Technical Reports

L2/08-060

Proposed Update to

Unicode Standard Annex #24

UNICODE SCRIPT ~~NAMES~~ PROPERTY

Version	Unicode 5.1.0 draft2
Authors	Mark Davis (mark.davis@google.com), Ken Whistler (ken@unicode.org)
Date	2007-11-06
This Version	http://www.unicode.org/reports/tr24/tr24-10.html
Previous Version	http://www.unicode.org/reports/tr24/tr24-9.html
Latest Version	http://www.unicode.org/reports/tr24/tr24
Revision	10

Summary

This annex specifies an assignment of script names to all Unicode code points. This information is useful in mechanisms such as regular expressions and other text processing tasks.

Status

*This is a **draft** document which may be updated, replaced, or superseded by other documents at any time. Publication does not imply endorsement by the Unicode Consortium. This is not a stable document; it is inappropriate to cite this document as other than a work in progress.*

> ***A Unicode Standard Annex (UAX)** forms an integral part of the Unicode Standard, but is published online as a separate document. The Unicode Standard may require conformance to normative content in a Unicode Standard Annex, if so specified in the Conformance chapter of that version of the Unicode Standard. The version number of a UAX document corresponds to the version of the Unicode Standard of which it forms a part.*

Please submit corrigenda and other comments with the online reporting form [Feedback]. Related information that is useful in understanding this annex is found in Unicode Standard Annex #41, "Common References for Unicode Standard Annexes." For the latest version of the Unicode Standard, see [Unicode]. For a list of current Unicode Technical Reports, see [Reports]. For more information about versions of the Unicode Standard, see [Versions].

Contents

1 Introduction
1.1 Classification of Text by Script Name

1.2 Scripts and Blocks
2 Usage Model
2.1 Handling Characters with the Common Script Property
2.2 Handling Nonspacing Marks
2.3 Using Script Names in Regular Expressions
2.4 Use of the Script Property in Rendering Systems
2.5 Limitations
2.6 Spoofing
3 Values
3.1 Relation to ISO 15924 Codes
3.2 Assignment of Script Values
3.3 Assignment of Script Names
4 Data File
4.1 Script Anomalies for East Asian Symbols
Acknowledgments
References
Modifications

1 Introduction

Script: A collection of symbols used to represent textual information in one or more writing systems.

The majority of characters encoded in the Unicode Standard [Unicode] are elements of collections called scripts. Exceptions include symbols, punctuation characters intended for use with multiple scripts, and characters that do not have a stand-alone script identity because they are intended to be used in combination with another character.

Therefore, a text in a given script is likely to consist of characters from that script, together with shared punctuation and characters whose script identity depends on the characters with which they are used.

1.1 Classification of Text by Script Name

The Unicode Character Database [UCD] provides a mapping from Unicode characters to script name values. This information is useful for a variety of tasks that need to analyze a piece of text and determine what parts of it are in which script. Examples include regular expressions or assigning different fonts to parts of a plain text stream based on the prevailing script.

These processes are similar to the task of bibliographers in cataloging documents by their script. However, bibliographers often ignore small inclusions of other scripts in the form of quoted material in cataloging. Conversely, significant differences in the writing style for the same script may be reflected in the bibliographical classification—for example, Fraktur or Gaelic for the Latin script.

Script information is also taken into consideration in collation. The data in the Default Unicode Collation Element Table (DUCET) are grouped by script, so that letters of different script values have different primary sort weights. However, numbers, symbols, and punctuation are not grouped with the letters. For the purposes of ordering, therefore, script is most significant for the letters. For more information, see Unicode Technical Standard #10, “Unicode Collation Algorithm” [UCA].

These examples demonstrate that the definition of *script* depends on the intended

purposes of the classification. *Table 1* summarizes some of the purposes for which text elements can be classified by script.

Table 1. Classification of Text by Script Name

Granularity	Classification	Purpose	Special Values
Document	Bibliographical	Record in which script a text is printed or published; subdivides some scripts—for example, Latin into normal, Fraktur, and Gaelic styles	**Unknown**
Character	Graphological/ typographical	Describe to which script a character belongs based on its origin	
	Orthographical	Describe with which script (or scripts) a character is used	**Common**, **Inherited**
	For collation	Group letters by script in collation element table	
Run	For font binding or search	Determine extent of run of like script in (potentially) mixed-script text	

Bibliographical, graphological, or historical classifications of scripts need different distinctions than the type of text-processing–related needs supported by Unicode script values. The requirements of the task not only affect how fine-grained the classification is, but also what kinds of special values are needed to make the system work. For example, when bibliographers are unable to determine the script of a document, they may classify it using a special value for **Unknown**. In text processing, the identities of all characters are normally known, but some characters may be shared across scripts or attached to any character, thus requiring special values for **Common** and **Inherited**.

