Chapter 9

**Middle East-I**

*Modern and Liturgical Scripts*

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

<table>
<thead>
<tr>
<th>Hebrew</th>
<th>Syriac</th>
<th>Mandaic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>Samaritan</td>
<td></td>
</tr>
</tbody>
</table>

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 Mandaic script was used in southern Mesopotamia in classical times for liturgical texts by adherents of the Mandaean gnostic religion. The Classical Mandaic and Neo-Mandaic languages are still in limited current use in modern Iran and Iraq and in the Mandaean diaspora.

The Middle Eastern scripts are mostly abjads, with small character sets. Words are demarcated by spaces. 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, Syriac and Mandaic are cursive scripts even when typeset, unlike Hebrew and Samaritan, where letters are unconnected. Most letters in Arabic, Syriac and Mandaic assume different forms depending on their position in a word. Shaping rules for the rendering of text are specified in Section 9.2, *Arabic*, Section 9.3, *Syriac* and Section 9.5, *Mandaic*. 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 pronunciation (“vocalization”) marks are encoded as combining characters, so support for
vocalized text necessitates use of composed character sequences. Yiddish and Syriac are normally written with vocalization; Hebrew, Samaritan, and Arabic are usually written unvocalized.
9.1 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 Pronunciation Marks.** 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 pronunciation marks.

*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 pronunciation marks: 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 23.2, 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 letter form used for one reading is referred to as qamats or qamats gadol; the new letter form 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 sup-
port Hebrew accents and vowel points may not necessarily support the special-usage characters U+05A2 HEBREW ACCENT ATNAH HAHKH and U+05C7 HEBREW POINT QAMATS QATAN.

Holam Male and Holam Hacer. 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 HASER FOR VAV 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.
9.2 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 Uyghur, 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. Other languages, such as Kurdish, Azerbaijani, Kazakh, and Uzbek have competing Arabic and Latin or Cyrillic orthographies in different countries.

The Arabic script is cursive, even in its printed form (see Figure 9-1). 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 *tashkil*, which are applied to consonantal base letters. In normal writing, however, these marks are omitted.

**Figure 9-1. Directionality and Cursive Connection**

| Memory representation: | ٢٥٥٥ |
| After reordering:      | ٥٥٥٥ |
| After joining:         | ٥٥٥٥ |

**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**: Diacritical 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 diacritical 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), *shaddah* (consonant gemination mark), and *sukun* (to mark lack of a following vowel). 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. In most modern languages using the Arabic script, it occurs only above *alef*, and in that combination represents the sound /ʔaa/. In Koranic Arabic, *maddah* occurs above *waw* or *yeh* to note vowel elongation. For this reason, and the shared use of *maddah* between Arabic and Syriac scripts, the precomposed combination U+0622 ARABIC LETTER ALEF WITH MADDA ABOVE is encoded, however the combining mark U+0653 ARABIC MADDA ABOVE is also encoded. U+0622 is given a canonical decomposition to the sequence of *alef* followed by the combining *maddah*. Some historical non-Arabic orthographies have also used *maddah* as an *ijam*. U+0653 should be used to represent those texts.

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 (or the /je/ sound used in Persian and Urdu for *ezaf*), 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. Sindhi uses U+2E41 reversed comma and U+204F reversed semicolon. Persian and some other languages use rounded forms of U+00AB left-pointing double angle quotation mark and U+00BB right-pointing double angle quotation mark.

For paired punctuation such as parentheses, the glyphs chosen to display for example, U+0028 left parenthesis and U+0029 right parenthesis, will depend on the direction of the rendered text. See “Paired Punctuation” in Section 6.2, General Punctuation, for more discussion.

**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 9-2.

**Figure 9-2. Using a Joiner**

Memory representation: ٨٩٦

After reordering: ٨٩٦

After joining: ٩٦

These connecting forms commonly occur in some abbreviations such as the marker for *hijri* dates, which consists of an initial form of *heh*: ٨.

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 9-3.

**Figure 9-3. Using a Non-joiner**

Memory representation: ٨٩٦

After reordering: ٨٩٦

After joining: ٩٦
Examples requiring the use of a non-joiner include the Persian plural suffix, some Persian proper names, and Ottoman Turkish vowels. This use of non-joiners is important for representation of text in such languages, and ignoring or removing them will result in text with a different meaning, or in meaningless text.

Joiners and non-joiners may also occur in combinations. The effects of such combinations are shown in Figure 9-4. For further discussion of joiners and non-joiners, see Section 23.2, Layout Controls.

Figure 9-4. Combinations of Joiners and Non-joiners

Memory representation: ٥۱۰۰۰۰۰۰۰۰ ۰
After reordering: ۰۰۰۰۰۰۰۰ ۵۰۰۰۰۰۰۰
After joining: ۰۵۰۰۰۰۰۰۰۰

Tashkil Nonspacing Marks. Tashkil are marks that indicate vowels or other modifications of consonant letters. In English, these marks are often referred to as “points.” They may also be called harakat, although technically, harakat refers to the subset of tashkil which denote short vowels. The code charts depict these tashkil in relation to a dotted 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 tashkil applied to the same Arabic base character. 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 9-5. 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 9-5.

Figure 9-5. Placement of Harakat
The shapes of the various *tashkil* marks may also depend on the style of writing. For example, *dammatan* can be written in at least three different styles:

- using a shape similar to that shown in the charts
- using two *dammus*, one of which is turned
- using two *dammus* vertically stacked

U+064C *Arabic dammatan* can be rendered in any of those three shapes. U+08F1 *Arabic open dammatan* is an alternative *dammatan* character for use in Quran orthographies which have two distinct forms of *dammatan* that convey a semantic difference.

**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 ٠١٢…). 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 Afghanistan, India, Iran, and Pakistan (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 9-1. A different set of number forms, called Rumi, was used in historical materials from Egypt to Spain, and is discussed in the subsection on “Rumi Numeral Symbols” in Section 22.3, *Numerals*.

