

Chapter 8

Middle Eastern Scripts

The scripts in this category 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 all alphabetic, 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 *Section 3.12, Bidirectional Behavior*).

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 the block descriptions for Arabic and 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 include 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 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 *Section 3.12, Bidirectional Behavior*.)

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 8859-8—Part 8. *Latin/Hebrew Alphabet*. The Unicode Standard encodes the Hebrew alphabetic characters in the same relative positions as in ISO 8859-8; however, there are no points or Hebrew punctuation characters in the 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 *Chapter 2.6, Combining Characters*, and *Section 3.10, 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. Use of the two dots with any other base character is an error.

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 (U+05C1 or U+05C2). 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 8859-8. These variant forms are generally used in place of the nominal letter forms 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 from the codeset 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. Note that 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, Mirrored—Normative*, for more information.

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 *yativ*, 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.

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 more often used 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.

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 *Hebrew Presentation Forms*, U+FB1D–U+FB4F, in the names list.

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

For additional punctuation to be used with this script, see *Section 6.1, General Punctuation*. The NEW SHEQEL SIGN (U+20AA) is encoded in the currency block.

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 and 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 *Section 3.12, Bidirectional Behavior*.)

Figure 8-1. Directionality and Cursive Connection

Memory Representation:	٥٥٥ ٥
Reversal:	٥ ٥٥٥
Joining:	٥ ههه

Standards. ISO 8859-6—Part 6. *Latin/Arabic Alphabet*. The Unicode Standard encodes the basic Arabic characters in the same relative positions as in ISO 8859-6. ISO 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 them 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 13.2, Layout Controls*.

Figure 8-2. Using Joiner

Memory Representation: ٥٥٥ ZW ٥
 Reversal: ٥ ZW ٥٥٥
 Joining: ٥ هه٥

Figure 8-3. Using Non-joiner

Memory Representation: ٥ ZW
NJ ٥٥ ٥
 Reversal: ٥ ٥٥ ZW
NJ ٥
 Joining: ٥ هه٥

Figure 8-4. Combinations of Joiner and Non-joiner

Memory Representation: ٥ ZW
NJ ZW ٥٥ ٥
 Reversal: ٥ ٥٥ ZW
NJ ZW ٥
 Joining: ٥ هه٥

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 the Combining Diacritical Marks block description. 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 *Chapter 2, General Structure*, and *Chapter 3, Conformance*.)

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 ·١٢٣...). 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 and Pakistan (here called *Eastern Arabic-Indic*) and those used in other parts of the Arabic world.

The Persian and Urdu variant digits are given separate codes in the Unicode Standard to account for the differences in appearance and directional treatment when rendering them. The Unicode Standard provides one sequence of digits for both Persian and Urdu. Their presentation depends on the font. (For a complete discussion of directional formatting in the Unicode Standard, see *Section 3.12, Bidirectional Behavior*.)

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 do not have any normative content; 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

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 spurious forms. The Unicode Standard encodes multiple forms of the Extended Arabic letters and variant digits 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 two enclosing marks, U+06DD and U+06DE, are used to enclose digits. When rendered, these digits appear 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.

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 backing store and the visual (or physical) order required by the display device. (See *Section 3.12, Bidirectional Behavior*, 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 8-2* (see *ArabicShaping.txt* on the CD-ROM for a complete list). *Table 8-3* defines derived superclasses of the primary Arabic joining classes; those superclasses are used in the cursive joining rules. In both tables, *right* and *left* refer to visual order.

Table 8-2. 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, TATWEEL (<i>kashida</i>) These characters are distinguished from the dual-joining characters in that they do not change shape.
Non-joining	U	ZERO WIDTH NON-JOINER and all spacing characters (other than the above), including HAMZA, HIGH HAMZA, spaces, digits, punctuation, non-Arabic letters, and so on
Transparent	T	All combining marks and format marks, including FATHATAN, DAMMATAN, FATHAH, DAMMA, KASRAH, SHADDAH, SUKUN, ALEF ABOVE, RIGHT-LEFT MARK, and so on

Table 8-3. 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 *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 8-4*.

