

# The Unicode® Standard

## Version 15.0 – Core Specification

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

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.

Unicode and the Unicode Logo are registered trademarks of Unicode, Inc., in the United States and other countries.

The authors and publisher have taken care in the preparation of this specification, 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.

© 2022 Unicode, Inc.

All rights reserved. This publication is protected by copyright, and permission must be obtained from the publisher prior to any prohibited reproduction. For information regarding permissions, inquire at <https://www.unicode.org/reporting.html>. For information about the Unicode terms of use, please see <https://www.unicode.org/copyright.html>.

The Unicode Standard / the Unicode Consortium; edited by the Unicode Consortium. — Version 15.0.

Includes index.

ISBN 978-1-936213-32-0 (<https://www.unicode.org/versions/Unicode15.0.0/>)

1. Unicode (Computer character set) I. Unicode Consortium.

QA268.U545 2022

ISBN 978-1-936213-32-0

Published in Mountain View, CA

September 2022

## Chapter 8

# *Europe-II*

## *Ancient and Other Scripts*

This chapter describes ancient scripts of Europe, as well as other historic and limited-use scripts of Europe not covered in *Chapter 7, Europe-I*. This includes the various ancient Mediterranean scripts, other early alphabets and sets of runes, some poorly attested historic scripts of paleographic interest, and more recently devised constructed scripts with significant usage. In particular, these include:

<i>Linear A</i>	<i>Old Italic</i>	<i>Caucasian Albanian</i>
<i>Linear B</i>	<i>Runic</i>	<i>Vithkuqi</i>
<i>Cypriot Syllabary</i>	<i>Old Hungarian</i>	<i>Old Permic</i>
<i>Cypro-Minoan</i>	<i>Gothic</i>	<i>Ogham</i>
<i>Anatolian Alphabets</i>	<i>Elbasan</i>	<i>Shavian</i>

Unicode encodes a number of ancient scripts, which have not been in normal use for a millennium or more, as well as historic scripts, whose usage ended in recent centuries. Although they are no longer used to write living languages, documents and inscriptions using these scripts exist, both for extinct languages and for 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 historic scripts are related to each other as well as to modern alphabets.

The Linear A script is an ancient writing system used from approximately 1700–1450 BCE on and around the island of Crete. The script contains more than ninety signs in regular use and a host of logograms. Surviving examples are inscribed on clay tablets, stone tables, and metals. The language of the inscriptions has not yet been deciphered.

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.

Cypro-Minoan is an undeciphered script from the late Bronze Age (circa 1550-1050 BCE) found on objects from the island of Cyprus, the ancient cities of Ugarit (modern-day Ras Shamra, Syria) and Tiryns, Greece. It is a syllabic script.

The ancient Anatolian alphabets Lycian, Carian, and Lydian all date from the first millennium BCE, and were used to write various ancient Indo-European languages of western and southwestern Anatolia. All are closely related to the Greek script.

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. One of the Old Italic alphabets of northern Italy may have influenced the development of the Runic script, which has a distinct angular appearance owing to its use in carving inscriptions in stone and wood.

Old Hungarian is another historical runiform script, used to write the Hungarian language in Central Europe. In recent decades it has undergone a significant revival in Hungary. It has developed casing, and is now used with modern typography to print significant amounts of material in the modern Hungarian language. It is laid out right-to-left.

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.

The Gothic script, like Cyrillic, was developed on the basis of Greek at a much later date than Old Italic.

Elbasan, Caucasian Albanian, and Old Permic are all simple alphabetic scripts. Elbasan is an historic alphabetic script invented in the middle of the eighteenth-century to write Albanian. It is named after the city where it originated. Caucasian Albanian dates from the early fifth century and is related to the modern Udi language. Old Permic was devised in the fourteenth century to write the Uralic languages Komi and Komi-Permyak. Its use for Komi extended into the seventeenth century.

Vithkuqi, a historical script for Albanian, was invented by Naum P. Veqilharxhi, and named for the town where it was created in the nineteenth century. A left-to-right alphabetic script, it is experiencing some modern revivalist efforts in artistic and cultural uses.

Shavian is a phonemic alphabet invented in the 1950s to write English. It was used to publish one book in 1962, but remains of some current interest.