**Table 9-1. Arabic Digit Names**

<table>
<thead>
<tr>
<th>Name</th>
<th>Code Points</th>
<th>Forms</th>
</tr>
</thead>
<tbody>
<tr>
<td>European</td>
<td>U+0030..U+0039</td>
<td>0123456789</td>
</tr>
<tr>
<td>Arabic-Indic</td>
<td>U+0660..U+0669</td>
<td></td>
</tr>
<tr>
<td>Eastern Arabic-Indic</td>
<td>U+06F0..U+06F9</td>
<td></td>
</tr>
<tr>
<td>Indic (Devanagari)</td>
<td>U+0966..U+096F</td>
<td>०१२३४५६७८९</td>
</tr>
</tbody>
</table>

There is substantial variation in usage of glyphs for the Eastern Arabic-Indic digits, especially for the digits four, five, six, and seven. Table 9-2 illustrates this variation with some
example glyphs for digits in languages of Afghanistan, India, Iran, and Pakistan. While some usage of the Persian glyph for U+06F7 extended Arabic-Indic digit seven can be documented for Sindhi, the form shown in Table 9-2 is predominant.

Table 9-2. Glyph Variation in Eastern Arabic-Indic Digits

<table>
<thead>
<tr>
<th>Code Point</th>
<th>Digit</th>
<th>Persian</th>
<th>Sindhi</th>
<th>Urdu</th>
</tr>
</thead>
<tbody>
<tr>
<td>U+06F4</td>
<td>4</td>
<td>ꝕ</td>
<td>ꝕ</td>
<td>ꝕ</td>
</tr>
<tr>
<td>U+06F5</td>
<td>5</td>
<td>Ꝗ</td>
<td>Ꝗ</td>
<td>Ꝗ</td>
</tr>
<tr>
<td>U+06F6</td>
<td>6</td>
<td>ꝗ</td>
<td>ꝗ</td>
<td>ꝗ</td>
</tr>
<tr>
<td>U+06F7</td>
<td>7</td>
<td>Ꝙ</td>
<td>Ꝙ</td>
<td>Ꝙ</td>
</tr>
</tbody>
</table>

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. 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+06F5 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 \u0673 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. The normal practice is that such marks be used at the end of words. In manuscripts, depending on the letter shapes present in the name and the calligraphic style in use, the honorific mark may appear over a letter in the middle of the word. If an exact representation of a manuscript is desired, the honorific mark may be represented as following that letter. The normalization algorithm does not move such word-level combining characters to the end of the word.

Spacing honorifics are also in wide use both in the Arabic script and among Muslim communities writing in other scripts. See "Word Ligatures" under Arabic Presentation Forms-A later in this section for more information.

**Arabic Mathematical Symbols.** A few Arabic mathematical symbols are encoded in this block. The Arabic mathematical radix signs, \u0606 Arabic-indic cube root and \u0607 Arabic-indic fourth root, differ from simple mirrored versions of \u221b cube root and \u221c fourth root, in that the digit portions of the symbols are written with Arabic-Indic digits and are not mirrored. \u0608 Arabic ray is a letterlike symbol used in Arabic mathematics.

**Date Separator.** \u060D 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 \u002f solidus, which is used, for example, as a separator in currency amounts.

**Full Stop.** \u061E 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.** \u060B 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 compatibility character, encoded for compatibility with Iranian standards. Ordinarily in Persian “rial” is simply spelled out as the sequence of letters, \u0631, 06CC, 0627, 0644>.

**Signs Spanning Numbers.** Several other special signs are written in association with numbers in the Arabic script. 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 their associated sequence of digits. In the text representation they precede the sequence of digits that they span, rather than follow a base character, as would be the case for a combining mark. Their General_Category value is Cf (format character). Unlike most other format 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. The characters have the Bidi_Class value of Arabic_Number to make them appear in the same run as the numbers following them.

A few similar signs spanning numbers or letters are associated with scripts other than Arabic. See the discussion of U+070F syriac abbreviation mark in Section 9.3, Syriac, and the discussion of U+110BD kaithi number sign in Section 15.2, Kaithi. All of these prefixed format controls, including the non-Arabic ones, are given the property value Pre-pended_Concatenation_Mark=True, to identify them as a class. They also have special behavior in text segmentation. (See Unicode Standard Annex #29, “Unicode Text Segmentation.”)

U+0600 arabic number sign signals the beginning of a number. It is followed by a sequence of one or more Arabic digits and is rendered below the digits of the number. The length of its rendered display may vary with the number of digits. The sequence terminates with the occurrence of any non-digit character.

U+0601 arabic sign sanah indicates a year (that is, as part of a date). This sign is also 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 9-6.

Figure 9-6. Arabic Year Sign

U+0602 arabic footnote marker is a specialized variant of number sign. Its use indicates that the number so marked represents a footnote number in the text.

U+0603 arabic sign safha is another specialized variant of number sign. It marks a page number.

U+0604 arabic sign samvat is a specialized variant of date sign used specifically to write dates of the Saka era. The shape of the glyph is a stylized abbreviation of the word samvat, the name of this calendar. It is seen in the Urdu orthography, where it contrasts with conventions used to display dates from the Gregorian or Islamic calendars.

U+0605 arabic number mark above is a specialized variant of number sign. It is used in Arabic text with Coptic numbers, such as in early astronomical tables. Unlike the other Arabic number signs, it extends across the top of the sequence of digits, and is used with Coptic digits, rather than with Arabic digits. (See also the discussion of supralineation and the numerical use of letters in Section 7.3, Coptic.)

U+06DD arabic end of ayah is another sign used to span numbers, but its rendering is somewhat different. Rather than extending below the following digits, this sign encloses the digit sequence. This sign is used conventionally to indicate numbered verses in the Koran.
U+08E2 ARABIC DISPUTED END OF AYAH is a specialized variant of the end of ayah. It is seen occasionally in Koranic text to mark a verse for which there is scholarly disagreement about the location of the end of the verse.

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, in almost every font style, 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 9-3. (See ArabicShaping.txt in the Unicode Character Database for a complete list.)

Table 9-3. Primary Arabic Joining Types

<table>
<thead>
<tr>
<th>Joining_Type</th>
<th>Examples and Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right_Joining (R)</td>
<td>ALEF, DAL, THAL, REH, ZAIN...</td>
</tr>
<tr>
<td>Left_Joining (L)</td>
<td>None (in Arabic)</td>
</tr>
<tr>
<td>Dual_Joining (D)</td>
<td>BEH, TEH, THEH, JEEM...</td>
</tr>
<tr>
<td>Join_Causing (C)</td>
<td>U+200D ZERO WIDTH JOINER and TATWEEL (U+0640). These characters are distinguished from the dual-joining characters in that they do not change shape themselves.</td>
</tr>
</tbody>
</table>
In Table 9-3, right and left refer to visual order, so a Joining_Type value of Right_Joining indicates that a character cursorily joins to a character displayed to its right in visual order. (For a discussion of the meaning of Joining_Type values in the context of a vertically rendered script, see “Cursive Joining” in Section 14.4, Phags-pa.) The Arabic characters with Joining_Type = Right_Joining are exemplified in more detail in Table 9-9, and those with Joining_Type = Dual_Joining are shown in Table 9-8. When characters do not join or cause joining (such as damma), they are classified as transparent.

