This PDF file is an excerpt from *The Unicode Standard, Version 4.0*, issued by the Unicode Consortium and published by Addison-Wesley. The material has been modified slightly for this online edition, however the PDF files have not been modified to reflect the corrections found on the Updates and Errata page (http://www.unicode.org/errata/). For information on more recent versions of the standard, see http://www.unicode.org/standard/versions/enumeratedversions.html.

Dai Kan-Wa Jiten used as the source of reference Kanji codes was written by Tetsuji Morohashi and published by Taishukan Shoten.

Cover and CD-ROM label design: Steve Mehallo, http://www.mehallo.com

The publisher offers discounts on this book when ordered in quantity for bulk purchases and special sales. For more information, customers in the U.S. please contact U.S. Corporate and Government Sales, (800) 382-3419, corpsales@pearsontechgroup.com. For sales outside of the U.S., please contact International Sales, +1 317 581 3793, international@pearsontechgroup.com

Visit Addison-Wesley on the Web: http://www.awprofessional.com

Library of Congress Cataloging-in-Publication Data
The Unicode Standard, Version 4.0 : the Unicode Consortium /Joan Aliprand... [et al.].
p. cm.
Includes bibliographical references and index.
ISBN 0-321-18578-1 (alk. paper)
1. Unicode (Computer character set). I. Aliprand, Joan.

QA268.U545 2004
005.7'2—dc21

2003052158

ISBN 0-321-18578-1
Text printed on recycled paper
1 2 3 4 5 6 7 8 9 10—CRW—0706050403
First printing, August 2003

Chapter 8

Middle Eastern Scripts

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

- Hebrew
- Arabic
- Syriac
- Thaana

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 scripts also function as major liturgical scripts and, therefore, are used worldwide by various religious groups. The Thaana script is used to write Dhivehi, the language of the Republic of Maldives, an island nation in the middle of the Indian Ocean.

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

Text in these scripts is written from right to left. Implementations of these scripts must conform to the Unicode bidirectional algorithm (see Unicode Standard Annex #9, "The Bidirectional Algorithm"). For more information about writing direction, see *Section 2.9, Writing Direction*.

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

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

8.1 Hebrew

Hebrew: U+0590–U+05FF

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

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

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

Vowels and Other Marks of Pronunciation. These combining marks, generically called *points* in the context of Hebrew, indicate vowels or other modifications of consonantal letters. General rules for applying combining marks are given in *Section 2.10, Combining Characters*, and *Section 3.11, Canonical Ordering Behavior*. Additional Hebrew-specific behavior is described below.

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

Dagesh, U+05BC HEBREW POINT DAGESH, has the form of a dot that appears inside the letter that it affects. It is not a vowel, but 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+05B9, U+05BB.

The remaining three points are *marks of pronunciation*: U+05BD HEBREW POINT METEG, U+05BF HEBREW POINT RAFE, and U+FB1E HEBREW POINT JUDEO-SPANISH VARIKA. *Meteg*, also known as *siluq*, goes below the base character; *rafe* and *varika* go above. 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 upon the direction of the rendered text. See *Section 4.7, Bidi Mirrored—Normative*, 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 there are complex typographic rules for positioning 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, on the other hand, 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, which 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.

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 also included. The Hebrew plus sign variant, U+FB29 HEBREW LETTER ALTERNATIVE PLUS SIGN, is used more often in handwriting than in print, but does occur in school textbooks. It is used by those who wish to avoid cross symbols, which can have religious and historical connotations.

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

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

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

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

8.2 Arabic

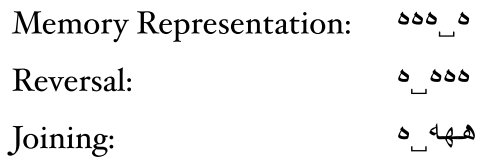
Arabic: U+0600–U+06FF

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

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

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

Figure 8-1. Directionality and Cursive Connection



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.