## 8.1 Linear A

### *Linear A: U+10600–U+1077F*

The Linear A script was used from approximately 1700–1450 BCE. It was mainly used on the island of Crete and surrounding areas to write a language which has not yet been identified. Unlike the later Linear B, which was used to write an early form of Greek, Linear A appears on a variety of media, such as clay tablets, stone offering tables, gold and silver hair pins, and pots.

**Encoding.** The repertoire of characters in the Unicode encoding of the Linear A script is broadly based on the GORILA catalog by Godart and Olivier (1976–1985), which is the basic set of signs used in decipherment efforts. All simple signs in that catalog are encoded as single characters. Composite signs consisting of vertically stacked parts or touching pieces are also encoded as single characters. Composite signs in the catalog which consist of side-by-side pieces that are not touching are treated as digraphs; the parts are individually encoded as characters, but the composite sign is not separately encoded.

**Structure.** Linear A contains more than ninety syllabic signs in regular use and a host of logograms. Some Linear A signs are also found in Linear B, although about 80% of the logograms in Linear A do not appear in Linear B.

**Character Names.** The Linear A character names are based on the GORILA catalog numbers.

**Directionality.** Linear A was written from left to right, though occasionally it appears right to left and, rarely, boustrophedon.

**Numbers.** Numbers in Linear A inscriptions are represented by characters in the Aegean Numbers block. Numbers are usually arranged in sets of five or fewer that are stacked vertically. The largest number recorded is 3,000. Linear A seems to use a series of unit fractions. Seven fractions are regularly used and are included in the Linear A block.

## 8.2 Linear B

### *Linear B Syllabary: U+1000–U+100F*

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+100FF*

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 ἥ 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.

## 8.3 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 8-1*. 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 8-1.** Similar Characters in Linear B and Cypriot

Linear B	Cypriot	Linear B	Cypriot
da	𐀀	ta	𐀁
na	𐀂	na	𐀃
pa	𐀄	pa	𐀅
ro	𐀆	lo	𐀇
se	𐀈	se	𐀉
ti	𐀊	ti	𐀋
to	𐀌	to	𐀍

For Aegean numbers, see the subsection “Aegean Numbers: U+10100–U+1013F” in *Section 8.2, Linear B*.

## 8.4 Cypro-Minoan

### *Cypro-Minoan: U+12F90–U+12FFF*

Cypro-Minoan is an undeciphered script found on approximately 250 objects from the island of Cyprus, the ancient cities of Ugarit (modern-day Ras Shamra, Syria) and Tiryns, Greece. The script dates to the late Bronze Age (circa 1550-1050 BCE). The name “Cypro-Minoan” was coined by Arthur Evans in 1909 because he believed Cypro-Minoan derived from the scripts of Minoan Crete.

Researchers have tentatively classified Cypro-Minoan into four categories, termed CM0, CM1, CM2, and CM3, based on temporal and geographical criteria. The repertoire in the Unicode Standard covers characters from the CM1, CM2, and CM3 groups, but does not cover CM0; it is largely based on Olivier 2007.

**Structure.** Cypro-Minoan is a syllabic script and has been encoded with left-to-right directionality.

**Names.** The character names are based on Olivier 2007.

**Glyphs.** The glyphs in the code charts generally follow the CM1 forms, but if no CM1 form exists, a CM2 or CM3 form is used. The glyphs follow Olivier 2007 generally, except for U+12F9C CYPRO-MINOAN SIGN CM013, which has been modified based on recent research. The code chart normalizes the glyphs into a more linear style.

**Punctuation.** A few Cypro-Minoan punctuation marks have been identified. Two script-specific signs are encoded: U+12FF1 CYPRO-MINOAN SIGN CM301 and U+12FF2 CYPRO-MINOAN SIGN CM302. Two other marks have been unified with two punctuation characters in the Aegean Numbers block: U+10100 AEGEAN WORD SEPARATOR LINE and U+10101 AEGEAN WORD SEPARATOR DOT.

**Numbers.** Numbers in Cypro-Minoan are known, but poorly attested. Users may choose to employ characters from the Aegean numbers block for Cypro-Minoan, but the exact relationship between the Cypro-Minoan and Aegean numbers remains uncertain.