The Phags-pa and Manichaean scripts have a few Left_Joining characters, which are otherwise unattested in the Arabic and Syriac scripts. See Section 10.5, Manichaean. For a discussion of the meaning of Joining_Type values in the context of a vertically rendered script, see “Cursive Joining” in Section 14.4, Phags-pa.

Table 9-4 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 9-3. Primary Arabic Joining Types (Continued)

<table>
<thead>
<tr>
<th>Joining_Type</th>
<th>Examples and Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non_Joining (U)</td>
<td>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 (U+0621), high hamza (U+0674), spaces, digits, punctuation, non-Arabic letters, and so on. Also, U+0600 arabic number sign, U+0605 arabic number mark above and U+06DD arabic end of ayah.</td>
</tr>
<tr>
<td>Transparent (T)</td>
<td>All nonspacing marks (General Category Mn or Me) and most format control characters (General Category Cf) are transparent to cursive joining. These include fathatan (U+064B) and other Arabic taskil, hamza below (U+0655), superscript alef (U+0670), combining Koranic annotation signs, and nonspacing marks from other scripts. Also U+070F syriac abbreviation mark.</td>
</tr>
</tbody>
</table>

### Table 9-4. Derived Arabic Joining Types

<table>
<thead>
<tr>
<th>Description</th>
<th>Derivation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right join-causing</td>
<td>Superset of dual-joining, left-joining, and join-causing</td>
</tr>
<tr>
<td>Left join-causing</td>
<td>Superset of dual-joining, right-joining, and join-causing</td>
</tr>
</tbody>
</table>

**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 9-5.

### Table 9-5. Arabic Glyph Types

<table>
<thead>
<tr>
<th>Glyph Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_n$</td>
<td>Non-joining glyph form that does not join on either side.</td>
</tr>
<tr>
<td>$X_r$</td>
<td>Right-joining glyph form (both right-joining and dual-joining characters may employ this form)</td>
</tr>
<tr>
<td>$X_l$</td>
<td>Left-joining glyph form (both left-joining and dual-joining characters may employ this form)</td>
</tr>
<tr>
<td>$X_m$</td>
<td>Dual-joining (medial) glyph form that joins on both left and right (only dual-joining characters employ this form)</td>
</tr>
</tbody>
</table>

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

$$MEEM_n + SHADDA_n + LAM_n \rightarrow MEEM_r + SHADDA_n + LAM_l$$

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

$$ALEF_n + TATWEEL_n \rightarrow ALEF_r + TATWEEL_n$$

**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:

$$TATWEEL_n + MEEM_n + TATWEEL_n \rightarrow TATWEEL_n + MEEM_m + TATWEEL_n$$

**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:

$$MEEM_n + TATWEEL_n \rightarrow MEEM_r + TATWEEL_n$$
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:

\[
\text{TATWEEL}_n + \text{MEEM}_n \rightarrow \text{TATWEEL}_n + \text{MEEM}_l
\]

\[
\phantom{+} + \rightarrow \phantom{+} + \rightarrow \phantom{+}
\]

If none of the preceding rules applies to a character \( X \), then it will adopt the non-joining 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.** The lam-alef type of ligatures are extremely common in the Arabic script. These ligatures occur in almost all font designs, except for a few modern styles. When supported by the style of the font, lam-alef ligatures are considered obligatory. This means that all character sequences rendered in that font, which match the rules specified in the following discussion, must form these ligatures. Many other Arabic ligatures are discretionary. Their use depends on the font design.

For the purpose of describing the obligatory Arabic ligatures, certain characters fall into two joining groups, as shown in Table 9-6. The complete list is available in ArabicShaping.txt in the Unicode Character Database.

<table>
<thead>
<tr>
<th>Joining Group</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALEF</td>
<td>MADDA-ON-ALEF, HAMZA ON ALEF, ...</td>
</tr>
<tr>
<td>LAM</td>
<td>LAM, LAM WITH SMALL V, LAM WITH DOT ABOVE, ...</td>
</tr>
</tbody>
</table>

**Ligature Rules.** The following rules describe the formation of obligatory 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 9-7.

**Table 9-7. Arabic Ligature Notation**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(X-Y)_n</td>
<td>Nominal ligature glyph form representing a combination of an X_r form and a Y_l form</td>
</tr>
<tr>
<td>(X-Y)_r</td>
<td>Right-joining ligature glyph form representing a combination of an X_r form and a Y_m form</td>
</tr>
</tbody>
</table>

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

\[
\text{ALEF}_r + \text{FATHA}_n + \text{LAM}_l \rightarrow (\text{LAM}-\text{ALEF})_n + \text{FATHA}_n
\]

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

\[
\begin{align*}
  \text{l} + \text{\textbar} & \rightarrow \text{\textbar\textbar} & (\text{not } \text{\textbar})
\end{align*}
\]

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

\[
\begin{align*}
  \text{l} + \text{\textbar} & \rightarrow \text{\textbar\textbar} & (\text{not } \text{\textbar})
\end{align*}
\]

**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 U+FB50..U+FDFF and U+FE70..U+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 9-8 exemplifies dual-joining Arabic characters and illustrates the forms taken by the letter skeletons and their *ijam* 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, or their dots may move to a different position. These joining groups include nya, farsi yeh, and burushaski yeh barree.

