

### ***Electronic Edition***

This file is part of the electronic edition of *The Unicode Standard, Version 5.0*, provided for online access, content searching, and accessibility. It may not be printed. Bookmarks linking to specific chapters or sections of the whole Unicode Standard are available at

<http://www.unicode.org/versions/Unicode5.0.0/bookmarks.html>

### ***Purchasing the Book***

For convenient access to the full text of the standard as a useful reference book, we recommend purchasing the printed version. The book is available from the Unicode Consortium, the publisher, and booksellers. Purchase of the standard in book format contributes to the ongoing work of the Unicode Consortium. Details about the book publication and ordering information may be found at

<http://www.unicode.org/book/aboutbook.html>

### ***Joining Unicode***

You or your organization may benefit by joining the Unicode Consortium: for more information, see [Joining the Unicode Consortium](http://www.unicode.org/consortium/join.html) at

<http://www.unicode.org/consortium/join.html>

This PDF file is an excerpt from *The Unicode Standard, Version 5.0*, issued by the Unicode Consortium and published by Addison-Wesley. The material has been modified slightly for this electronic 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/versions/enumeratedversions.html>.

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 the publisher was aware of a trademark claim, the designations have been printed with initial capital letters or in all capitals.

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.

The authors and publisher have taken care in the 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.

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

The publisher offers excellent discounts on this book when ordered in quantity for bulk purchases or special sales, which may include electronic versions and/or custom covers and content particular to your business, training goals, marketing focus, and branding interests. For more information, please contact U.S. Corporate and Government Sales, (800) 382-3419, [corpsales@pearsoned.com](mailto:corpsales@pearsoned.com). For sales outside the United States please contact International Sales, [international@pearsoned.com](mailto:international@pearsoned.com)

Visit us on the Web: [www.awprofessional.com](http://www.awprofessional.com)

*Library of Congress Cataloging-in-Publication Data*

The Unicode Standard / the Unicode Consortium ; edited by Julie D. Allen ... [et al.]. — Version 5.0.  
p. cm.

Includes bibliographical references and index.

ISBN 0-321-48091-0 (hardcover : alk. paper)

1. Unicode (Computer character set) I. Allen, Julie D.

II. Unicode Consortium.

QA268.U545 2007

005.7'22—dc22

2006023526

Copyright © 1991–2007 Unicode, Inc.

All rights reserved. Printed in the United States of America. This publication is protected by copyright, and permission must be obtained from the publisher prior to any prohibited reproduction, storage in a retrieval system, or transmission in any form or by any means, electronic, mechanical, photocopying, recording, or likewise. For information regarding permissions, write to Pearson Education, Inc., Rights and Contracts Department, 75 Arlington Street, Suite 300, Boston, MA 02116. Fax: (617) 848-7047

ISBN 0-321-48091-0

Text printed in the United States on recycled paper at Courier in Westford, Massachusetts.

First printing, October 2006

## Chapter 14

# *Archaic Scripts*

The following historic scripts are encoded in Version 5.0 of the Unicode Standard:

<i>Ogham</i>	<i>Linear B</i>	<i>Ugaritic</i>
<i>Old Italic</i>	<i>Cypriot</i>	<i>Old Persian</i>
<i>Runic</i>	<i>Phoenician</i>	<i>Sumero-Akkadian</i>
<i>Gothic</i>		

Unicode encodes a number of historic scripts. Although they are no longer used to write living languages, documents and inscriptions using these scripts exist, both for extinct and precursors of modern languages. The primary user communities for these scripts are scholars interested in studying the scripts and the languages written in them. Some of the historical scripts are related to each other and to modern alphabets.

The Ogham script is indigenous to Ireland. While its originators may have been aware of the Latin or Greek scripts, it seems clear that the sound values of Ogham letters were suited to the phonology of a form of Primitive Irish.

Old Italic was derived from Greek and was used to write Etruscan and other languages in Italy. It was borrowed by the Romans and is the immediate ancestor of the Latin script now used worldwide. Old Italic had other descendants, too: The Alpine alphabets seem to have been influential in devising the Runic script, which has a distinct angular appearance owing to its use in carving inscriptions in stone and wood. Gothic, like Cyrillic, was developed on the basis of Greek at a much later date than Old Italic.

The two historic scripts of northwestern Europe, Runic and Ogham, have a distinct appearance owing to their primary use in carving inscriptions in stone and wood. They are conventionally rendered from left to right in scholarly literature, but on the original stone carvings often proceeded in an arch tracing the outline of the stone.

Both Linear B and Cypriot are syllabaries that were used to write Greek. Linear B is the older of the two scripts, and there are some similarities between a few of the characters that may not be accidental. Cypriot may descend from Cypro-Minoan, which in turn may descend from Linear B.

The Phoenician alphabet was used in various forms around the Mediterranean. It is ancestral to Latin, Greek, Hebrew, and many other scripts both modern and historical.