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+061F ARABIC QUESTION MARK, and U+066A ARABIC PERCENT SIGN. For paired punctuation such as parentheses, the glyphs chosen to represent U+0028 LEFT PARENTHESIS and U+0029 RIGHT PARENTHESIS will depend upon the direction of the rendered text.

The Non-joiner and the Joiner. The Unicode Standard provides two user-selectable formatting codes: U+200C ZERO WIDTH NON-JOINER and U+200D ZERO WIDTH JOINER (see *Figure 8-2*, *Figure 8-3*, and *Figure 8-4*). The use of a non-joiner between two letters prevents those letters from forming a cursive connection with each other when rendered. Examples

include the Persian plural suffix, some Persian proper names, and Ottoman Turkish vowels. For further discussion of joiners and non-joiners, see *Section 15.2, Layout Controls*.

Figure 8-2. Using a Joiner

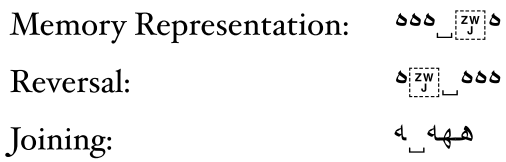


Figure 8-3. Using a Non-joiner

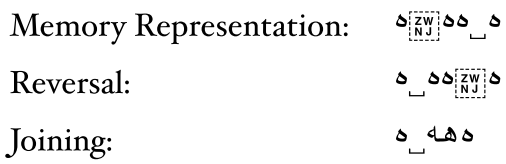
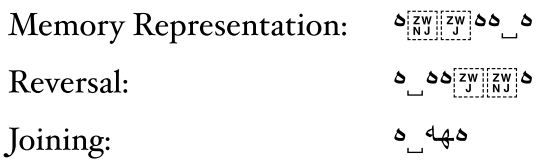


Figure 8-4. Combinations of Joiners and Non-joiners



Harakat (Vowel) Nonspacing Marks. *Harakat* are marks that indicate vowels or other modifications of consonant letters. The occurrence of a character in the harakat range and its depiction in relation to a dashed circle constitute an assertion that this character is intended to be applied via some process *to the character that precedes it* in the text stream, the base character. General rules for applying nonspacing marks are given in *Section 7.7, Combining Marks*. The few marks that are placed after (to the left of) the base character are treated as ordinary spacing characters in the Unicode Standard. The Unicode Standard does not specify a sequence order in case of multiple harakat applied to the same Arabic base character, as there is no possible ambiguity of interpretation. For more information about the canonical ordering of nonspacing marks, see *Section 2.10, Combining Characters,* and *Section 3.11, Canonical Ordering Behavior*.

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

The Unicode Standard includes both *Indic* digits (including forms used with different Indic scripts), *Arabic* digits (with forms used in most of the Arabic world), and *European* digits (now used internationally). Because of this decision, the traditional names could not be retained without confusion. In addition, there are two main variants of the Arabic digits—those used in Iran, Pakistan, and Afghanistan (here called *Eastern Arabic-Indic*) and those used in other parts of the Arabic world.

In summary, the Unicode Standard uses the names shown in *Table 8-1*. These names have been chosen to reduce the confusion involved in the use of the decimal number forms. They are not intended to show any preferences in name usage; as with the choice of any other names, they are meant to be unique distinguishing labels and should not be viewed as favoring one culture over another.

Table 8-1. Digit Names

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

There is substantial variation among the languages in the glyphs for the Eastern Arabic-Indic digits, especially for the digits four, five, six, and seven. *Table 8-2* illustrates this variation with some example glyphs for digits in languages of Iran, Pakistan, and India. While some usage of the Persian glyph for U+06F7 EXTENDED ARABIC-INDIC DIGIT SEVEN can be documented for Sindhi, the form shown in *Table 8-2* is predominant.