### Table 9-8. Dual-Joining Arabic Characters

<table>
<thead>
<tr>
<th>Joining Group</th>
<th>$X_n$</th>
<th>$X_r$</th>
<th>$X_m$</th>
<th>$X_l$</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEH</td>
<td>ب</td>
<td>ب</td>
<td></td>
<td>ب</td>
<td>Includes TEH and THEH.</td>
</tr>
<tr>
<td>NOON</td>
<td>ن ن</td>
<td>ن ن</td>
<td></td>
<td></td>
<td>Includes NOON GHUNNA.</td>
</tr>
<tr>
<td>AFRICAN NOON</td>
<td>ن ن</td>
<td>ن ن</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NYA</td>
<td>ن ن</td>
<td>ن ن</td>
<td></td>
<td></td>
<td>Jawi NYA.</td>
</tr>
<tr>
<td>YEH</td>
<td>ي ي</td>
<td>ي ي</td>
<td></td>
<td></td>
<td>Includes ALEF MAKSURA.</td>
</tr>
<tr>
<td>Farsi Yeh</td>
<td>ي ي</td>
<td>ي ي</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Burushaski Yeh Barree</td>
<td>ي ي</td>
<td>ي ي</td>
<td></td>
<td></td>
<td>Dual joining, as opposed to Yeh Barree</td>
</tr>
<tr>
<td>HAH</td>
<td>ح ح</td>
<td>ح ح</td>
<td></td>
<td></td>
<td>Includes KHAH and JEEM.</td>
</tr>
<tr>
<td>SEEN</td>
<td>س س</td>
<td>س س</td>
<td></td>
<td></td>
<td>Includes SHEEN.</td>
</tr>
</tbody>
</table>
Table 9-8. Dual-Joining Arabic Characters (Continued)

<table>
<thead>
<tr>
<th>Joining Group</th>
<th>Xn</th>
<th>Xr</th>
<th>Xm</th>
<th>Xl</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAD</td>
<td>ص</td>
<td>ص</td>
<td>ص</td>
<td>ص</td>
<td>Includes DAD.</td>
</tr>
<tr>
<td>TAH</td>
<td>ط</td>
<td>ط</td>
<td>ط</td>
<td>ط</td>
<td>Includes ZAH.</td>
</tr>
<tr>
<td>AIN</td>
<td>ع</td>
<td>ع</td>
<td>ع</td>
<td>ع</td>
<td>Includes GHAIN.</td>
</tr>
<tr>
<td>FEH</td>
<td>ف</td>
<td>ف</td>
<td>ف</td>
<td>ف</td>
<td></td>
</tr>
<tr>
<td>AFRICAN FEH</td>
<td>ف</td>
<td>ف</td>
<td>ف</td>
<td>ف</td>
<td></td>
</tr>
<tr>
<td>QAF</td>
<td>ق</td>
<td>ق</td>
<td>ق</td>
<td>ق</td>
<td></td>
</tr>
<tr>
<td>AFRICAN QAF</td>
<td>ق</td>
<td>ق</td>
<td>ق</td>
<td>ق</td>
<td></td>
</tr>
<tr>
<td>MEEM</td>
<td>م</td>
<td>م</td>
<td>م</td>
<td>م</td>
<td></td>
</tr>
<tr>
<td>HEH</td>
<td>ه</td>
<td>ه</td>
<td>ه</td>
<td>ه</td>
<td></td>
</tr>
<tr>
<td>KNOTTED HEH</td>
<td>ه</td>
<td>ه</td>
<td>ه</td>
<td>ه</td>
<td></td>
</tr>
<tr>
<td>HEH GOAL</td>
<td>ه</td>
<td>ه</td>
<td>ه</td>
<td>ه</td>
<td>Includes HAMZA ON HEH GOAL.</td>
</tr>
<tr>
<td>KAF</td>
<td>ك</td>
<td>ك</td>
<td>ك</td>
<td>ك</td>
<td></td>
</tr>
<tr>
<td>SWASH KAF</td>
<td>ك</td>
<td>ك</td>
<td>ك</td>
<td>ك</td>
<td></td>
</tr>
<tr>
<td>GAF</td>
<td>ك</td>
<td>ك</td>
<td>ك</td>
<td>ك</td>
<td>Includes KEHEH.</td>
</tr>
<tr>
<td>LAM</td>
<td>ل</td>
<td>ل</td>
<td>ل</td>
<td>ل</td>
<td></td>
</tr>
</tbody>
</table>
Right-Joining. *Table 9-9* 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 9-9. Right-Joining Arabic Characters**

<table>
<thead>
<tr>
<th>Joining Group</th>
<th>X₀</th>
<th>Xᵣ</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALEF</td>
<td>ﺞ</td>
<td>ﺞ</td>
<td></td>
</tr>
<tr>
<td>WAW</td>
<td>ﺝ</td>
<td>ﺝ</td>
<td></td>
</tr>
<tr>
<td>STRAIGHT WAW</td>
<td>ﺝ</td>
<td>ﺝ</td>
<td>Tatar straight waw</td>
</tr>
<tr>
<td>DAL</td>
<td>ﻒ</td>
<td>ﻒ</td>
<td>Includes thal.</td>
</tr>
<tr>
<td>REH</td>
<td>ﺕ</td>
<td>ﺕ</td>
<td>Includes zain.</td>
</tr>
<tr>
<td>TEH MARBUTA</td>
<td>ﻳ</td>
<td>ﻳ</td>
<td>Includes hamza on heh.</td>
</tr>
<tr>
<td>TEH MARBUTA GOAL</td>
<td>ﻳ</td>
<td>ﻳ</td>
<td></td>
</tr>
<tr>
<td>YEH WITH TAIL</td>
<td>ﺩ</td>
<td>ﺩ</td>
<td></td>
</tr>
<tr>
<td>YEH BARREE</td>
<td>ﻳ</td>
<td>ﻳ</td>
<td></td>
</tr>
<tr>
<td>ROHINGYA YEH</td>
<td>ﻳ</td>
<td>ﻳ</td>
<td>Isolated form does not occur.</td>
</tr>
</tbody>
</table>

Some characters occur only at the end of words or morphemes when correctly spelled; these 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. Other characters, such as *alef maksura*, are considered trailing in modern Arabic, but are dual-joining in Koranic Arabic and languages like Uyghur. These are classified as dual-joining.

**Letter heh.** In the case of U+0647 *Arabic letter heh*, the glyph ﺙ 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 isolated forms of U+0647 *Arabic letter heh* and U+06C1 *Arabic letter heh goal* both look like U+06D5 *Arabic letter ae*.

U+06BE *Arabic letter heh doachashmee* is used to represent any *heh*-like letter that appears with stems at both sides in all contextual forms. The exact contextual shapes of the
letter depend on the language and the style of writing. The forms shown in Table 9-8 for
knotted heh are used in certain styles of writing in South Asia. Other South Asian styles
may use different medial and final forms. The style used in China and Central Asia for lan-
guages such as Uyghur uses medial and final forms for heh doachashmee that are visually
similar to the medial form of heh shown in Table 9-8.

**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 Uni-
code 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 9-10.