Despite these differences, the vast majority of Unicode script values correspond more or less directly to the script identifiers used by bibliographers and others. Unicode script values are therefore mapped to their equivalents in the registry of script identifiers defined by [ISO15924].

1.2 Scripts and Blocks

Unicode characters are also divided into non-overlapping ranges called blocks [Blocks]. Many of these blocks have the same name as one of the scripts because characters of that script are primarily encoded in that block. However, blocks and scripts differ in the following ways:

- Blocks are simply ranges, and often contain code points that are unassigned.
- Characters from the same script may be in several different blocks.
- Characters from different scripts may be in the same block.

As a result, for mechanisms such as regular expressions, using script values produces more meaningful results than simple matches based on block names.

For more information, see *Annex A, Character Blocks,* in Unicode Technical Standard #18,

"Unicode Regular Expressions" [RegEx].

2 Usage Model

The script values form a full partition of the codespace: every code point is assigned a single script value. This value either a specific script value, such as **Cyrillic,** or one of the following three special values:

- **Inherited**—for characters that may be used with multiple scripts, and inherit their script from the preceding characters. Includes nonspacing marks, enclosing marks, and the zero width joiner/non-joiner characters.
- **Common**—for other characters that may be used with multiple scripts.
- **Unknown**—for unassigned, private-use, ~~and~~ noncharacter, and surrogate code points.

As new scripts are added to the standard, more script values will be added. See *Section 3.2, Assignment of Script Values*.

A character is assigned a specific Unicode script value (as opposed to **Common** or **Inherited**) only when it is clearly not used with other scripts. This facilitates the use of the Script property for common tasks such as regular expressions, but means that some characters that are definite members of a given script by their graphology nevertheless are assigned one of the generic values. As more data on the usage of individual characters is collected, the script value assigned to a character may change. If it becomes established that a character is regularly used with more than one script, it will be assigned the **Common**
value, where previously it would have had a more specific script value. However, the opposite type of change is possible as well.

2.1 Handling Characters with the Common Script Property

In determining the boundaries of a run of text in a given script, programs must resolve any of the special script values, such as **Common,** based on the context of the surrounding characters. A simple heuristic uses the script of the preceding character, which works well in many cases. However, this may not always produce optimal results. For example, in the text "... gamma (γ) is ...", this heuristic would cause matching parentheses to be in different scripts.

Generally, paired punctuation, such as brackets or quotation marks, belongs to the enclosing or outer level of the text and should therefore match the script of the enclosing text. In addition, opening and closing elements of a pair resolve to the same script values, where possible. The use of quotation marks is language dependent; therefore it is not possible to tell from the character code alone whether a particular quotation mark is used as an opening or closing punctuation. For more information, see *Section 6.2, General Punctuation,* of [Unicode].

Some characters that are normally used as paired punctuation may also be used singly. An example is U+2019 RIGHT SINGLE QUOTATION MARK, which is also used as *apostrophe,* in which case it no longer acts as an enclosing punctuation. An example from physics would be <ψ| or |ψ>, where the enclosing punctuation characters may not form consistent pairs.

2.2 Handling Nonspacing Marks

Implementations that determine the boundaries between characters of given scripts should

never break between a nonspacing mark (a character with General_Category value of Mn or Me) and its base character. Thus, for boundary determinations and similar sorts of processing, a nonspacing mark—whatever its script value—should inherit the script value of its base character.

Normally, a nonspacing mark has the **Inherited** script value to reflect this. However, in cases where the best interpretation of a nonspacing mark *in isolation* would be a specific script, its script property value may be different from **Inherited**. For example, the Hebrew marks and accents are used only with Hebrew characters and are therefore assigned the **Hebrew** script value.

In cases where the base character itself has the **Common** script, and it is followed by one or more nonspacing marks with a specific script values, such as those Hebrew marks, it may be best for processing to let the base acquire the script value from the mark. This would be the case, for example, if using a graphic symbol as a base to illustrate the placement of nonspacing marks in a particular script.