Table 8-4. 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)

Table 8-4. Arabic Glyph Types (Continued)

Glyph Types	Description
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 + SHADDAH.N + LAM.N → MEEM.R + SHADDAH.N + LAM.L

م + ّ + ل → م + ّ + ل

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

| + _ → | + _

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

_ + م + _ → _ + م + _

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

م + _ → م + _

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

_ + م → _ + م

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.

For the purpose of describing the obligatory Arabic ligatures, certain Unicode characters fall into the following classes (see the end of this block description 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 below as *ALEF* and *LAM*, respectively.

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

In the following rules, if X and Y refer to characters, then various glyph types representing combinations of these characters are referred to as shown in *Table 8-5*.

Table 8-5. 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:

$ALEF.R + FATHAH.N + LAM.L \rightarrow LAM-ALEF.N + FATHAH.N$

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

$\text{ا} + \text{ل} \rightarrow \text{لا}$ (not لا)

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

$\text{ا} + \text{ل} \rightarrow \text{لا}$ (not لا)

- From the perspective of logical (or reading) order, the preceding $(ALEF.LAM)$ ligature forms are referred to as $LAM-ALEF$ forms. The difference reflects the use of visual order rather than logical order to state ligature rules.

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..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 8-6, Table 8-7, and Table 8-8 provide a detailed list of the Arabic characters that are either right-joining or dual-joining. All other Arabic characters (aside from `TATWEEL`) are non-joining, including `U+06D5 ARABIC LETTER AE`. For brevity in the names, the words `TWO`, `THREE`, and `FOUR` are abbreviated using digits.

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

In some cases, 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 classed as right-joining, even when similarly shaped characters are dual-joining.

When characters do not join or cause joining (such as `DAMMATAN`), they are classified as transparent.

- In the case of `U+0647 HEH`, the glyph `HEH1` 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 nominal form of `HEH` is the isolate form, which looks like `U+06D5 ARABIC LETTER AE`. In the case of `U+06C1 HEH GOAL`, the nominal form is not even listed; it also resembles `ARABIC LETTER AE`.

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 `Arabic-Shaping.txt` on the CD-ROM.

Table 8-6. Dual-Joining Arabic Characters

Group	X _n	X _r	X _m	X _l	Other Characters with Similar Shaping Behavior
BEH	ب	ب	ب	ب	All letters based on the BEH form, including <code>TEH</code> , <code>THEH</code> , and diacritic variants of these. 0628, 062A, 062B, 0679..0680.
NOON	ن	ن	ن	ن	All letters based on the NOON form. 0646, 06B9..06BD.
YEH	ي	ي	ي	ي	All letters based on the YEH form, including <code>ALEF MAKSURA</code> . 0626, 0649, 064A, 0678, 06CC, 06CE, 06D0, 06D1.
HAH	ح	ح	ح	ح	All letters based on the HAH form, including <code>KHAH</code> , <code>JEEM</code> , and diacritic variants of these. 062C..062E, 0681..0687, 06BF.
SEEN	س	س	س	س	All letters based on the SEEN form, including <code>SHEEN</code> and diacritic variants of these. 0633, 0634, 069A..069C, 06FA.
SAD	ص	ص	ص	ص	All letters based on the SAD form, including <code>DAD</code> and diacritic variants of these. 0635, 0636, 069D, 069E, 06FB.

Table 8-6. Dual-Joining Arabic Characters (Continued)