### Table 9-10. Forms of the Arabic Letter yeh

<table>
<thead>
<tr>
<th>Character</th>
<th>Joining Group</th>
<th>X_n</th>
<th>X_r</th>
<th>X_m</th>
<th>X_l</th>
</tr>
</thead>
<tbody>
<tr>
<td>U+0649 alef maksura yeh</td>
<td>YEH</td>
<td>☐☐☐</td>
<td>☐☐☐</td>
<td>☐☐☐</td>
<td>☐☐☐</td>
</tr>
<tr>
<td>U+064A yeh</td>
<td>YEH</td>
<td>☐☐☐</td>
<td>☐☐☐</td>
<td>☐☐☐</td>
<td>☐☐☐</td>
</tr>
<tr>
<td>U+06CC Farsi yeh</td>
<td>Farsi yeh</td>
<td>☐☐☐</td>
<td>☐☐☐</td>
<td>☐☐☐</td>
<td>☐☐☐</td>
</tr>
<tr>
<td>U+0777 yeh with digit four below</td>
<td>YEH</td>
<td>☐☐☐</td>
<td>☐☐☐</td>
<td>☐☐☐</td>
<td>☐☐☐</td>
</tr>
<tr>
<td>U+0620 kashmiri yeh</td>
<td>YEH</td>
<td>☐☐☐</td>
<td>☐☐☐</td>
<td>☐☐☐</td>
<td>☐☐☐</td>
</tr>
<tr>
<td>U+06D2 yeh barree</td>
<td>Yeh Barree</td>
<td>☐☐☐</td>
<td>☐☐☐</td>
<td>☐☐☐</td>
<td>☐☐☐</td>
</tr>
<tr>
<td>U+077A yeh barree with digit two above</td>
<td>Burushaski yeh Barree</td>
<td>☐☐☐</td>
<td>☐☐☐</td>
<td>☐☐☐</td>
<td>☐☐☐</td>
</tr>
<tr>
<td>U+08AC rohingya yeh</td>
<td>Rohingya yeh</td>
<td>☐☐☐</td>
<td>☐☐☐</td>
<td>☐☐☐</td>
<td>☐☐☐</td>
</tr>
</tbody>
</table>

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 9-10. 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 9-10. 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 THREE ABOVE and U+077B ARABIC LETTER YEH Barree with extended ARABIC-INDIC DIGIT TWO 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 9-10.

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 9-10. However, when Kashmiri is written in Nastaliq style, the final and isolated forms of kashmiri yeh usually appear as truncated yeh shapes (ज़) 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.

**Noon Ghunna.** The letter noon ghunna is used to mark nasalized vowels at the ends of words and some morphemes in Urdu, Balochi, and other languages of Southern Asia. It is represented by U+06BA ARABIC LETTER NOON GHUNNA. The noon ghunna has the shape of a dotless noon and typically appears only in final and isolated contexts in these languages. In the middle of words and morphemes, the normal noon, U+0646 ARABIC LETTER NOON,
is used instead. To avoid ambiguity, sometimes a special mark, \( \text{U+0658 ARABIC MARK NOON GHUNNA} \), is added to the dotted noon to indicate nasalization.

\( \text{U+06BA ARABIC LETTER NOON GHUNNA} \) is also used as a dotless noon for the noon skeleton in all four of its contextual forms. As such, it is used in representation of early Arabic and Koranic Arabic texts. Rendering systems should display \( \text{U+06BA} \) as a dual-joining letter, with all four contextual forms shown dotless, regardless of the language of the text.

Advanced text entry applications for Urdu, Balochi, and other languages using noon ghunna may include specialized logic for its handling. For example, they might detect mid-word usage of the noon ghunna key and emit the regular dotted noon character (\( \text{U+0646} \)) instead, as appropriate for spelling in that context.

**Combining Hamza Above.** \( \text{U+0654 ARABIC HAMZA ABOVE} \) is intended both for the representation of hamza semantics in combination with certain Arabic letters, and as a diacritical 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.

\( \text{U+0654 ARABIC HAMZA ABOVE} \) should not be used with \( \text{U+0649 ARABIC LETTER ALEF MAKSURA} \). Instead, the precomposed \( \text{U+0626 ARABIC LETTER YEH WITH HAMZA ABOVE} \) should be used to represent a yeh-shaped base with no dots in any positional form, and with a hamza above. Because \( \text{U+0626} \) is canonically equivalent to the sequence \(<\text{U+064A ARABIC LETTER YEH}, \text{U+0654 ARABIC HAMZA ABOVE}\>\), when \( \text{U+0654} \) is applied to \( \text{U+064A ARABIC LETTER YEH} \), 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, \( \text{U+08A8 ARABIC LETTER YEH WITH TWO DOTS BELOW AND HAMZA ABOVE} \), 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 Cameroon, to represent a palatal implosive.

In most other cases when a hamza is needed as a mark above for an Arabic letter, \( \text{U+0654 ARABIC HAMZA ABOVE} \) can be freely used in combination with basic Arabic letters. Three exceptions are the extended Arabic letters \( \text{U+0681 ARABIC LETTER HAH WITH HAMZA ABOVE}, \text{U+076C ARABIC LETTER REH WITH HAMZA ABOVE}, \text{and U+08A1 ARABIC LETTER BEH WITH HAMZA ABOVE} \), where the hamza mark is functioning as an ijam (diacritic), rather than as a normal hamza. In those three cases, the extended Arabic letters have no canonical decompositions; consequently, the preference is to use those precomposed forms, rather than applying \( \text{U+0654 ARABIC HAMZA ABOVE} \) to hah, reh, or beh respectively.

These interactions between various letters and the hamza are summarized in Table 9-11.