## 8.5 Ancient Anatolian Alphabets

**Lycian:** U+10280–U+1029F

**Carian:** U+102A0–U+102DF

**Lydian:** U+10920–U+1093F

The Anatolian scripts described in this section all date from the first millennium BCE, and were used to write various ancient Indo-European languages of western and southwestern Anatolia (now Turkey). All are closely related to the Greek script and are probably adaptations of it. Additional letters for some sounds not found in Greek were probably either invented or drawn from other sources. However, development parallel to, but independent of, the Greek script cannot be ruled out, particularly in the case of Carian.

**Lycian.** Lycian was used from around 500 BCE to about 200 BCE. The term “Lycian” is now used in place of “Lycian A” (a dialect of Lycian, attested in two texts in Anatolia, is called “Lycian B”, or “Milyan”, and dates to the first millennium BCE). The Lycian script appears on some 150 stone inscriptions, more than 200 coins, and a few other objects.

Lycian is a simple alphabetic script of 29 letters, written left-to-right, with frequent use of word dividers. The recommended word divider is U+205A TWO DOT PUNCTUATION. *Scriptio continua* (a writing style without spaces or punctuation) also occurs. In modern editions U+0020 SPACE is sometimes used to separate words.

**Carian.** The Carian script is used to write the Carian language, and dates from the first millennium BCE. While a few texts have been found in Caria, most of the written evidence comes from Carian communities in Egypt, where they served as mercenaries. The repertoire of the Carian texts is well established. Unlike Lycian and Lydian, Carian does not use a single standardized script, but rather shows regional variation in the repertoire of signs used and their form. Although some of the values of the Carian letters remain unknown or in dispute, their distinction from other letters is not. The Unicode encoding is based on the standard “Masson set” catalog of 45 characters, plus 4 recently-identified additions. Some of the characters are considered to be variants of others—and this is reflected in their names—but are separately encoded for scholarly use in discussions of decipherment.

The primary direction of writing is left-to-right in texts from Caria, but right-to-left in Egyptian Carian texts. However, both directions occur in the latter, and left-to-right is favored for modern scholarly usage. Carian is encoded in Unicode with left-to-right directionality. Word dividers are not regularly employed; *scriptio continua* is common. Word dividers which are attested are U+00B7 MIDDLE DOT (or U+2E31 WORD SEPARATOR MIDDLE DOT), U+205A TWO DOT PUNCTUATION, and U+205D TRICOLON. In modern editions U+0020 SPACE may be found.

**Lydian.** While Lydian is attested from inscriptions and coins dating from the end of the eighth century (or beginning of the seventh) until the third century BCE, the longer well-preserved inscriptions date to the fifth and fourth centuries BCE.

Lydian is a simple alphabetic script of 26 letters. The vast majority of Lydian texts have right-to-left directionality (the default direction); a very few texts are left-to-right and one is boustrophedon. Most Lydian texts use U+0020 SPACE as a word divider. Rare examples have been found which use *scriptio continua* or which use dots to separate the words. In the latter case, U+003A COLON and U+00B7 MIDDLE DOT (or U+2E31 WORD SEPARATOR MIDDLE DOT) can be used to represent the dots. U+1093F LYDIAN TRIANGULAR MARK is thought to indicate quotations, and is mirrored according to text directionality.

## 8.6 Old Italic

### *Old Italic: U+10300–U+1032F*

The Old Italic script is used to represent a number of related historical alphabets located on the Italian peninsula. Some of these were used for non-Indo-European languages (Etruscan, Raetic, and probably North Picene), and some for various Indo-European languages belonging to the Italic branch (Faliscan and members of the Sabellian group, including Oscan, Umbrian, and South Picene) the Celtic branch (Cisalpine Celtic), and the Venetic branch. The ultimate source for the alphabets in ancient Italy is Euboean Greek used at Ischia and Cumae in the bay of Naples in the eighth century BCE. Unfortunately, no Greek abecedaries from southern Italy have survived. The native alphabets of Faliscan, Oscan, Umbrian, North Picene, South Picene, Venetic, and Cisalpine Celtic all derive from an Etruscan form of the alphabet. Raetic, or another Old Italic alphabet of northern Italy, may have influenced the historical development of Runic. (See *Section 8.7, Runic.*)

There are some 10,000 inscriptions in Etruscan. By the time of the earliest Etruscan inscriptions, circa 700 BCE, local distinctions are already found in the use of the alphabet. Three major stylistic divisions are identified: the Northern, Southern, and Caere/Veii. Use of Etruscan can be divided into two stages, owing largely to the phonological changes that occurred: the “archaic Etruscan alphabet,” used from the seventh to the fifth centuries BCE, and the “neo-Etruscan alphabet,” used from the fourth to the first centuries BCE. Glyphs for eight of the letters differ between the two periods; additionally, neo-Etruscan abandoned the letters KA, KU, and EKS.