Group	X _n	X _r	X _m	X _l	Other Characters with Similar Shaping Behavior
TAH	ط	ط	ط	ط	All letters based on the TAH form, including ZAH and diacritic variants of these. 0637, 0638, 069F.
AIN	ع	ع	ع	ع	All letters based on the AIN form, including GHAIN and diacritic variants of these. 0639, 063A, 06A0, 06FC.
FEH	فا	فا	فا	فا	All letters based on the FEH form. 0641, 06A1..06A6.
QAF	ق	ق	قا	قا	All letters based on the QAF form. 0642, 06A7, 06A8.
MEEM	م	م	م	م	All letters based on the MEEM form. 0645.
HEH	ه	ه	ه	ه	All letters based on the HEH form. 0647.
KNOTTED HEH	ه	ه	ه	ه	All letters based on the KNOTTED HEH form. 06BE.
HEH GOAL	ه	ه	ه	ه	All letters based on the HEH GOAL form, but excluding HAMZA ON HEH GOAL. 06C1.
KAF	ك	ك	ك	ك	All letters based on the KAF form. 0643, 06AC..06AE.
SWASH KAF	ك	ك	ك	ك	All letters based on the SWASH KAF form. 06AA.
GAF	گ	گ	گ	گ	All letters based on the GAF form. 06A9, 06AB, 06AF..06B4.
LAM	ل	ل	ل	ل	All letters based on the LAM form. 0644, 06B5..06B8.

Table 8-7. 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.
REH	ر	ر	All letters based on the REH form, including ZAIN and diacritic variants of these. 0631, 0632, 0691..0699.
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.

Table 8-7. Right-Joining Arabic Characters (Continued)

Group	X _n	X _r	Other Characters with Similar Shaping Behavior
HAMZA ON HEH GOAL	ه	ه	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	ي	ي	The tailed form of YEH. 06CD.
YEH BARREE	ي	ي	All letters based on the YEH BARREE form (Urdu). 06D2, 06D3.

Table 8-8. Other Arabic Character Joining Classes

Class	Members
Join-causing	ZERO WIDTH JOINER (200D) and TATWEEL (0640).
Non-joining	ZERO WIDTH NON-JOINER (200C), and all spacing characters (other than those mentioned in <i>Table 8-6</i> or <i>Table 8-7</i>) are non-joining, including HAMZA (0621), HIGH HAMZA (0674), spaces, digits, punctuation, non-Arabic letters, and so on.
Transparent	All combining marks and other format controls are irrelevant, including FATHATAN (064B) and other points, MADDAH ABOVE (0653), HAMZA ABOVE (0654), HAMZA BELOW (0655), SUPERSCRIPT ALEF (0670), combining Koranic annotation signs, RIGHT-TO-LEFT MARK (200F), and so on.

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

The alternative (ornate) forms of parentheses for use with the Arabic script are not 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 the Specials block (U+FF00..U+FFFF).

8.3 Syriac

Syriac: U+0700–U+074F

The Syriac language belongs to the Aramaic branch of the family of Semitic 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.

Syriac Languages. 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.

The Syriac Script. 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 characters.

Directionality. The Syriac script is written from right to left. Conformant implementations of Syriac script must use the Unicode bidirectional algorithm (see *Section 3.12, Bidirectional Behavior*).

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, 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, and Garshuni 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 user-selectable 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


ⲛⲉ	= 15 (number in letters)
ⲛⲉ	= on the 15th (number with prefix)
ⲛⲉ	= ⲛⲉⲛⲉⲛⲉ
ⲛⲉ	= ⲛⲉⲛⲉⲛⲉ

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

Backing Store:  SAM 

Reversed: 

Rendered: 

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 (like 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), *galath*, *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 used 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 numbers 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, 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. As shown in the preceding examples, 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 Diacritic 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 rukakha points, as well as the dotted vowel points.
U+0308	This is the plural marker. It is also used in Garshuni for the Arabic <i>teh marbuta</i> .
U+030A, U+0325	These are two other forms for the indication of qushshaya and rukakha. They are used interchangeably with U+0741 SYRIAC QUSHSHAYA and U+0742 SYRIAC RUKAKHA, 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 <i>digit markers</i> .
U+032E	This is a mark used in late and modern East Syriac texts, as well as Swadaya, to indicate a fricative <i>pe</i> .

Use of Characters of the Arabic Block. Syriac makes use of several characters from the Arabic block, including tatweel (U+0640). 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 <i>except</i> 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. The shaping classes are included in ArabicShaping.txt on the CD-ROM.