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

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

The Unicode Standard provides a single, complete sequence of digits for Persian, Sindhi, and Urdu to account for the differences in appearance and directional treatment when rendering them. (For a complete discussion of directional formatting of numbers in the Unicode Standard, see Unicode Standard Annex #9, "The 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 relatively obscure languages, such as Baluchi and Lahnda, which have little tradition in printed typography. As a result, the set of characters encoded in this section unavoidably contains a few spurious forms. The Unicode Standard encodes multiple forms of the Extended Arabic letters because the character forms and usages are not well documented for a number of languages. This approach was felt to be the most practical in the interest of minimizing the risk of omitting valid characters.

Koranic Annotation Signs. These characters are used in the Koran to mark pronunciation and other annotation. The enclosing mark U+06DE is used to enclose a digit. When rendered, the digit appears in a smaller size.

Languages. The languages using a given character are occasionally indicated, even though this information is incomplete. When such an annotation ends with an ellipsis (…), then the languages cited are merely the known principal ones among many.

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/), while languages such as Urdu and Farsi have 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). On the other hand, the use of these two diacritics in Arabic is redundant, merely emphasizing that the underlying vowel is long.

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

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

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

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

U+0601 ARABIC SIGN SANAH indicates a year (that is, as part of a date). This sign also renders 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, 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 the following example:

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

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

Like U+06DD ARABIC END OF AYAH, all of these signs can span multiple-digit numbers, rather than just a single digit. They are not formally considered *combining marks* in the sense used by the Unicode Standard, although they clearly interact graphically with the sequence of digits that follows them. They *precede* the sequence of digits that they span, rather than following a base character, as would be the case for a combining mark. Their General Category value is Cf (format control character), but unlike most other format con-

trol characters, they should be rendered with a visible glyph, even in circumstances where no suitable digit or sequence of digits follows them in logical order.

Poetic Verse Sign. U+060E ARABIC POETIC 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.

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

At a minimum, a display process must also select an appropriate glyph to depict each Arabic letter according to its immediate *joining* context; furthermore, it must substitute certain ligature glyphs for sequences of Arabic characters. The remainder of this section specifies a minimum set of rules that provide legible Arabic joining and ligature substitution behavior.

Cursive Joining

Joining Classes. 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 its joining class and the joining class of adjacent characters. Each Arabic character falls into one of the classes shown in *Table 8-3*. (See ArabicShaping.txt in the Unicode Character Database for a complete list.) In this table, *right* and *left* refer to visual order. The characters of the right-joining class are exemplified in more detail in *Table 8-8*, and those of the dual-joining class in *Table 8-7*.

Table 8-3. Primary Arabic Joining Classes

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

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

Table 8-4. Derived Arabic Joining Classes

Joining Class	Members
Right join-causing	Superset of dual-joining, left-joining, and join-causing
Left join-causing	Superset of dual-joining, right-joining, and join-causing

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

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

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

Table 8-5. Arabic Glyph Types

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

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

MEEM.N + SHADDA.N + LAM.N → MEEM.R + SHADDA.N + LAM.L

م + ّ + ل → مـ + ّ + ا → لّم

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

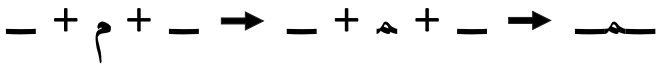
ALEF.N + TATWEEL.N → ALEF.R + TATWEEL.N

ا + ـ → ـا + ـ → ـا

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 → TATWEEL.N + MEEM.M + TATWEEL.N

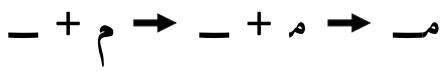


R5 ***A dual-joining character X that has a right join-causing character on the right and no left join-causing character on the left will adopt the form X_r. For example:***

MEEM.N + TATWEEL.N → MEEM.R + TATWEEL.N

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

TATWEEL.N + MEEM.N → TATWEEL.N + MEEM.L



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

As just noted, the ZERO WIDTH NON-JOINER may be used to prevent joining, as in the Persian (Farsi) plural suffix or Ottoman Turkish vowels.