The unification of these alphabets into a single Old Italic script requires language-specific fonts because the glyphs most commonly used may differ somewhat depending on the language being represented.

Most of the languages have added characters to the common repertoire: Etruscan and Faliscan add LETTER EF; Oscan adds LETTER EF, LETTER II, and LETTER UU; Umbrian adds LETTER EF, LETTER ERS, and LETTER CHE; North Picene adds LETTER UU; South Picene adds LETTER II, LETTER UU, and LETTER ESS; Venetic adds LETTER YE; and Raetic adds NORTH-ERN TSE and SOUTHERN TSE.

The Latin script itself derives from a south Etruscan model, probably from Caere or Veii, around the mid-seventh century BCE or a bit earlier. However, because there are significant differences between Latin and Faliscan of the seventh and sixth centuries BCE in terms of formal differences (glyph shapes, directionality) and differences in the repertoire of letters used, this warrants a distinctive character block. Fonts for early Latin should use the *uppercase* code positions U+0041..U+005A.

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:], while [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 Etruscan texts from the seventh to six centuries BCE were written from right-to-left, but left-to-right was not uncommon, and is found in approximately ten percent of the texts from this period. From the fifth to the first centuries BCE, right-to-left was the standard, and left-to-right directionality was extremely rare. The other local varieties of Old Italic 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.

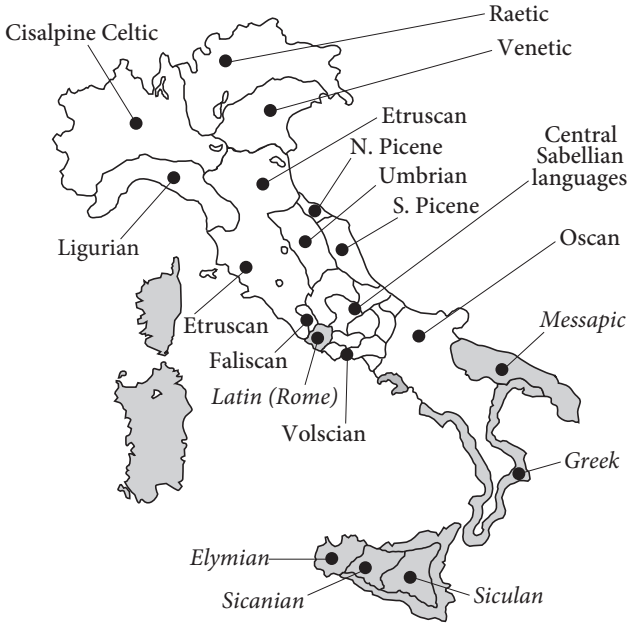
**Punctuation.** The earliest inscriptions are written with no space between words in what is called *scriptio continua*. There are numerous Etruscan inscriptions with dots separating word forms, attested as early as the second quarter of the seventh century BCE. This punctuation is sometimes, but only rarely, used to mark certain types of syllables and not to separate 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 abecedarium (mid-seventh century BCE). Note that the phonetic values for U+10317 OLD ITALIC LETTER EKS [ks] and U+10319 OLD ITALIC LETTER KHE [k<sup>h</sup>] show the influence of western, Euboean Greek; eastern Greek has U+03A7 GREEK CAPITAL LETTER CHI [k<sup>h</sup>] and U+03A8 GREEK CAPITAL LETTER PSI [ps] instead.

The geographic distribution of the Old Italic script is shown in *Figure 8-1*. 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 italics. In particular, note that the ancient Greek colonies of the southern Italian and Sicilian coasts used the Greek script proper. Rome, of course, is shown in gray, because Latin was written with the Latin alphabet, now encoded in the Latin script.

**Figure 8-1.** Distribution of Old Italic



## 8.7 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.

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.

