern Dhivehi language of the Republic of Maldives, a group of atolls in the Indian Ocean. Like the Arabic script, Thaana is written from right to left and uses vowel signs, but it is not cursive. The basic Thaana letters have been extended by a small set of dotted letters used to transcribe Arabic. The use of modified Thaana letters to write Arabic began in the middle of the twentieth century. Loan words

from Arabic may be written in the Arabic script, although this custom is not very prevalent today. (See *Section 8.2, Arabic.*)

While Thaana's glyphs were borrowed in part from Arabic (letters *haa* through *vaavu* were based on the Arabic-Indic digits, for example), and while vowels and *sukun* are marked with combining characters as in Arabic, Thaana is properly considered an alphabet, rather than an abjad, because writing the vowels is obligatory.

Directionality. The Thaana script is written from right to left. Conformant implementations of Thaana script must use the Unicode Bidirectional Algorithm (see Unicode Standard Annex #9, "Unicode Bidirectional Algorithm").

Vowels. Consonants are always written with either a vowel sign (U+07A6..U+07AF) or the null vowel sign (U+07B0 THAANA SUKUN). U+0787 THAANA LETTER ALIFU with the null vowel sign denotes a glottal stop. The placement of the Thaana vowel signs is shown in *Table 8-19*.

Table 8-19. Thaana Glyph Placement

Syllable	Display
tha	◌ަ
thaa	◌ާ
thi	◌ި
thee	◌ީ
thu	◌ު
thoo	◌ެ
the	◌ެ
they	◌ެ
tho	◌ެ
thoa	◌ެ
th	◌ެ

Numerals. Both European (U+0030..U+0039) and Arabic digits (U+0660..U+0669) are used. European numbers are used more commonly and have left-to-right display directionality in Thaana. Arabic numeric punctuation is used with digits, whether Arabic or European.

Punctuation. The Thaana script uses spaces between words. It makes use of a mixture of Arabic and European punctuation, though rules of usage are not clearly defined. Sentence-final punctuation is now generally shown with a single period (U+002E "." FULL STOP) but may also use a sequence of two periods (U+002E followed by U+002E). Phrases may be separated with a comma (usually U+060C ARABIC COMMA) or with a single period (U+002E). Colons, dashes, and double quotation marks are also used in the Thaana script. In addition, Thaana makes use of U+061F ARABIC QUESTION MARK and U+061B ARABIC SEMICOLON.

Character Names and Arrangement. The character names are based on the names used in the Republic of Maldives. The character name at U+0794, *yaa,* is found in some sources as *yaviyani*, but the former name is more common today. Characters are listed in Thaana alphabetical order from *haa* to *ttaa* for the Thaana letters, followed by the extended characters in Arabic alphabetical order from *hhaa* to *waavu*.