Ligatures

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

For the purpose of describing the obligatory Arabic ligatures, certain Unicode characters fall into the following classes (see *Table 8-7* and *Table 8-8* for a complete list):

Alef-types:	MADDA-ON-ALEF, HAMZA ON ALEF, ...
Lam-types:	LAM, LAM WITH SMALL V, LAM WITH DOT ABOVE, ...

These two classes are designated in the joining type tables as *ALEF* and *LAM*, respectively.

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

In the ligature rules, if X and Y refer to characters, then various glyph types representing combinations of these characters are referred to as shown in *Table 8-6*.

Table 8-6. Ligature Notation

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

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

ALEF.R + FATHA.N + LAM.L → LAM-ALEF.N + 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:***

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

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

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

Arabic Character Joining Types. *Table 8-7* and *Table 8-8* provide a detailed list of the Arabic characters that are either dual-joining or right-joining, respectively. All other Arabic characters (aside from TATWEEL) are non-joining.

Most of the extended Arabic characters are merely variations on the basic Arabic shapes, with additional or different diacritic marks. When characters do not join or cause joining (such as DAMMATAN), they are classified as transparent.

The characters in these tables are grouped by shape and not by standard Arabic alphabetical order. For a machine-readable version of the information in these tables, see ArabicShaping.txt in the Unicode Character Database.

Table 8-7. Dual-Joining Arabic Characters

Group	X_n	X_r	X_m	X_l	Other Characters with Similar Shaping Behavior
BEH	ب	ب	بـ	بـ	All letters based on the BEH form, including TEH, THEH, and diacritic variants of these. 0628, 062A, 062B, 0679..0680.
NOON	ن	ن	نـ	نـ	All letters based on the NOON form. 0646, 06B9..06BD.
YEH	ي	ي	يـ	يـ	All letters based on the YEH form, including ALEF MAKSURA. 0626, 0649, 064A, 0678, 06CC, 06CE, 06D0, 06D1.
HAH	ح	ح	حـ	حـ	All letters based on the HAH form, including KHAH, JEEM, and diacritic variants of these. 062C..062E, 0681..0687, 06BF.
SEEN	س	س	سـ	سـ	All letters based on the SEEN form, including SHEEN and diacritic variants of these. 0633, 0634, 069A..069C, 06FA.
SAD	ص	ص	صـ	صـ	All letters based on the SAD form, including DAD and diacritic variants of these. 0635, 0636, 069D, 069E, 06FB.
TAH	ط	ط	طـ	طـ	All letters based on the TAH form, including ZAH and diacritic variants of these. 0637, 0638, 069F.
AIN	ع	ع	عـ	عـ	All letters based on the AIN form, including GHAIN and diacritic variants of these. 0639, 063A, 06A0, 06FC.
FEH	ف	ف	فـ	فـ	All letters based on the FEH form. 0641, 06A1..06A6.
QAF	ق	ق	قـ	قـ	All letters based on the QAF form. 0642, 06A7, 06A8.
MEEM	م	م	مـ	مـ	All letters based on the MEEM form. 0645.
HEH	ه	ه	هـ	هـ	All letters based on the HEH form. 0647.
KNOTTED HEH	ھ	ھ	ھـ	ھـ	All letters based on the KNOTTED HEH form. 06BE, 06FF.
HEH GOAL	ہ	ہ	ہـ	ہـ	All letters based on the HEH GOAL form, but excluding HAMZA ON HEH GOAL. 06C1.
KAF	ك	ك	كـ	كـ	All letters based on the KAF form. 0643, 06AC..06AE.
SWASH KAF	ک	ک	کـ	کـ	All letters based on the SWASH KAF form. 06AA.
GAF	گ	گ	گـ	گـ	All letters based on the GAF form. 06A9, 06AB, 06AF..06B4.
LAM	ل	ل	لـ	لـ	All letters based on the LAM form. 0644, 06B5..06B8.