Three ancient cuneiform scripts are described in this chapter: Ugaritic, Old Persian, and Sumero-Akkadian. The largest and oldest of these is Sumero-Akkadian. The other two scripts are not derived directly from the Sumero-Akkadian tradition but had common writing technology, consisting of wedges indented into clay tablets with reed styluses. Ugaritic texts are about as old as the earliest extant Biblical texts. Old Persian texts are newer, dating from the fifth century BCE.

---

## 14.1 Ogham

### **Ogham: U+1680–U+169F**

Ogham is an alphabetic script devised to write a very early form of Irish. Monumental Ogham inscriptions are found in Ireland, Wales, Scotland, England, and on the Isle of Man. Many of the Scottish inscriptions are undeciphered and may be in Pictish. It is probable that Ogham (Old Irish “Ogam”) was widely written in wood in early times. The main flowering of “classical” Ogham, rendered in monumental stone, was in the fifth and sixth centuries CE. Such inscriptions were mainly employed as territorial markers and memorials; the more ancient examples are standing stones.

The script was originally written along the edges of stone where two faces meet; when written on paper, the central “stemlines” of the script can be said to represent the edge of the stone. Inscriptions written on stemlines cut into the face of the stone, instead of along its edge, are known as “scholastic” and are of a later date (post-seventh century). Notes were also commonly written in Ogham in manuscripts as recently as the sixteenth century.

**Structure.** The Ogham alphabet consists of 26 distinct characters (*fedá*), the first 20 of which are considered to be primary and the last 6 (*forfedá*) supplementary. The four primary series are called *aicmí* (plural of *aicme*, meaning “family”). Each *aicme* was named after its first character, (*Aicme Beithe*, *Aicme Uatha*, meaning “the B Family,” “the H Family,” and so forth). The character names used in this standard reflect the spelling of the names in modern Irish Gaelic, except that the acute accent is stripped from *Úr*, *Éabhadh*, *Ór*, and *Ifin*, and the mutation of *nGéadal* is not reflected.

**Rendering.** Ogham text is read beginning from the bottom left side of a stone, continuing upward, across the top, and down the right side (in the case of long inscriptions). Monumental Ogham was incised chiefly in a bottom-to-top direction, though there are examples of left-to-right bilingual inscriptions in Irish and Latin. Manuscript Ogham accommodated the horizontal left-to-right direction of the Latin script, and the vowels were written as vertical strokes as opposed to the incised notches of the inscriptions. Ogham should therefore be rendered on computers from left to right or from bottom to top (never starting from top to bottom).

**Forfedá (Supplementary Characters).** In printed and in manuscript Ogham, the fonts are conventionally designed with a central stemline, but this convention is not necessary. In implementations without the stemline, the character U+1680 OGHAM SPACE MARK should



Venetic, Rhaetic, Lepontic, and Gallic alphabets, has not yet been proposed for addition to the Unicode Standard but is considered to differ enough from both Old Italic and Latin to warrant independent encoding. The Alpine script is thought to be the source for Runic, which is encoded at U+16A0..U+16FF. (See *Section 14.3, Runic.*)

Character names assigned to the Old Italic block are unattested but have been reconstructed according to the analysis made by Sampson (1985). While the Greek character names (ALPHA, BETA, GAMMA, and so on) were borrowed directly from the Phoenician names (modified to Greek phonology), the Etruscans are thought to have abandoned the Greek names in favor of a phonetically based nomenclature, where stops were pronounced with a following -e sound, and liquids and sibilants (which can be pronounced more or less on their own) were pronounced with a leading e- sound (so [k], [d] became [ke:], [de:] became [l:], [m:] became [el], [em]). It is these names, according to Sampson, which were borrowed by the Romans when they took their script from the Etruscans.

**Directionality.** Most early Etruscan texts have right-to-left directionality. From the third century BCE, left-to-right texts appear, showing the influence of Latin. Oscan, Umbrian, and Faliscan also generally have right-to-left directionality. Boustrophedon appears rarely, and not especially early (for instance, the Forum inscription dates to 550–500 BCE). Despite this, for reasons of implementation simplicity, many scholars prefer left-to-right presentation of texts, as this is also their practice when transcribing the texts into Latin script. Accordingly, the Old Italic script has a default directionality of strong left-to-right in this standard. If the default directionality of the script is overridden to produce a right-to-left presentation, the glyphs in Old Italic fonts should also be mirrored from the representative glyphs shown in the code charts. This kind of behavior is not uncommon in archaic scripts; for example, archaic Greek letters may be mirrored when written from right to left in boustrophedon.

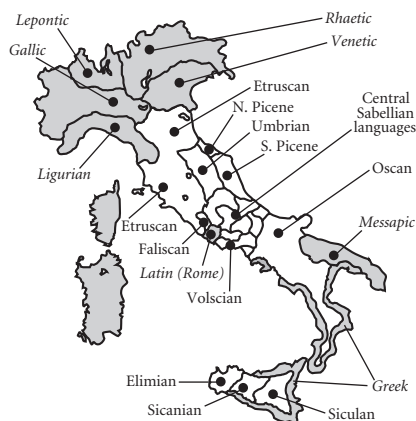
**Punctuation.** The earliest inscriptions are written with no space between words in what is called *scriptio continua*. There are numerous Etruscan inscriptions with dots separating word forms, attested as early as the second quarter of the seventh century BCE. This punctuation is sometimes, but only rarely, used to separate syllables rather than words. From the sixth century BCE, words were often separated by one, two, or three dots spaced vertically above each other.