This approach can be generalized by letting an entire combining character sequence, including spacing combining marks, acquire the script value of the first non-**Inherited**, non-**Common** character in the sequence. Rendering generally works best if an entire combining character sequence can be treated as a segment having a single script, using one set of orthographic rules, and ideally using a single font for display. Exceptional fallback for rendering may be required for defective combining character sequences or in cases where a base character and a combining mark have different specific script values. For example, there may simply be no felicitous way to display a Devanagari combining vowel on a Mongolian consonant base.

2.3 Using Script Names in Regular Expressions

The script property is useful in regular expression syntax for easy specification of spans of text that consist of a single script or mixture of scripts. In general, regular expressions should use specific script values only in conjunction with both **Common** and **Inherited**. For example, to distinguish a sequence of characters appropriate for **Greek**, one would use

```
((Greek | Common) (Inherited | Me | Mn)?)*
```

The preceding expression matches all characters that are either in **Greek** or in **Common** and which are optionally followed by characters in **Inherited**. For completeness, the regular expression also allows any nonspacing or enclosing mark.

Some languages commonly use multiple scripts, so for **Japanese** one might use

```
((Hiragana | Katakana | Han | Latin | Common) (Inherited | Me | Mn)?)*
```

Note that while it is necessary to include Latin in the preceding expression to ensure that it can cover the typical script use found in many Japanese texts, doing so would make it difficult to isolate a run of Japanese inside an English document, for example. For more information, see Unicode Technical Standard #18, “Unicode Regular Expressions” [RegEx].

2.4 Use of the Script Property in Rendering Systems

In rendering systems, it is generally necessary to respect a certain set of orthographic and typographic rules, which vary across the world. For example, the placement of some

diacritics which are nominally rendered above their base may be adjusted to be slightly on the side, as is normally the case for Greek. Another example of variation in rendering is the treatment of spaces in justification. In the absence of an explicit specification of those rules, the script property value of the characters involved provides a good first approximation. Typically, a rendering system will partition a text string into segments of homogeneous script (after resolution of the **Common** and **Inherited** occurrences along the lines described in the previous sections), and then apply the rules appropriate to the script of each segment.

2.5 Limitations

The script values form a full partition of the Unicode codespace, but that partition does not exhaust the possibilities for useful and relevant script-like subsets of Unicode characters.

For example, a user might wish to define a regular expression to span typical mathematical expressions, but the subset of Unicode characters used in mathematics does not correspond to any particular script. Instead, it requires use of the **Math** property, other character properties, and particular subsets of Latin, Greek, and Cyrillic letters. For information on other character properties, see the [UCD].

In texts of an academic, scientific, or engineering nature, the use of isolated Greek characters is common—for example, Ω for ohm; α, β, and γ for types of radioactive decays or in names of chemical compounds; π for 3.1415..., and so on. It is generally undesirable to treat such usage the same as ordinary text in the Greek script. Some commonly used characters, such as μ, already exist twice in the Unicode Standard, but with different script values.

2.6 Spoofing

The script property values may also be useful in providing users feedback to signal possible spoofing, where visually similar characters (*confusable characters*) are substituted in an attempt to mislead a user. For example, a domain name such as `macchiato.com` could be spoofed with `macchiato.com` (using U+03BF GREEK LETTER SMALL LETTER OMICRON for the first “o”) or `macchiato.com` (using U+0441 CYRILLIC SMALL LETTER ES for the first two “c”s). The user can be alerted to odd cases by displaying mixed scripts with different colors, highlighting, or boundary marks: `macchiato.com` or `macchiato.com`, for example.

Possible spoofing is not limited to mixtures of scripts. Even in ASCII, there are confusable characters such as 0 and O, or 1 and l. For a more complete approach, the use of script values needs to be augmented with other information such as General_Category values and lists of individual characters that are not distinguished by other Unicode properties. For additional information, see Unicode Technical Report #36, “Unicode Security Considerations” [Security].

3 Values

Table 2
illustrates some of the script values used in the data file. The short name for the Unicode script value matches the ISO 15924 code. Further subdivisions of scripts by ISO 15924 into varieties are shown in parentheses. For a complete list of values and short names, see the Property Value Aliases [PropValue]. As with all property value aliases, the values in the file are not case sensitive, and the presence of hyphen or underscore is optional. The order in which the scripts are listed here or in the data file is not significant.