**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 *futhork* from the name of its first six letters. The original *old futhork* contained 24 runes:

ƿ ᚋ ᚏ ᚦ ᚱ < X ƿ ᚠ ᚢ ᚣ ᚤ ᚥ ᚦ ᚧ ᚨ ᚩ ᚪ ᚫ ᚬ ᚭ ᚮ ᚯ ᚰ ᚱ

They are usually transliterated in this way:

f u þ a r k g w h n i j ð 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 *futhork* changed in a different way; in the eighth century, the simplified younger *futhork* appeared. It consists of only 16 runes, some of which are used in two different forms. The long-branch form is shown here:

ƿ ᚋ ᚏ ᚦ ᚱ ᚲ \* 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 *futhork* was influenced by the Latin alphabet and new runes were invented so that there was full correspondence with the Latin letters.

**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.

**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-Frisian 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 † RUNIC LETTER GER is actually a rare form, as the more common form is already used for U+16E1 ‡ 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.** 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.



## 8.8 Old Hungarian

### *Old Hungarian: U+10C80–U+10CFF*

The Old Hungarian script is a runiform script that is used to write the Hungarian language. Old Hungarian is mentioned in a written account of the late 13th century and has been found on short stone-carved inscriptions. The script was probably developed and in use earlier. Modern use has increased dramatically in the last two decades, with some uses being simply decorative. There are also currently publications of books, magazines, and teaching materials.

**Structure.** Old Hungarian is an alphabetic script. The consonants traditionally bore an inherent vowel. Vowel signs were only explicitly written in final position, where vowels were long, and for disambiguation. In later phases of script usage, all vowels were written explicitly. The script is rendered linearly, but traditionally used a large set of ligatures.

Casing is not part of the traditional Old Hungarian script. However, modern practice has introduced casing into many publications. Uppercase letters appear as larger size variants of lowercase letters.

**Directionality.** The primary direction of writing is right-to-left both in historical sources and in modern use. Conformant implementations of Old Hungarian script must use the Unicode Bidirectional Algorithm (see Unicode Standard Annex #9, “Unicode Bidirectional Algorithm”).

**Punctuation and Numbers.** Traditional texts separate words with spaces or with one, two, or four dots. Modern users punctuate Old Hungarian with U+0020 SPACE, U+2E41 REVERSED COMMA, and U+2E42 DOUBLE LOW-REVERSED-9 QUOTATION MARK, with some use of U+2E31 WORD SEPARATOR MIDDLE DOT, U+205A TWO DOT PUNCTUATION, U+205D TRICOLON, and U+205E VERTICAL FOUR DOTS as well.

Old Hungarian numbers have their origin in a tally system which was widely used throughout Hungary until the nineteenth century. Since the twentieth century, these numbers have been used regularly with Old Hungarian. The numbers are built up from elements in a multiplicative-additive system. Old Hungarian numbers are encoded at U+10CFA..U+10CFF.

## 8.9 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:

**SYƆ ƆAMƆLIƆ IST İN ESÄİIN ƆRANƆETAN**  
*swe gameliþ ist 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 both sides 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:

•Ḅ or Ḅ̅̅̅ 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.

## 8.10 Elbasan

### *Elbasan: U+10500–U+1052F*

The earliest alphabet devised for the Albanian language was created around 1750 for the Elbasan Gospel manuscript, which is the only known example of the script. This manuscript, preserved at the State Archives in Tirana, records the earliest-known Albanian-language text in an original alphabet. Most of the letters in the Elbasan alphabet seem to be new creations, although some of their shapes may have been influenced by Greek and Cyrillic.

**Structure.** Elbasan is a simple alphabetic script written from left to right horizontally. The alphabet consists of forty letters.

Three characters have an inherent diacritical dot: U+10505 ǀ ELBASAN LETTER NDE is used to indicate a pre-nasalized U+10504 ǀ /d/; U+10511 ǎ ELBASAN LETTER LLE is used to indicate a geminate U+10510 ǎ /l/; U+1051A ǁ ELBASAN LETTER RRE is used to indicate a geminate U+10519 ǁ /r/. In many cases the dot on *nde* is written like a small *ne*. In one instance in the manuscript *gje* is written with a dot above to indicate prenasalized /j/.

Two different letters are used for /n/: U+10513 N ELBASAN LETTER NE is used generally, and U+10514 ǂ ELBASAN LETTER NA is typically used in prenasalized position as in ǂǂ /ng/ and ǂǂ /nj/. Two letters, which are rare and appear in Greek loanwords, are used for /γ/, U+10525 Ɔ ELBASAN LETTER GHE and U+10526 L ELBASAN LETTER GHAMMA.