Table 8-8. Right-Joining Arabic Characters

Group	X_n	X_r	Other Characters with Similar Shaping Behavior
ALEF	ا	ا	All letters based on the ALEF form. 0622, 0623, 0625, 0627, 0671, 0672, 0673, 0675.
WAW	و	و	All letters based on the WAW form. 0624, 0648, 0676, 0677, 06C4..06CB, 06CF.
DAL	د	د	All letters based on the DAL form, including THAL and diacritic variants of these. 062F, 0630, 0688..0690, 06EE.
REH	ر	ر	All letters based on the REH form, including ZAIN and diacritic variants of these. 0631, 0632, 0691..0699, 06EF.
TEH MARB-UTA	ة	ة	All letters based on the HEH form that show a knotted form on right-joining, including HAMZA ON HEH. 0629, 06C0, 06D5.
HAMZA ON HEH GOAL	ۂ	ۂ	All letters based on the HEH form that show a goal form on right-joining (Urdu), including HAMZA ON HEH GOAL. 06C2, 06C3.
YEH WITH TAIL	ۍ	ۍ	The tailed form of YEH. 06CD.
YEH BARREE	ے	ے	All letters based on the YEH BARREE form (Urdu). 06D2, 06D3.

In some cases, characters occur only at the end of words in correct spelling; they are called *trailing characters*. Examples include TEH MARBUTA, ALEF MAKSURA, 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.

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

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 when it is in the initial or medial position. 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.

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

This block contains a list of presentation forms (glyphs) encoded as characters for compatibility. At the time of publication, there are no known implementations of all of these presentation forms. As with most other compatibility encodings, these characters have a preferred encoding that makes use of noncompatibility characters.

The presentation forms in this block consist of contextual (positional) variants of Extended Arabic letters, contextual variants of Arabic letter ligatures, spacing forms of Arabic diacritic combinations, contextual variants of certain Arabic letter/diacritic combinations, and Arabic phrase ligatures. The ligatures include a large set of presentation forms. How-

ever, the set of ligatures appropriate for any given Arabic font will generally not match this set precisely. Fonts will often include only a subset of these glyphs, and they may also include glyphs outside of this set. These glyphs are generally not accessible as characters and are used only by rendering engines.

The alternative, ornate forms of parentheses (U+FD3E ORNATE LEFT PARENTHESIS and U+FD3F ORNATE RIGHT PARENTHESIS) for use with the Arabic script are not considered to be compatibility characters.

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

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

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

Zero Width No-Break Space. This character (U+FEFF), which is not an Arabic presentation form, is described in *Section 15.9, Specials*.

8.3 Syriac

Syriac: U+0700–U+074F

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

Syriac is divided into two dialects. West Syriac is used by the Syrian Orthodox, Maronites, and Syrian Catholics. East Syriac is used by the Assyrians (that is, Ancient Church of the East) and Chaldaeans. The two dialects are very similar with almost no difference 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* (known also 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, "The 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 writ-

ing 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, as well as 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 centuries. It 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. Additionally, it accommodates letters and diacritics used in Neo-Aramaic languages, Christian Palestinian Aramaic, Garshuni, Persian, and Sogdian languages. *Examples are supplied in the Serto type style, except where otherwise noted.*

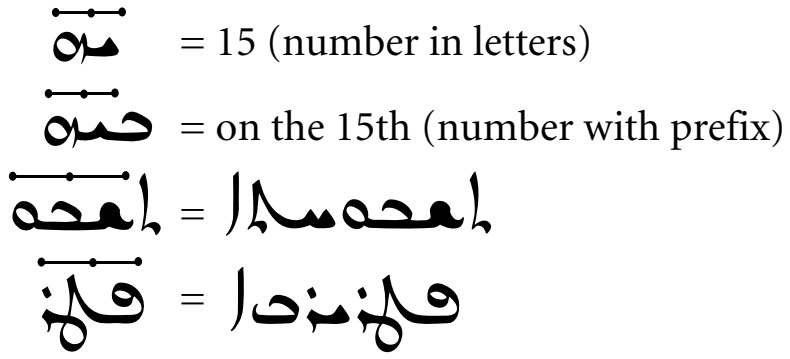
Character Names. Character names follow the East Syriac convention for naming the letters of the alphabet. Diacritical points use a descriptive naming—for example, SYRIAC DOT ABOVE.