The first five entries in Table 9-11 show the cases where the hamza above can be freely used, and where there is a canonical equivalence to the precomposed characters. The last four entries show the exceptions, where use of the hamza above is inappropriate, and where only the precomposed characters should be used.
Table 9-11. Arabic Letters With Hamza Above

<table>
<thead>
<tr>
<th>Code Point</th>
<th>Name</th>
<th>Decomposition</th>
</tr>
</thead>
<tbody>
<tr>
<td>0623</td>
<td>alef with hamza above</td>
<td>0627 0654</td>
</tr>
<tr>
<td>0624</td>
<td>waw with hamza above</td>
<td>0648 0654</td>
</tr>
<tr>
<td>0626</td>
<td>yeh with hamza above</td>
<td>064A 0654</td>
</tr>
<tr>
<td>06C2</td>
<td>heh goal with hamza above</td>
<td>06C1 0654</td>
</tr>
<tr>
<td>06D3</td>
<td>yeh barree with hamza above</td>
<td>06D2 0654</td>
</tr>
<tr>
<td>0681</td>
<td>hah with hamza above</td>
<td>None</td>
</tr>
<tr>
<td>076C</td>
<td>reh with hamza above</td>
<td>None</td>
</tr>
<tr>
<td>08A1</td>
<td>beh with hamza above</td>
<td>None</td>
</tr>
<tr>
<td>08A8</td>
<td>yeh with 2 dots below and hamza above</td>
<td>None</td>
</tr>
</tbody>
</table>

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 Sorani. Some of those extensions are shared with Persian, Urdu, or other languages; for example, U+06C6 ARABIC LETTER OE, which represents the Kurdish vowel [ə]. Sorani also makes other unusual adaptations of the Arabic script, including the use of a digraph waw+waw to represent the long Kurdish vowel [uː]. 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 Sorani 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 Sorani 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 Sorani 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 20th 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 characters of Joining_Group=ALEF.

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. Characters used for other languages are annotated in the character names list.

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

This block contains a list of Arabic presentation forms encoded as characters primarily for compatibility reasons. These characters have a preferred representation that makes use of a normal (noncompatibility) Arabic character, or in many cases a sequence of Arabic characters. Presentation form is a mostly obsolete term for a contextually shaped glyph (for a single character) or for a ligature glyph (for a sequence of 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. The included 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 *ijam* or *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. The Arabic pedagogical symbols do not partake of any Arabic shaping behavior. Their Joining_Type is Non_Joining, so if used in juxtaposition with an Arabic letter skeleton, they will break the cursive connection and render after the letter, instead of above or below it.

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.

**Word Ligatures.** The signs and symbols encoded at U+FDF0..U+FDFD are word ligatures sometimes treated as a unit. Most of them are encoded for compatibility with older character sets and are rarely used, except the following:

U+FDF2 Arabic ligature allah isolated form is a very common ligature, used to display the name of God. When the formation of the *allah* ligature is desired, the recommended way to represent the word would be `<aelf, lam, lam, shadda, superscript aelf, heh>` <0627, 0644, 0644, 0651, 0670, 0647>. In non-Arabic languages, other forms of *heh*, such as *heh goal* (U+06C1), may also form the ligature. Extra care should be taken not to form the ligature in the absence of the *shadda* and the *superscript aelf*, as the sequences `<aelf, lam, lam, heh>` and `<aelf, lam, lam, shadda, heh>` exist in Persian and other languages with different meanings or pronunciations, where the formation of the ligature would be incorrect and inappropriate.

U+FDFA Arabic ligature sallallahou alayhe wasallam and U+FDFB Arabic ligature jallajalalouhou are honorifics, commonly used after the name of the prophet Muhammad or God. Their usage is comparable to the honorifics found at U+0610..U+0613, except that these are spacing characters. The same characters are sometimes used by Muslims writing in other scripts such as Latin and Cyrillic.

U+FDFD Arabic ligature bismillah ar-rahman ar-raheem is a special ligated form of the *basmala*, a common opening phrase used by Muslims. The ligature is written in a multitude of ways. Its usage is common in writings by Muslims in non-Arabic scripts, even more than the honorifics mentioned above. It can be displayed as a unit above text in several different scripts, such as Bengali and Thaana. Unlike the other Arabic word ligatures, this character does not have a compatibility decomposition.
U+FDFC rial sign is a condensed version of the word rial, the Iranian currency. The character was invented by a typewriter standardization committee in 1973 and is encoded in the Unicode Standard as a compatibility character, as it continues to be specified in Iranian national standards for character sets and keyboard layouts, including ISIRI 9147:2007. Except for a short life during the typewriter era, it has not received widespread usage outside standards, as Iranians prefer to spell out the word as <reh, farsi yeh, alef, lam>.

**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 some of the obligatory lam-alef ligatures. They are included here for compatibility with preexisting standards and legacy implementations that use these forms as characters. Instead of these, letters from the Arabic block (U+0600..U+06FF) should be used for interchange. Implementations should 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 23.8, Specials.
9.3 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 diacritical 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 9-7.
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 9-8.

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.

Diacritical Marks and Vowels. The function of the diacritical 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 sys-
tem, whereas West Syriac texts (especially later ones and in modern times) employ a mixture of the two systems.

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

1. U+0741 SYRIAC QUSHSHAHA and U+0742 SYRIAC RUHKAKHA 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 NONJOINER 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 9-12.
Middle East-I 398

Use of Characters of the Arabic Block. Syriac makes use of several characters from the Arabic block, including U+0640 \text{arabic tatweel}. Modern texts use U+060C \text{arabic comma}, U+061B \text{arabic semicolon}, and U+061F \text{arabic question mark}. The \text{shadda} (U+0651) is also used in the core part of literary Syriac on top of a \text{waw} in the word “O”. Arabic \text{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 characters. 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 \text{syriac Table 9-12. Miscellaneous Syriac Diacritic Use}

<table>
<thead>
<tr>
<th>Code Points</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>U+0303, U+0330</td>
<td>These are used in Swadaya to indicate letters not found in Syriac.</td>
</tr>
<tr>
<td>U+0304, U+0320</td>
<td>These are used for various purposes ranging from phonological to grammatical to orthographic markers.</td>
</tr>
<tr>
<td>U+0307, U+0323</td>
<td>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.</td>
</tr>
<tr>
<td>U+0308</td>
<td>This is the plural marker. It is also used in Garshuni for the Arabic teh marbuta.</td>
</tr>
<tr>
<td>U+030A, U+0325</td>
<td>These are two other forms for the indication of qushshaya and rukkakha. They are used interchangeably with U+0741 \text{syriac qushshaya} and U+0742 \text{syriac rukkakha}, especially in West Syriac grammar books.</td>
</tr>
<tr>
<td>U+0324</td>
<td>This diacritical mark is found in ancient manuscripts. It has a grammatical and phonological function.</td>
</tr>
<tr>
<td>U+032D</td>
<td>This is one of the digit markers.</td>
</tr>
<tr>
<td>U+032E</td>
<td>This is a mark used in late and modern East Syriac texts as well as in Swadaya to indicate a fricative pe.</td>
</tr>
</tbody>
</table>
LETTER ALAPH in final position. The additional glyph types associated with final alaph are listed in Table 9-13.