**Accents and Other Marks.** The Elbasan manuscript contains breathing accents, similar to those used in Greek. Those accents do not appear regularly in the orthography and have not been fully analyzed yet. Raised vertical marks also appear in the manuscript, but are not specific to the script. Generic combining characters from the Combining Diacritical Marks block can be used to render these accents and other marks.

**Names.** The names used for the characters in the Elbasan block are based on those of the modern Albanian alphabet.

**Numerals and Punctuation.** There are no script-specific numerals or punctuation marks. A separating dot and spaces appear in the Elbasan manuscript, and may be rendered with U+00B7 MIDDLE DOT and U+0020 SPACE, respectively. For numerals, a Greek-like system of letter and combining overline is in use. Overlines also appear above certain letters in abbreviations, such as  $\overline{\text{Zot}}$  to indicate *Zot* (Lord). The overlines in numerals and abbreviations can be represented with U+0305 COMBINING OVERLINE.

## 8.11 Caucasian Albanian

### *Caucasian Albanian: U+10530–U+1056F*

The Caucasian Albanian script was identified as a unique script in 1937 on the basis of an alphabet list in an Armenian manuscript in the Matenadaran collection in Yerevan and confirmed by a few inscriptions on artifacts excavated in northwest Azerbaijan around 1950. In the 1990s two palimpsest manuscripts containing the Caucasian Albanian script were discovered in St. Catherine's Monastery on Mount Sinai. These undated manuscripts appear to have been written during the seventh century CE. The palimpsests were deciphered and the Caucasian Albanian language and script was determined to be closely related to, if not an ancestor of, the present-day Udi language.

**Structure.** Caucasian Albanian is a simple alphabetic script written from left to right horizontally. Spaces are not used to separate words in the manuscript, though modern editions use spaces for the better legibility.

**Abbreviations.** An abbreviatory convention occurs, using a line above spanning two letters. This line above has a titlo-appearance, with small fillets at the ends of the strokes. This convention is similar to that seen in Coptic. For Caucasian Albanian, use of U+035E COMBINING DOUBLE MACRON is recommended to represent such abbreviations, with the font design dealing with the swash ends of the line, for styles that require it.

**Numerals.** Script-specific numerals are not known. Letters used as numbers are marked with a line above and/or below the letter, so  $\overline{\text{Բ}}$  or  $\underline{\text{Բ}}$  or  $\overline{\underline{\text{Բ}}} = 2$ . When more than two or three letters are associated with a numeric mark, a continuous line is drawn above or below all of them. These lines above and/or below can be represented with U+0304 COMBINING MACRON, U+0331 COMBINING MACRON BELOW, or with various combinations of combining half macrons and conjoining macrons from the Combining Half Marks block (U+FE20..U+FE2F), as needed. (See the discussion of supralineation in *Section 7.3, Coptic.*)

**Punctuation.** One special mark, U+1056F CAUCASIAN ALBANIAN CITATION MARK, is used to indicate text that is a citation from the psalms.

## 8.12 Vithkuqi

### *Vithkuqi: U+10570–U+105BF*

Vithkuqi, a historical script for Albanian, was invented by Naum P. Veqilharxhi, and named for the town where it was created. The script also has been known by different spellings of the town's name: BÛthakukye or Beitha Kukju. Use of this alphabetic script arose between 1824 and 1845. The earliest use of Vithkuqi was in a nineteenth-century spelling book that was the basis for spelling books used in the regions of Bulgaria and Albania. A copy of the early Vithkuqi spelling book can be found in the Gennadius Library in Athens.

There are revivalist efforts in artistic and cultural uses of Vithkuqi, notably in the script's use in modern tattoos. Vithkuqi glyphs visually resemble cursive Armenian.

**Structure.** Vithkuqi is a left-to-right alphabetic script. There is no ligation, and non-productive diacritics are encoded atomically.

**Casing.** Casing is used in the Vithkuqi script.

**Numerals and Punctuation.** Vithkuqi uses European numbers and standard Latin punctuation.

## 8.13 Old Permic

### *Old Permic: U+10350–U+1037F*

The Old Permic script was devised in the 14th century by the Russian missionary Stefan of Perm, and was used to write the Uralic languages Komi and Komi-Permyak. It was modeled on the Greek and Cyrillic alphabets, but many glyphs were taken from the “Tamga signs” used in indigenous Komi religious practices. Stefan translated Russian and Greek liturgical and biblical texts into Komi. There are a few surviving medieval documents in the script, chiefly in the form of icons, glosses, and inscriptions on monuments.