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

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

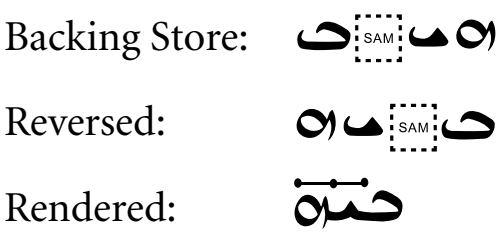
The Syriac abbreviation is used for letter numbers and contractions. A Syriac abbreviation generally extends from the last tall character in the word until the end of the word. A common exception to this rule is found with letter numbers that are preceded by a preposition character, as seen in *Figure 8-5*.

Figure 8-5. Syriac Abbreviation



A SAM is placed before the character where the abbreviation begins. The Syriac abbreviation begins over the character following the SAM and continues until the end of the word. Use of the SAM is demonstrated in *Figure 8-6*.

Figure 8-6. Use of SAM



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

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

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

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

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

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

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

Punctuation. Most punctuation marks used with Syriac are found in the Latin-1 and Arabic blocks. The other ones 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 METOBELOS marks the end of such sections.

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

Semkath. Unlike other letters, the joining mechanism of *semkath* varies through the course of history from right-joining to dual-joining. It is necessary to enter a U+200C ZERO WIDTH NON-JOINER character after the *semkath* to obtain the right-joining form where required. There are two common variants of this character: 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 also used in Syriac text. Their usage is explained in *Table 8-9*.

Table 8-9. Miscellaneous Syriac Diacritic Use

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

Use of Characters of the Arabic Block. Syriac makes use of several characters from the Arabic block, including U+0640 ARABIC TATWEEL. Modern texts use U+060C ARABIC COMMA, U+061B ARABIC SEMICOLON, and U+061F ARABIC QUESTION MARK. The *shadda* (U+0651)

is also used in the core part of literary Syriac on top of a *waw* in the word "O". Arabic *harakat* are used in Garshuni to indicate the corresponding Arabic vowels and diacritics.

Syriac Shaping

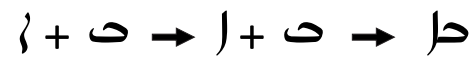
Minimum Rendering Requirements. Rendering requirements for Syriac are similar to those for Arabic. The remainder of this section specifies a minimum set of rules that provides legible Syriac joining and ligature substitution behavior.

Joining Classes. Each Syriac character is represented by up to four possible contextual glyph forms. The form used is determined by its joining class and the joining class of the letter on each side. These classes are identical in behavior to those outlined for Arabic, with the addition of three extra classes that determine the behavior of final *alaphs*. See *Table 8-10*.

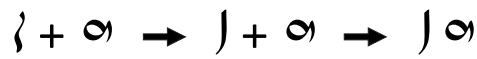
Table 8-10. Additional Syriac Joining Classes

Joining Class	Description
Afj	Final joining (alaph only)
Afn	Final non-joining *except* following dalath and rish (alaph only)
Afx	Final non-joining following dalath and rish (alaph only)

R1 ***An* alaph *that has a left-joining character to its right and a word breaking character to its left will take the form of A_{fj}.***



R2 ***An* alaph *that has a non-left-joining character to its right, except for a* dalath *or* rish*, and a word breaking character to its left will take the form of A_{fn}.***



R3 ***An* alaph *that has a* dalath *or* rish *to its right and a word breaking character to its left will take the form of A_{fx}.***

The above example is in the East Syriac font style.