**Numerals.** Etruscan numerals are not well attested in the available materials, but are employed in the same fashion as Roman numerals. Several additional numerals are attested, but as their use is at present uncertain, they are not yet encoded in the Unicode Standard.

**Glyphs.** The default glyphs in the code charts are based on the most common shapes found for each letter. Most of these are similar to the Marsiliana abecedary (mid-seventh century BCE). Note that the phonetic values for U+10317 OLD ITALIC LETTER EKS [ks] and U+10319 OLD ITALIC LETTER KHE [kh] show the influence of western, Euboean Greek; eastern Greek has U+03A7 GREEK CAPITAL LETTER CHI [x] and U+03A8 GREEK CAPITAL LETTER PSI [ps] instead.

The geographic distribution of the Old Italic script is shown in *Figure 14-1*. In the figure, the approximate distribution of the ancient languages that used Old Italic alphabets is shown in white. Areas for the ancient languages that used other scripts are shown in gray, and the labels for those languages are shown in oblique type. In particular, note that the ancient Greek colonies of the southern Italian and Sicilian coasts used the Greek script proper. Also, languages such as Ligurian, Venetic, and so on, of the far north of Italy made use of alphabets of the Alpine script. Rome, of course, is shown in gray, because Latin was written with the Latin alphabet, now encoded in the Latin script.

**Figure 14-1.** Distribution of Old Italic



## 14.3 Runic

### **Runic:** *U+16A0–U+16F0*

The Runic script was historically used to write the languages of the early and medieval societies in the German, Scandinavian, and Anglo-Saxon areas. Use of the Runic script in various forms covers a period from the first century to the nineteenth century. Some 6,000 Runic inscriptions are known. They form an indispensable source of information about the development of the Germanic languages.

**Historical Script.** The Runic script is an historical script, whose most important use today is in scholarly and popular works about the old Runic inscriptions and their interpretation. The Runic script illustrates many technical problems that are typical for this kind of script. Unlike many other scripts in the Unicode Standard, which predominantly serve the needs of the modern user community—with occasional extensions for historic forms—the encoding of the Runic script attempts to suit the needs of texts from different periods of time and from distinct societies that had little contact with one another.

**Direction.** Like other early writing systems, runes could be written either from left to right or from right to left, or moving first in one direction and then the other (*boustrophedon*), or following the outlines of the inscribed object. At times, characters appear in mirror image, or upside down, or both. In modern scholarly literature, Runic is written from left to right. Therefore, the letters of the Runic script have a default directionality of strong left-to-right in this standard.

**The Runic Alphabet.** Present-day knowledge about runes is incomplete. The set of graphemically distinct units shows greater variation in its graphical shapes than most modern scripts. The Runic alphabet changed several times during its history, both in the number and the shapes of the letters contained in it. The shapes of most runes can be related to some Latin capital letter, but not necessarily to a letter representing the same sound. The most conspicuous difference between the Latin and the Runic alphabets is the order of the letters.

The Runic alphabet is known as the *futhark* from the name of its first six letters. The original *old futhark* contained 24 runes:

Ʊ ᚋ ᚔ ᚕ ᚖ ᚗ ᚘ ᚙ ᚛ ᚜ ᚝ ᚞ ᚟ ᚠ ᚡ ᚢ ᚣ ᚤ ᚥ ᚦ ᚧ ᚨ ᚩ ᚪ ᚫ

They are usually transliterated in this way:

f u þ a r k g w h n i j ï p z s t b e m l ŋ d o

In England and Friesland, seven more runes were added from the fifth to the ninth century.

In the Scandinavian countries, the *futhark* changed in a different way; in the eighth century, the simplified younger *futhark* appeared. It consists of only 16 runes, some of which are used in two different forms. The long-branch form is shown here:

Ʊ ᚋ ᚔ ᚕ ᚖ ᚗ ᚘ ᚙ ᚛ ᚜ ᚝ ᚞ ᚟ ᚠ ᚡ ᚢ ᚣ ᚤ ᚥ ᚦ ᚧ ᚨ ᚩ ᚪ ᚫ

f u þ o r k h n i a s t b m l r

The use of runes continued in Scandinavia during the Middle Ages. During that time, the *futhark* was influenced by the Latin alphabet and new runes were invented so that there was full correspondence with the Latin letters.