Old Permic continued to be used for Komi until the 17th century. In addition, the script was used cryptographically from the 15th century to write Russian, because it was unknown to most readers of Russian.

**Structure.** Old Permic is a simple, caseless, alphabetic script, read from left to right in horizontal lines running from top to bottom.

**Combining Letters.** A small number of letters, encoded from U+10376 to U+1037A, appear as superscript letters in abbreviations, in the same way that letters are used in Latin and Cyrillic.

**Combining Marks.** Old Permic employs a number of combining marks, as shown in Table 8-2. U+0483 COMBINING CYRILLIC TITLO indicates an abbreviation, typically with a specific set of holy words. The *combining grave accent* was sometimes used to mark consonant palatalization, and the *combining diaeresis* at times distinguishes [i] and [j]. However, the use of these combining marks is not always clear, and may have no phonetic value at all.

**Table 8-2.** Combining Marks Used in Old Permic

U+0300	COMBINING GRAVE ACCENT
U+0306	COMBINING BREVE
U+0307	COMBINING DOT ABOVE
U+0308	COMBINING DIAERESIS
U+0313	COMBINING COMMA ABOVE
U+0483	COMBINING CYRILLIC TITLO
U+20DB	COMBINING THREE DOTS ABOVE

**Numerals.** Script-specific numerals are not known. Letters of the alphabet can be marked with U+0483 COMBINING CYRILLIC TITLO to indicate numeric use.

**Punctuation.** Old Permic does not have any script-specific punctuation, but uses middle dot, colon, and apostrophe. Spaces are used to separate words in manuscripts.

## 8.14 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 be given its conventional width and simply left blank like U+0020 SPACE. U+169B OGHAM FEATHER MARK and U+169C OGHAM REVERSED FEATHER MARK are used at the beginning and the end of Ogham text, particularly in manuscript Ogham. In some cases, only the *Ogham feather mark* is used, which can indicate the direction of the text.

The word *latheirt* >π<      < shows the use of the feather marks. This word was written in the margin of a ninth-century Latin grammar and means “massive hangover,” which may be the scribe’s apology for any errors in his text.

## 8.15 Shavian

### **Shavian: U+10450–U+1047F**

The playwright George Bernard Shaw (1856–1950) was an outspoken critic of the idiosyncrasies of English orthography. In his will, he directed that Britain’s Public Trustee seek out and publish an alphabet of no fewer than 40 letters to provide for the phonetic spelling of English. The alphabet finally selected was designed by Kingsley Read and is variously known as Shavian, Shaw’s alphabet, and the Proposed British Alphabet. Also in accordance with Shaw’s will, an edition of his play, *Androcles and the Lion*, was published and distributed to libraries, containing the text both in the standard Latin alphabet and in Shavian.

As with other attempts at spelling reform in English, the alphabet has met with little success. Nonetheless, it has its advocates and users. The normative version of Shavian is taken to be the version in *Androcles and the Lion*.

**Structure.** The alphabet consists of 48 letters and 1 punctuation mark. The letters have no case. The digits and other punctuation marks are the same as for the Latin script. The one additional punctuation mark is a “name mark,” used to indicate proper nouns. U+00B7 MIDDLE DOT should be used to represent the “name mark.” The letter names are intended to be indicative of their sounds; thus the sound /p/ is represented by U+10450 ᶏ SHAVIAN LETTER PEEP.

The first 40 letters are divided into four groups of 10. The first 10 and second 10 are 180-degree rotations of one another; the letters of the third and fourth groups often show a similar relationship of shape.

The first 10 letters are tall letters, which ascend above the x-height and generally represent unvoiced consonants. The next 10 letters are “deep” letters, which descend below the baseline and generally represent voiced consonants. The next 20 are the vowels and liquids. Again, each of these letters usually has a close phonetic relationship to the letter in its matching set of 10.

The remaining 8 letters are technically ligatures, the first 6 involving vowels plus /r/. Because ligation is not optional, these 8 letters are included in the encoding.

**Collation.** The problem of collation is not addressed by the alphabet’s designers.