Table 8-11. Right-Joining Syriac Characters

Character	X _n	X _r
DALATH	ܱ	ܲ
DOTLESS DALATH RISH	ܱ	ܳ
HE	ܵ	ܶ
WAW	ܷ	ܸ

Table 8-11. Right-Joining Syriac Characters (Continued)

Character	X _n	X _r
ZAIN	ܐ	ܐ
YUDH HE	ܘܗ	ܘܗ
SADHE	ܘܫ	ܘܫ
RISH	ܘܫ	ܘܫ
TAW	ܐ	ܐ

Table 8-12. 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	ܗ	ܗ	ܗ	ܗ
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	ܦ	ܦ	ܦ	ܦ
SEMKATH FINAL	ܦ	ܦ	ܦ	ܦ
E	ܐ	ܐ	ܐ	ܐ
PE	ܦ	ܦ	ܦ	ܦ
PE REVERSED	ܦ	ܦ	ܦ	ܦ
QAPH	ܩ	ܩ	ܩ	ܩ
SHIN	ܫ	ܫ	ܫ	ܫ

Table 8-13. Alph-Joining Syriac Characters

Type Style	A _n	A _r	A _{fj}	A _{fn}	A _{fx}
Estrangela	Ⲁ	Ⲁ	Ⲁ	Ⲁ	Ⲁ
Serto (West Syriac)	Ⲁ	Ⲁ	Ⲁ	Ⲁ	Ⲁ
East Syriac	Ⲁ	Ⲁ	Ⲁ	Ⲁ	Ⲁ

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 dates from 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*.)

Directionality. The Thaana script is written from right to left. Conformant implementations of Thaana script must use the Unicode bidirectional algorithm (see *Section 3.12, Bidirectional Behavior*).

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.

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

Character Names and Arrangement. The character names are based on the names used in the Republic of Maldives. The character name at U+0794, *yaa*, is found in some sources as *yaviyani*, but the former 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*.

This PDF file is an excerpt from *The Unicode Standard, Version 3.0*, 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 (see <http://www.unicode.org/unicode/uni2errata/UnicodeErrata.html>). More recent versions of the Unicode standard exist (see <http://www.unicode.org/unicode/standard/versions/>).

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and Addison-Wesley was aware of a trademark claim, the designations have been printed in initial capital letters. However, not all words in initial capital letters are trademark designations.

The authors and publisher have taken care in preparation of this book, but make no expressed or implied warranty of any kind and assume no responsibility for errors or omissions. No liability is assumed for incidental or consequential damages in connection with or arising out of the use of the information or programs contained herein.

The *Unicode Character Database* and other files are provided as-is by Unicode®, Inc. No claims are made as to fitness for any particular purpose. No warranties of any kind are expressed or implied. The recipient agrees to determine applicability of information provided.

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

ISBN 0-201-61633-5

Copyright © 1991-2000 by Unicode, Inc.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior written permission of the publisher or Unicode, Inc.

This book is set in Minion, designed by Rob Slimbach at Adobe Systems, Inc. It was typeset using FrameMaker 5.5 running under Windows NT. ASMUS, Inc. created custom software for chart layout. The Han radical-stroke index was typeset by Apple Computer, Inc. The following companies and organizations supplied fonts:

Apple Computer, Inc.
Atelier Fluxus Virus
Beijing Zhong Yi (Zheng Code) Electronics Company
DecoType, Inc.
IBM Corporation
Monotype Typography, Inc.
Microsoft Corporation
Peking University Founder Group Corporation
Production First Software

Additional fonts were supplied by individuals as listed in the *Acknowledgments*.

The Unicode® Consortium is a registered trademark, and Unicode™ is a trademark of Unicode, Inc. The Unicode logo is a trademark of Unicode, Inc., and may be registered in some jurisdictions.

All other company and product names are trademarks or registered trademarks of the company or manufacturer, respectively.

The publisher offers discounts on this book when ordered in quantity for special sales. For more information please contact:

Corporate, Government, and Special Sales
Addison Wesley Longman, Inc.
One Jacob Way
Reading, Massachusetts 01867

Visit A-W on the Web: <http://www.awl.com/cseng/>

First printing, January 2000.