**Representative Glyphs.** The known inscriptions can include considerable variations of shape for a given rune, sometimes to the point where the nonspecialist will mistake the shape for a different rune. There is no dominant main form for some runes, particularly for many runes added in the Anglo-Friesian and medieval Nordic systems. When transcribing a Runic inscription into its Unicode-encoded form, one cannot rely on the idealized *representative glyph* shape in the character charts alone. One must take into account to which of the four Runic systems an inscription belongs and be knowledgeable about the permitted form variations within each system. The representative glyphs were chosen to provide an image that distinguishes each rune visually from all other runes in the same system. For actual use, it might be advisable to use a separate font for each Runic system. Of particular note is the fact that the glyph for U+16C4 Ʊ RUNIC LETTER GER is actually a rare form, as the more common form is already used for U+16E1 Ʊ RUNIC LETTER IOR.



**Unifications.** When a rune in an earlier writing system evolved into several different runes in a later system, the unification of the earlier rune with one of the later runes was based on similarity in graphic form rather than similarity in sound value. In cases where a substantial change in the typical graphical form has occurred, though the historical continuity is undisputed, unification has not been attempted. When runes from different writing systems have the same graphic form but different origins and denote different sounds, they have been coded as separate characters.

**Long-Branch and Short-Twig.** Two sharply different graphic forms, the *long-branch* and the *short-twig* form, were used for 9 of the 16 Viking Age Nordic runes. Although only one form is used in a given inscription, there are runologically important exceptions. In some cases, the two forms were used to convey different meanings in later use in the medieval system. Therefore the two forms have been separated in the Unicode Standard.

**Staveless Runes.** Staveless runes are a third form of the Viking Age Nordic runes, a kind of Runic shorthand. The number of known inscriptions is small and the graphic forms of many of the runes show great variability between inscriptions. For this reason, staveless runes have been unified with the corresponding Viking Age Nordic runes. The corresponding Viking Age Nordic runes must be used to encode these characters—specifically the short-twig characters, where both short-twig and long-branch characters exist.

**Punctuation Marks.** The wide variety of Runic punctuation marks has been reduced to three distinct characters based on simple aspects of their graphical form, as very little is known about any difference in intended meaning between marks that look different. Any other punctuation marks have been unified with shared punctuation marks elsewhere in the Unicode Standard.

**Golden Numbers.** Runes were used as symbols for Sunday letters and golden numbers on calendar staves used in Scandinavia during the Middle Ages. To complete the number series 1–19, three more calendar runes were added. They are included after the punctuation marks.

**Encoding.** A total of 81 characters of the Runic script are included in the Unicode Standard. Of these, 75 are Runic letters, 3 are punctuation marks, and 3 are Runic symbols. The order of the Runic characters follows the traditional *futhark* order, with variants and derived runes being inserted directly after the corresponding ancestor.

Runic character names are based as much as possible on the sometimes several traditional names for each rune, often with the Latin transliteration at the end of the name.

---

## 14.4 Gothic

### **Gothic: U+10330–U+1034F**

The Gothic script was devised in the fourth century by the Gothic bishop, Wulfila (311–383 CE), to provide his people with a written language and a means of reading his translation of

the Bible. Written Gothic materials are largely restricted to fragments of Wulfila’s translation of the Bible; these fragments are of considerable importance in New Testament textual studies. The chief manuscript, kept at Uppsala, is the Codex Argenteus or “the Silver Book,” which is partly written in gold on purple parchment. Gothic is an East Germanic language; this branch of Germanic has died out and thus the Gothic texts are of great importance in historical and comparative linguistics. Wulfila appears to have used the Greek script as a source for the Gothic, as can be seen from the basic alphabetical order. Some of the character shapes suggest Runic or Latin influence, but this is apparently coincidental.

**Diacritics.** The tenth letter U+10339 GOTHIC LETTER EIS is used with U+0308 COMBINING DIAERESIS when word-initial, when syllable-initial after a vowel, and in compounds with a verb as second member as shown below:

SYE ƆAMELIƆ IST IN ESAIIN PRAUFETAU  
*swe gameliþ Ɔst in esaiin praufetau*  
 “as is written in Isaiah the prophet”

To indicate contractions or omitted letters, U+0305 COMBINING OVERLINE is used.

**Numerals.** Gothic letters, like those of other early Western alphabets, can be used as numbers; two of the characters have only a numeric value and are not used alphabetically. To indicate numeric use of a letter, it is either flanked on one side by U+00B7 MIDDLE DOT or followed by both U+0304 COMBINING MACRON and U+0331 COMBINING MACRON BELOW, as shown in the following example:

•Ɔ or Ɔ̄ means “5”

**Punctuation.** Gothic manuscripts are written with no space between words in what is called *scriptio continua*. Sentences and major phrases are often separated by U+0020 SPACE, U+00B7 MIDDLE DOT, or U+003A COLON.

## 14.5 Linear B

### *Linear B Syllabary: U+10000–U+1007F*

The Linear B script is a syllabic writing system that was used on the island of Crete and parts of the nearby mainland to write the oldest recorded variety of the Greek language. Linear B clay tablets predate Homeric Greek by some 700 years; the latest tablets date from the mid- to late thirteenth century BCE. Major archaeological sites include Knossos, first uncovered about 1900 by Sir Arthur Evans, and a major site near Pylos. The majority of currently known inscriptions are inventories of commodities and accounting records.