Table 9-13. Syriac Final Alaph Glyph Types

<table>
<thead>
<tr>
<th>Glyph Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₀ᶠ</td>
<td>Final joining (alaph only)</td>
</tr>
<tr>
<td>A₀ᶠ₉</td>
<td>Final non-joining except following dalath and rish (alaph only)</td>
</tr>
<tr>
<td>A₀₉</td>
<td>Final non-joining following dalath and rish (alaph only)</td>
</tr>
</tbody>
</table>

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 9.2, 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 A₀ᶠ₉.

\[
\begin{array}{c}
| + & \rightarrow & | + & \rightarrow & | \\
\end{array}
\]

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 A₀₉.

\[
\begin{array}{c}
| + & \rightarrow & | + & \rightarrow & | \\
\end{array}
\]

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 A₀ₑₙ.

\[
\begin{array}{c}
| + & \rightarrow & | + & \rightarrow & | \\
\end{array}
\]

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 9.2, Arabic.

Table 9-14 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 9-14. Dual-Joining Syriac Characters

<table>
<thead>
<tr>
<th>Joining Group</th>
<th>X_n</th>
<th>X_r</th>
<th>X_m</th>
<th>X_l</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>BETH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Includes PERSIAN BHETH</td>
</tr>
<tr>
<td>GAMAL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Includes GAMAL GARSHUNI and PERSIAN GHAMAL</td>
</tr>
<tr>
<td>HETH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TETH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Includes TETH GARSHUNI</td>
</tr>
<tr>
<td>YUDH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KAPH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KHAPH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sogdian</td>
</tr>
<tr>
<td>LAMADH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MIM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NUN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEMKATH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FINAL_SEMKATH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>REVERSED_PE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sogdian</td>
</tr>
<tr>
<td>QAPH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SHIN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 9-15 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 9-15. Right-Joining Syriac Characters

<table>
<thead>
<tr>
<th>Joining Group</th>
<th>X₀</th>
<th>Xₗ</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>DALATH_RISH</td>
<td>‹️</td>
<td>›️</td>
<td>Includes RISH, DOTLESS DALATH RISH, and PERSIAN DHALATH</td>
</tr>
<tr>
<td>HE</td>
<td>🅒</td>
<td>🅒</td>
<td></td>
</tr>
<tr>
<td>SYRIAC_WAW</td>
<td>🅐</td>
<td>🅐</td>
<td></td>
</tr>
<tr>
<td>ZAIN</td>
<td>🅓</td>
<td>🅓</td>
<td></td>
</tr>
<tr>
<td>ZHAIN</td>
<td>🅒</td>
<td>🅒</td>
<td>Sogdian</td>
</tr>
<tr>
<td>YUDH_HE</td>
<td>🅒</td>
<td>🅒</td>
<td></td>
</tr>
<tr>
<td>SADHE</td>
<td>🅒</td>
<td>🅒</td>
<td></td>
</tr>
<tr>
<td>TAW</td>
<td>🅒</td>
<td>🅒</td>
<td></td>
</tr>
</tbody>
</table>

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 9-16 illustrates all of the glyph forms for alaph in each of the three major Syriac type styles.

Table 9-16. Syriac Alaph Glyph Forms

<table>
<thead>
<tr>
<th>Type Style</th>
<th>X₀</th>
<th>Xₗ</th>
<th>Aᶠj</th>
<th>Aᶠm</th>
<th>Aᶠx</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estrangela</td>
<td>🅒</td>
<td>🅒</td>
<td>🅒</td>
<td>🅒</td>
<td>🅒</td>
</tr>
<tr>
<td>Serto (West Syriac)</td>
<td>₪</td>
<td>₪</td>
<td>₪</td>
<td>₪</td>
<td>₪</td>
</tr>
<tr>
<td>East Syriac</td>
<td>₪</td>
<td>₪</td>
<td>₪</td>
<td>₪</td>
<td>₪</td>
</tr>
</tbody>
</table>

Ligature Classes. As in other scripts, ligatures in Syriac vary depending on the font style. Table 9-17 identifies the principal valid ligatures for each font style. When applicable, these ligatures are obligatory, unless denoted with an asterisk (*).
Table 9-17. Syriac Ligatures

<table>
<thead>
<tr>
<th>Characters</th>
<th>Estrangela</th>
<th>Serto (West Syriac)</th>
<th>East Syriac</th>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALAPH LAMADH</td>
<td>N/A</td>
<td>Dual-joining</td>
<td>N/A</td>
<td>Beth Gazo</td>
</tr>
<tr>
<td>GAMAL LAMADH</td>
<td>N/A</td>
<td>Dual-joining*</td>
<td>N/A</td>
<td>Armalah</td>
</tr>
<tr>
<td>GAMAL E</td>
<td>N/A</td>
<td>Dual-joining*</td>
<td>N/A</td>
<td>Armalah</td>
</tr>
<tr>
<td>HE YUDH</td>
<td>N/A</td>
<td>N/A</td>
<td>Right-joining*</td>
<td>Qdom</td>
</tr>
<tr>
<td>YUDH TAW</td>
<td>N/A</td>
<td>Right-joining*</td>
<td>N/A</td>
<td>Armalah*</td>
</tr>
<tr>
<td>KAPH LAMADH</td>
<td>N/A</td>
<td>Dual-joining*</td>
<td>N/A</td>
<td>Shhimo</td>
</tr>
<tr>
<td>KAPH TAW</td>
<td>N/A</td>
<td>Right-joining*</td>
<td>N/A</td>
<td>Armalah</td>
</tr>
<tr>
<td>LAMADH SPACE ALAPH</td>
<td>N/A</td>
<td>Right-joining*</td>
<td>N/A</td>
<td>Nomocanon</td>
</tr>
<tr>
<td>LAMADH ALAPH</td>
<td>Right-joining*</td>
<td>Right-joining*</td>
<td>BFBS</td>
<td></td>
</tr>
<tr>
<td>LAMADH LAMADH</td>
<td>N/A</td>
<td>Dual-joining*</td>
<td>N/A</td>
<td>Shhimo</td>
</tr>
<tr>
<td>NUN ALAPH</td>
<td>N/A</td>
<td>Right-joining*</td>
<td>N/A</td>
<td>Shhimo</td>
</tr>
<tr>
<td>SEMAKATH TETH</td>
<td>N/A</td>
<td>Dual-joining*</td>
<td>N/A</td>
<td>Qurobo</td>
</tr>
<tr>
<td>SADHE NUN</td>
<td>Right-joining*</td>
<td>Right-joining*</td>
<td>Right-joining*</td>
<td>Mushhoto</td>
</tr>
<tr>
<td>BISH SEYAME</td>
<td>Right-joining</td>
<td>Right-joining</td>
<td>Right-joining</td>
<td>BFBS</td>
</tr>
<tr>
<td>TAW ALAPH</td>
<td>Right-joining*</td>
<td>N/A</td>
<td>Right-joining*</td>
<td>Qdom</td>
</tr>
<tr>
<td>TAW YUDH</td>
<td>N/A</td>
<td>N/A</td>
<td>Right-joining*</td>
<td></td>
</tr>
</tbody>
</table>