Syriac Cursive Joining

Table 8-11, *Table 8-12*, and *Table 8-13* provide listings of how each character is shaped in the appropriate joining type. Syriac characters not shown are non-joining. These tables are in the Serto (West Syriac) font style, whereas the code charts in *Chapter 16, Code Charts*, are in the Estrangela font style. The shaping classes are included in ArabicShaping.txt in the Unicode Character Database.

Table 8-11. Dual-Joining Syriac Characters

Character	X_n	X_r	X_m	X_l
BETH	ܒ	ܒ	ܒ	ܒ
GAMAL	ܓ	ܓ	ܓ	ܓ
GAMAL GARSHUNI	ܓ	ܓ	ܓ	ܓ
HETH	ܚ	ܚ	ܚ	ܚ
TETH	ܛ	ܛ	ܛ	ܛ
TETH GARSHUNI	ܛ	ܛ	ܛ	ܛ
YUDH	ܝ	ܝ	ܝ	ܝ
KAPH	ܟ	ܟ	ܟ	ܟ
LAMADH	ܠ	ܠ	ܠ	ܠ
MIM	ܡ	ܡ	ܡ	ܡ
NUN	ܢ	ܢ	ܢ	ܢ
SEMKATH	ܣ	ܣ	ܣ	ܣ
SEMKATH FINAL	ܣ	ܣ	ܣ	ܣ
E	ܥ	ܥ	ܥ	ܥ
PE	ܦ	ܦ	ܦ	ܦ
PE REVERSED	ܦ	ܦ	ܦ	ܦ
QAPH	ܩ	ܩ	ܩ	ܩ
SHIN	ܫ	ܫ	ܫ	ܫ

Table 8-12. Right-Joining Syriac Characters

Character	X_n	X_r
DALATH	ܕ	ܕ
DOTLESS DALATH RISH	ܪ	ܪ
HE	ܗ	ܗ
WAW	ܘ	ܘ
ZAIN	ܙ	ܙ
YUDH HE	ܝܗ	ܝܗ
SADHE	ܨ	ܨ
RISH	ܪ	ܪ
TAW	ܬ	ܬ

Table 8-13. Alaph-Joining Syriac Characters

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

Ligatures

Ligature Classes. As in other scripts, ligatures in Syriac vary depending upon the font style. *Table 8-14* identifies the principal valid ligatures for each font style. When applicable, these ligatures are obligatory, unless denoted with an asterisk (*).

Table 8-14. Syriac Ligatures

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

8.4 Thaana

Thaana: U+0780–U+07BF

The Thaana script is used to write the modern Dhivehi language of the Republic of Maldives, a group of atolls in the Indian Ocean. Like the Arabic script, Thaana is written from right to left and uses vowel signs, but it is not cursive. The basic Thaana letters have been extended by a small set of dotted letters used to transcribe Arabic. The use of modified Thaana letters to write Arabic began in the middle of the twentieth century. Loan words from Arabic may be written in the Arabic script, although this custom is not very prevalent today. (See *Section 8.2, Arabic*.)

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

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

Vowels. Consonants are always written with either a vowel sign (U+07A6..U+07AF) or the null vowel sign (U+07B0 THAANA SUKUN). U+0787 THAANA LETTER ALIFU with the null vowel sign denotes a glottal stop. The placement of the Thaana vowel signs is shown in *Table 8-15*.

Table 8-15. Thaana Glyph Placement

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

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

Punctuation. The Thaana script uses spaces between words. It makes use of a mixture of Arabic and European punctuation, with rules of usage not clearly defined. Sentence-final punctuation is now generally shown with a single period (U+002E "." FULL STOP), but may also use a sequence of two periods (U+002E followed by U+002E). Phrases may be sepa-

rated with a comma (usually U+060C ARABIC COMMA) or with a single period (U+002E). Colons, dashes, and double quotation marks are also used in the Thaana script. In addition, Thaana makes use of U+061F ARABIC QUESTION MARK and U+061B ARABIC SEMICOLON.

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