Early attempts to decipher the script failed until Michael Ventris, an architect and amateur decipherer, came to the realization that the language might be Greek and not, as previously thought, a completely unknown language. Ventris worked together with John Chadwick, and decipherment proceeded quickly. The two published a joint paper in 1953.

Linear B was written from left to right with no nonspacing marks. The script mainly consists of phonetic signs representing the combination of a consonant and a vowel. There are about 60 known phonetic signs, in addition to a few signs that seem to be mainly free variants (also known as Chadwick’s optional signs), a few unidentified signs, numerals, and a number of ideographic signs, which were used mainly as counters for commodities. Some ligatures formed from combinations of syllables were apparently used as well. Chadwick gives several examples of these ligatures, the most common of which are included in the Unicode Standard. Other ligatures are the responsibility of the rendering system.

**Standards.** The catalog numbers used in the Unicode character names for Linear B syllables are based on the Wingspread Convention, as documented in Bennett (1964). The letter “B” is prepended arbitrarily, so that name parts will not start with a digit, thus conforming to ISO/IEC 10646 naming rules. The same naming conventions, using catalog numbers based on the Wingspread Convention, are used for Linear B ideograms.

### ***Linear B Ideograms: U+10080–U+108FF***

The Linear B Ideograms block contains the list of Linear B signs known to constitute ideograms (logographs), rather than syllables. When generally agreed upon, the names include the meaning associated with them—for example, U+10080 Ṛ LINEAR B IDEOGRAM B100 MAN. In other instances, the names of the ideograms simply carry their catalog number.

### ***Aegean Numbers: U+10100–U+1013F***

The signs used to denote Aegean whole numbers (U+10107..U+10133) derive from the non-Greek Linear A script. The signs are used in Linear B. The Cypriot syllabary appears to use the same system, as evidenced by the fact that the lower digits appear in extant texts. For measurements of agricultural and industrial products, Linear B uses three series of signs: liquid measures, dry measures, and weights. No set of signs for linear measurement has been found yet. Liquid and dry measures share the same symbols for the two smaller subunits; the system of weights retains its own unique subunits. Though several of the signs originate in Linear A, the measuring system of Linear B differs from that of Linear A. Linear B relies on units and subunits, much like the imperial “quart,” “pint,” and “cup,” whereas Linear A uses whole numbers and fractions. The absolute values of the measurements have not yet been completely agreed upon.

## **14.6 Cypriot Syllabary**

### ***Cypriot Syllabary: U+10800–U+1083F***

The Cypriot syllabary was used to write the Cypriot dialect of Greek from about 800 to 200 BCE. It is related to both Linear B and Cypro-Minoan, a script used for a language that has not yet been identified. Interpretation has been aided by the fact that, as use of the Cypriot syllabary died out, inscriptions were carved using both the Greek alphabet and the Cypriot

syllabary. Unlike Linear B and Cypro-Minoan, the Cypriot syllabary was usually written from right to left, and accordingly the characters in this script have strong right-to-left directionality.

Word breaks can be indicated by spaces or by separating punctuation, although separating punctuation is also used between larger word groups.

Although both Linear B and the Cypriot syllabary were used to write Greek dialects, Linear B has a more highly abbreviated spelling. Structurally, the Cypriot syllabary consists of combinations of up to 12 initial consonants and 5 different vowels. Long and short vowels are not distinguished. The Cypriot syllabary distinguishes among a different set of initial consonants than Linear B; for example, unlike Linear B, Cypriot maintained a distinction between [l] and [r], though not between [d] and [t], as shown in *Table 14-1*. Not all of the 60 possible consonant-vowel combinations are represented. As is the case for Linear B, the Cypriot syllabary is well understood and documented.

**Table 14-1.** Similar Characters in Linear B and Cypriot

Linear B	Cypriot
da    𐀄	ta    𐀄
na    𐀅	na    𐀅
pa    𐀆	pa    𐀆
ro    𐀇	lo    𐀇
se    𐀈	se    𐀈
ti    𐀉	ti    𐀉
to    𐀊	to    𐀊

For Aegean numbers, see the subsection “Aegean Numbers: U+10100–U+1013F” in *Section 14.5, Linear B*.

---

## 14.7 Phoenician

### *Phoenician: U+10900–U+1091F*

The Phoenician alphabet and its successors were widely used over a broad area surrounding the Mediterranean Sea. Phoenician evolved over the period from about the twelfth century BCE until the second century BCE, with the last neo-Punic inscriptions dating from about the third century CE. Phoenician came into its own from the ninth century BCE. An older form of the Phoenician alphabet is a forerunner of the Greek, Old Italic (Etruscan),

Latin, Hebrew, Arabic, and Syriac scripts among others, many of which are still in modern use. It has also been suggested that Phoenician is the ultimate source of Kharoshthi and of the Indic scripts descending from Brahmi.