**Syriac Supplement: U+0860–U+086F**

The Syriac Supplement block contains characters used to write a dialect of Malayalam called Suriyani Malayalam, which is also known as Garshuni (Karshoni) or Syriac Malayalam.
9.4 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 10.3, Phoenician). In contrast, the more recent square Hebrew script associated with Judaism derives from the Imperial Aramaic script (see Section 10.4, 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-alaf, are used to indicate a pharyngeal voiced fricative /f/. 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 MASHEAAT 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 ("test") 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 annaaau 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 9-18.

Table 9-18. Samaritan Performative Punctuation Marks

<table>
<thead>
<tr>
<th>Code Point</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0833</td>
<td>bau</td>
<td>request, prayer, humble petition</td>
</tr>
<tr>
<td>0834</td>
<td>atmaau</td>
<td>expression of surprise</td>
</tr>
<tr>
<td>0835</td>
<td>shiyyaalaa</td>
<td>question</td>
</tr>
<tr>
<td>0838</td>
<td>ziqaa</td>
<td>shout, cry</td>
</tr>
<tr>
<td>083A</td>
<td>zaef</td>
<td>outburst indicating vehemence or anger</td>
</tr>
<tr>
<td>083B</td>
<td>turu</td>
<td>didactic expression, a “teaching”</td>
</tr>
<tr>
<td>083C</td>
<td>arkaanu</td>
<td>expression of submissiveness</td>
</tr>
</tbody>
</table>
9.5 Mandaic

**Mandaic: U+0840–U+085F**

The origins of the Mandaic script are unclear, but it is thought to have evolved between the second and seventh century CE from a cursivized form of the Aramaic script (as did the Syriac script) or from the Parthian chancery script. It was developed by adherents of the Mandaean gnostic religion of southern Mesopotamia to write the dialect of Eastern Aramaic they used for liturgical purposes, which is referred to as Classical Mandaic.

The religion has survived into modern times, with more than 50,000 Mandaeans in several communities worldwide (most having left what is now Iraq). In addition to the Classical Mandaic still used within some of these communities, a variety known as Neo-Mandaic or Modern Mandaic has developed and is spoken by a small number of people. Mandaeans consider their script sacred, with each letter having specific mystic properties, and the script has changed very little over time.

**Letter It.** The character U+0847 🅾️ mandaic letter it is a pharyngeal, pronounced [hu]. It can appear at the end of personal names or at the end of words to indicate the third person singular suffix.

**Structure.** Mandaic is unusual among Semitic scripts in being a true alphabet; the letters halqa, ushenna, aksa, and in are used to write both long and short forms of vowels, instead of functioning as consonants also used to write long vowels (*matres lectionis*), in the manner characteristic of other Semitic scripts. This is possible because some consonant sounds represented by the corresponding letters in other Semitic scripts are not used in the Mandaic language.

The character U+0856 🅶️ mandaic letter dushenna, also called adu, has a morphemic function. It is used to write the relative pronoun and the genitive exponent *di*. Dushenna is a digraph derived from an old ligature for *ad + aksa*. It is thus an addition to the usual Semitic set of 22 characters. The Mandaic alphabet is traditionally represented as the 23 letters halqa through dushenna, with halqa appended again at the end to form a symbolically-important cycle of 24 letters.

Two additional Mandaic characters are encoded in the Unicode Standard: U+0857 🅷️ mandaic letter kad is derived from an old ligature of *ak + dushenna*; it is a digraph used to write the word *kd*, which means “when, as, like”. The second additional character, U+0858 🅸️ mandaic letter ain, is a borrowing from U+0639 arabic letter ain.

Three diacritical marks are used in teaching materials to differentiate vowel quality; they may be omitted from ordinary text. U+0859 🅹️ mandaic affrication mark is used to extend the character set for foreign sounds (whether affrication, lenition, or another sound). U+085A 🅸️ mandaic vocalization mark is used to distinguish vowel quality of halqa, ushenna, and aksa. U+085B 🅺️ mandaic gemination mark is used to indicate what native writers call a “hard” pronunciation.
Punctuation. Sentence punctuation is used sparsely. A single script-specific punctuation mark is encoded: U+085E MANDAIC PUNCTUATION. It is used to start and end text sections, and is also used in colophons—the historical lay text added to the religious text—where it is typically displayed in a smaller size.

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

Shaping and Layout Behavior. Mandaic has fully-developed joining behavior, with forms as shown in Table 9-19 and Table 9-20. In these tables, $X_n$, $X_r$, $X_m$, and $X_l$ designate the nominal, right-joining, dual-joining (medial), and left-joining forms respectively, just as in Table 9-7, Table 9-8, and Table 9-9.

<table>
<thead>
<tr>
<th>Character</th>
<th>$X_n$</th>
<th>$X_r$</th>
<th>$X_m$</th>
<th>$X_l$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AD</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AH</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>USHENNA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AK</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AN</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IN</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASZ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AQ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Line Breaking. Spaces provide the primary line break opportunity. When text is fully justified, words may be stretched as in Arabic. U+0640 ARABIC TATWEEL may be inserted for this purpose.
Table 9-20. Right-Joining Mandaic Characters

<table>
<thead>
<tr>
<th>Character</th>
<th>$X_n$</th>
<th>$X_r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>HALQA</td>
<td>٥٠</td>
<td>٥٠</td>
</tr>
<tr>
<td>AZ</td>
<td></td>
<td>٥٠</td>
</tr>
<tr>
<td>IT</td>
<td></td>
<td>٥٠</td>
</tr>
<tr>
<td>AKSA</td>
<td>٥٠</td>
<td>٥٠</td>
</tr>
<tr>
<td>ASH</td>
<td>٥٠</td>
<td>٥٠</td>
</tr>
<tr>
<td>Middle East-I</td>
<td>408</td>
<td>9.5 Mandaic</td>
</tr>
</tbody>
</table>