Phoenician is an historic script, and as for many other historic scripts, which often saw continuous change in use over periods of hundreds or thousands of years, its delineation as a script is somewhat problematic. This issue is particularly acute for historic Semitic scripts, which share basically identical repertoires of letters, which are historically related to each other, and which were used to write closely related Semitic languages.

In the Unicode Standard, the Phoenician script is intended for the representation of text in Palaeo-Hebrew, Archaic Phoenician, Phoenician, Early Aramaic, Late Phoenician cursive, Phoenician papyri, Siloam Hebrew, Hebrew seals, Ammonite, Moabite, and Punic. The line from Phoenician to Punic is taken to constitute a single continuous branch of script evolution, distinct from that of other related but separately encoded Semitic scripts.

The earliest Hebrew language texts were written in the Palaeo-Hebrew alphabet, one of the forms of writing considered to be encompassed within the Phoenician script as encoded in the Unicode Standard. The Samaritans who did not go into exile continued to use Palaeo-Hebrew forms, eventually developing them into the distinct Samaritan script, which is not yet encoded in the Unicode Standard. The Jews in exile gave up the Palaeo-Hebrew alphabet and instead adopted Imperial Aramaic writing, which was a descendant of the Early Aramaic form of the Phoenician script. Later, they transformed Imperial Aramaic into the “Jewish Aramaic” script now called (Square) Hebrew, separately encoded in the Hebrew block in the Unicode Standard.

Some scholars conceive of the language written in the Palaeo-Hebrew form of the Phoenician script as being quintessentially Hebrew and consistently transliterate it into Square Hebrew. In such contexts, Palaeo-Hebrew texts are often considered to simply *be* Hebrew, and because the relationship between the Palaeo-Hebrew letters and Square Hebrew letters is one-to-one and quite regular, the transliteration is conceived of as simply a font change. Other scholars of Phoenician transliterate texts into Latin. The encoding of the Phoenician script in the Unicode Standard does not invalidate such scholarly practice; it is simply intended to make it possible to represent Phoenician, Punic, and similar textual materials directly in the historic script, rather than as specialized font displays of transliterations in modern Square Hebrew.

**Directionality.** Phoenician is written horizontally from right to left. The characters of the Phoenician script are all given strong right-to-left directionality.

**Punctuation.** Inscriptions and other texts in the various forms of the Phoenician script generally have no space between words. Dots are sometimes found between words in later exemplars—for example, in Moabite inscriptions—and U+1091F PHOENICIAN WORD SEPARATOR should be used to represent this punctuation.

**Stylistic Variation.** The letters for Phoenician proper and especially for Punic have very exaggerated descenders. These descenders help distinguish the main line of Phoenician

script evolution toward Punic, as contrasted with the Hebrew forms, where the descenders instead grew shorter over time.

**Numerals.** Phoenician numerals are built up from four elements in combination. Like the letters, Phoenician numbers are written from right to left:  $\text{|||}\text{𐤌}\text{𐤎}$  means 143 (100 + 20 + 20 + 1 + 1 + 1). This practice differs from modern Semitic scripts like Hebrew and Arabic, which use decimal numbers written from left to right.

**Names.** The names used for the characters here are those reconstructed by Theodor Nöldeke in 1904, as given in Powell (1996).

## 14.8 Ugaritic

### *Ugaritic: U+10380–U+1039F*

The city state of Ugarit was an important seaport on the Phoenician coast (directly east of Cyprus, north of the modern town of Minet el-Beida) from about 1400 BCE until it was completely destroyed in the twelfth century BCE. The site of Ugarit, now called Ras Shamra (south of Latakia on the Syrian coast), was apparently continuously occupied from Neolithic times (circa 5000 BCE). It was first uncovered by a local inhabitant while plowing a field in 1928 and subsequently excavated by Claude Schaeffer and Georges Chenet beginning in 1929, in which year the first of many tablets written in the Ugaritic script were discovered. They later proved to contain extensive portions of an important Canaanite mythological and religious literature that had long been sought and that revolutionized Biblical studies. The script was first deciphered in a remarkably short time jointly by Hans Bauer, Edouard Dhorme, and Charles Virolleaud.

The Ugaritic language is Semitic, variously regarded by scholars as being a distinct language related to Akkadian and Canaanite, or a Canaanite dialect. Ugaritic is generally written from left to right horizontally, sometimes using U+1039F  $\text{𐤀}$  UGARITIC WORD DIVIDER. In the city of Ugarit, this script was also used to write the Hurrian language. The letters U+1039B  $\text{𐤁}$  UGARITIC LETTER I, U+1039C  $\text{𐤂}$  UGARITIC LETTER U, and U+1039D  $\text{𐤃}$  UGARITIC LETTER SSU are used for Hurrian.

**Variant Glyphs.** There is substantial variation in glyph representation for Ugaritic. Glyphs for U+10398  $\text{𐤄}$  UGARITIC LETTER THANNA, U+10399  $\text{𐤅}$  UGARITIC LETTER GHAIN, and U+1038F  $\text{𐤆}$  UGARITIC LETTER DHAL differ somewhat between modern reference sources, as do some transliterations. U+10398  $\text{𐤄}$  UGARITIC LETTER THANNA is most often displayed with a glyph that looks like an occurrence of U+10393  $\text{𐤃}$  UGARITIC LETTER AIN overlaid with U+10382  $\text{𐤂}$  UGARITIC LETTER GAMLA.

**Ordering.** The ancient Ugaritic alphabetical order, which differs somewhat from the modern Hebrew order for similar characters, has been used to encode Ugaritic in the Unicode Standard.

**Character Names.** Some of the Ugaritic character names have been reconstructed; others appear in an early fragmentary document.

---

## 14.9 Old Persian

### *Old Persian: U+103A0–U+103DF*

The Old Persian script is found in a number of inscriptions in the Old Persian language dating from the Achaemenid Empire. Scholars today agree that the character inventory of Old Persian was invented for use in monumental inscriptions of the Achaemenid king, Darius I, by about 525 BCE. Old Persian is an alphabetic writing system with some syllabic aspects. While the shapes of some Old Persian letters look similar to signs in Sumero-Akkadian Cuneiform, it is clear that only one of them, U+103BE 𐎠 OLD PERSIAN SIGN LA, was actually borrowed. It was derived from the New Assyrian historic variant 𐎠 of Sumero-Akkadian U+121B7 𐎠 CUNEIFORM SIGN LA, because *la* is a foreign sound not used in the Old Persian language.

**Directionality.** Old Persian is written from left to right.

**Repertoire.** The repertoire contains 36 signs. These represent consonants, vowels, or consonant plus vowel syllables. There are also five numbers, one word divider, and eight ideograms. It is considered unlikely that any additional characters will be discovered.

**Numerals.** The attested numbers are built up by stringing the base numbers (1, 2, 10, 20, and 100) in sequences.

**Variants.** The signs U+103C8 OLD PERSIAN SIGN AURAMAZDAA and U+103C9 OLD PERSIAN SIGN AURAMAZDAA-2, and the signs U+103CC OLD PERSIAN SIGN DAHYAAUSH and U+103CD OLD PERSIAN SIGN DAHYAAUSH-2, have been encoded separately because their conventional attestation in the corpus of Old Persian texts is quite limited and scholars consider it advantageous to distinguish the forms in plain text representation.

---

## 14.10 Sumero-Akkadian

### *Cuneiform: U+12000–U+123FF*

Sumero-Akkadian Cuneiform is a logographic writing system with a strong syllabic component. It was written from left to right on clay tablets.

**Early History of Cuneiform.** The earliest stage of Mesopotamian Cuneiform as a complete system of writing is first attested in Uruk during the so-called Uruk IV period (circa 3500–3200 BCE) with an initial repertoire of about 700 characters or “signs” as Cuneiform scholars customarily call them.

Late fourth millennium ideographic tablets were also found at Susa and several other sites in western Iran, in Assyria at Nineveh (northern Iraq), at Tell Brak (northwestern Syria), and at Habuba Kabira in Syria. The writing system developed in Sumer (southeastern Iraq) was repeatedly exported to peripheral regions in the third, second, and first millennia BCE. Local variations in usage are attested, but the core of the system is the Sumero-Akkadian writing system.

Writing emerged in Sumer simultaneously with a sudden growth in urbanization and an attendant increase in the scope and scale of administrative needs. A large proportion of the elements of the early writing system repertoire was devised to represent quantities and commodities for bureaucratic purposes.

At this earliest stage, signs were mainly pictographic, in that a relatively faithful facsimile of the thing signified was traced, although some items were strictly ideographic and represented by completely arbitrary abstractions, such as the symbol for sheep  $\oplus$ . Some scholars believe that the abstract symbols were derived from an earlier “token” system of accounting, but there is no general agreement on this point. Where the pictographs are concerned, interpretation was relatively straightforward. The head of a bull was used to denote “cattle”; an ear of barley was used to denote “barley.” In some cases, pictographs were also interpreted logographically, so that meaning was derived from the symbol by close conceptual association. For example, the representation of a bowl might mean “bowl,” but it could indicate concepts associated with bowls, such as “food.” Renditions of a leg might variously suggest “leg,” “stand,” or “walk.”

By the next chronological period of south Mesopotamian history (the Uruk III period, 3200–2900 BCE), logographic usage seems to have become much more widespread. In addition, individual signs were combined into more complex designs to express other concepts. For example, a head with a bowl next to it was used to denote “eat” or “drink.” This is the point during script development at which one can truly speak of the first Sumerian texts. In due course, the early graphs underwent change, conditioned by factors such as the most widely available writing medium and writing tools, and the need to record information more quickly and efficiently from the standpoint of the bureaucracy that spawned the system.

Clay was the obvious writing medium in Sumer because it was widely available and easily molded into cushion- or pillow-shaped tablets. Writing utensils were easily made for it by sharpening pieces of reed. Because it was awkward and slow to inscribe curvilinear lines in a piece of clay with a sharpened reed (called a *stylus*), scribes tended to approximate the pictographs by means of short, wedge-shaped impressions made with the edge of the stylus. These short, mainly straight shapes gave rise to the modern word “cuneiform” from the Latin *cuneus*, meaning “wedge.” Cuneiform proper was common from about 2700 BCE, although experts use the term “cuneiform” to include the earlier forms as well.

**Geographic Range.** The Sumerians did not live in complete isolation, and there is very early evidence of another significant linguistic group in the area immediately north of Sumer known as Agade or Akkad. Those peoples spoke a Semitic language whose dialects are subsumed by scholars under the heading “Akkadian.” In the long run, the Akkadian



speakers became the primary users and promulgators of Cuneiform script. Because of their trade involvement with their neighbors, Cuneiform spread through Babylonia (the umbrella term for Sumer and Akkad) to Elam, Assyria, eastern Syria, southern Anatolia, and even Egypt. Ultimately, many languages came to be written in Cuneiform script, the most notable being Sumerian, Akkadian (including Babylonian, Assyrian, Eblaite), Elamite, Hittite, and Hurrian.

Periods of script usage are defined according to geography and primary linguistic representation, as shown in *Table 14-2*.

**Table 14-2. Cuneiform Script Usage**

Archaic Period (to 2901 BCE)		Elamite (2100–360 BCE)
Early Dynastic (2900–2335 BCE)		
Old Akkadian (2334–2154 BCE)		
Ur III (NeoSumerian) (2112–2095 BCE)		
Old Assyrian (1900–1750 BCE)	Old Babylonian (2004–1595 BCE)	
Middle Assyrian (1500–1000 BCE)	Middle Babylonian (1595–627 BCE)	
Neo-Assyrian (1000–609 BCE)		
	Neo-Babylonian (626–539 BCE)	
Hittite (1570–1220 BCE)		

**Sources and Coverage.** The base character repertoire for the Cuneiform block was distilled from the list of Ur III signs compiled by the Cuneiform Digital Library Initiative (UCLA) in union with the list constructed independently by Miguel Civil. This repertoire is comprehensive from the Ur III period onward. Old Akkadian, Early Dynastic, and Archaic Cuneiform are not covered by this repertoire.



**Simple Signs.** Most Cuneiform signs are simple units; each sign of this type is represented by a single character in the standard.

**Complex and Compound Signs.** Some Cuneiform signs are categorized as either complex or compound signs. Complex signs are made up of a primary sign with one or more secondary signs written within it or conjoined to it, such that the whole is generally treated by scholars as a unit; this includes linear sequences of two or more signs or wedge-clusters where one or more of those clusters have not been clearly identified as characters in their own right. Complex signs, which present a relative visual unity, are assigned single individual code points irrespective of their components.

Compound signs are linear sequences of two or more signs or wedge-clusters generally treated by scholars as a single unit, when each and every such wedge-cluster exists as a clearly identified character in its own right. Compound signs are encoded as sequences of their component characters. Signs that shift from compound to complex, or vice versa, generally have been treated according to their Ur III manifestation.

**Mergers and Splits.** Over the long history of Cuneiform, an number of signs have simplified and merged; in other cases, a single sign has diverged and developed into more than one distinct sign. The choice of signs for encoding as characters was made at the point of maximum differentiation in the case of either mergers or splits to enable the most comprehensive set for the representation of text in any period.

Fonts for the representation of Cuneiform text may need to be designed distinctly for optimal use for different historic periods. Fonts for some periods will contain duplicate glyphs depending on the status of merged or split signs at that point of the development of the writing system.

**Glyph Variants Acquiring Independent Semantic Status.** Glyph variants such as U+122EC  CUNEIFORM SIGN TA ASTERISK, a Middle Assyrian form of the sign U+122EB  CUNEIFORM SIGN TA, which in Neo-Assyrian usage has its own logographic interpretation, have been assigned separate code positions. They are to be used only when the new interpretation applies.





**Formatting.** Cuneiform was often written between incised lines or in blocks surrounded by drawn boxes known as *case rules*. These boxes and lines are considered formatting and are not part of the script. Case ruling and the like are not to be treated as punctuation.

**Ordering.** The characters are encoded in the Unicode Standard in Latin alphabetical order by primary sign name. Complex signs based on the primary sign are organized according to graphic principles; in some cases, these correspond to the native analyses.

**Other Standards.** There is no standard legacy encoding of Cuneiform primarily because it was not possible to encode the huge number of characters in the pre-Unicode world of 8-bit fonts.

### **Cuneiform Numbers and Punctuation: U+12400–U+1247F**

**Cuneiform Punctuation.** A small number of signs are occasionally used in Cuneiform to indicate word division, repetition, or phrase separation.

**Cuneiform Numerals.** In general, numerals have been encoded separately from signs that are visually identical but semantically different (for example, U+1244F  CUNEIFORM NUMERIC SIGN ONE BAN2, U+12450  CUNEIFORM NUMERIC SIGN TWO BAN2, and so on, versus U+12226  CUNEIFORM SIGN MASH, U+1227A  CUNEIFORM SIGN PA, and so on).