Table 2. Unicode Script Values and ISO 15924 Codes

Script Value	ISO 15924
Common	*Zyyy*
Inherited	*Qaai*
Unknown	Zzzz
LATIN	Latn (Latf, Latg)
CYRILLIC	Cyrl (Cyrs)
ARMENIAN	Armn
HEBREW	Hebr
ARABIC	Arab
SYRIAC	Syrc (Syrj, Syrn, Syre)
BRAILLE	Brai
...	...

Although Braille is not a script in the same sense as Latin or Greek, it is given a script value in [Data24]. This is useful for various applications for which these script values are intended, such as matching spans of similar characters in regular expressions.

3.1 Relation to ISO 15924 Codes

ISO 15924: *Code for the Representation of Names of Scripts* [ISO15924] provides an enumeration of four-letter script codes. In the [UCD] file [PropValue], corresponding codes from [ISO15924] are provided as short names for the scripts.

In some cases the match between these script values and the ISO 15924 codes is not precise, because the goals are somewhat different. ISO 15924 is aimed primarily at the bibliographic identification of scripts; consequently, it occasionally identifies varieties of scripts that may be useful for book cataloging, but that are not considered distinct scripts in the Unicode Standard. For example, ISO 15924 has separate script codes for the Fraktur and Gaelic varieties of the Latin script.

Where there are no corresponding ISO 15924 codes, the private-use ones starting with Q are used. Such values are likely to change in the future. In such a case, the Q-names will be retained as aliases in the [PropValue] for backward compatibility.

3.2 Assignment of Script Values

New characters and scripts are continually added to the Unicode Standard. The following methodology is used to assign script values when new characters are added to the Unicode Standard:

A. If a character is used in only one script, assign it to that script
B. Otherwise, nonspacing marks (Mn, Me) and zero width joiner/non-joiner are **Inherited**
C. Otherwise, use **Common**

Script values are not immutable. As more data on the usage of individual characters is collected, script values may be reassigned using the above methodology.

3.3 New Script Names

The following methodology is used to create names for new scripts added to the Unicode

Standard. Script names are limited to

A. Latin letters A–Z or a–z
B. Digits 0–9
C. SPACE and medial HYPHEN-MINUS

Script names are guaranteed to be unique, even when ignoring case differences and the presence of SPACE or HYPHEN-MINUS. Underscores are not used when assigning script names. Similar restrictions apply to block names.

4 Data File

The Scripts.txt data file is available at [Data24]. The format of the file is similar to that of Blocks.txt [Blocks]. The fields are separated by semicolons. The first field contains either a single code point or the first and last code points in a range separated by "..". The second field provides the script value for that range. The comment (after a #) indicates the General_Category and the character name. For each range, it gives the character count in square brackets and uses the names for the first and last characters in the range. For example:

```
0B01;       ORIYA # Mn ORIYA SIGN CANDRABINDU
0B02..0B03; ORIYA # Mc [2] ORIYA SIGN ANUSVARA..ORIYA SIGN VISARGA
```

The value **Unknown**
is the default value, given to all code points that are not explicitly mentioned in the data file.

4.1Script Anomalies for East Asian Symbols

There are a number of compatibility symbols derived from East Asian character sets which have the script value **Common** but whose compatibility decompositions contain characters with other script values. In particular, the parenthesized ideographs, circled ideographs, Japanese era name symbols, and Chinese telegraph symbols in the 3200..33FF range contain Han ideographs, and the squared Latin abbreviation symbols in the same range contain Latin (and occasional Greek) letters. Some of these characters have different scripts in their compatibility decompositions. What this means is that script extents calculated on the basis of the script property value of the symbols themselves will differ from script extents calculated on NFKD normalized text, in which these characters decompose into sequences including the Han and/or Latin characters.

The UTC has determined that since these symbols may be used with multiple scripts in Chinese, Japanese, and/or Korean contexts, their script value should simply be left as **Common**. There are other, more reliable clues about the behavior of these compatibility symbols, such as their association with East Asian character sets, which can be used by rendering systems to assure their appropriate display and appropriate font choice. This determination is somewhat different from that for the more script-specific parenthesized and circled Hangul and Katakana symbols in the same range, which *are* given specific script values. At this point keeping the script value stable for these compatibility symbols is more useful for implementers than attempting to reconcile this distinction in treatment by modifying script values for them. Implementations that wish to have script values that are preserved over compatibility equivalence would tailor the script values for these characters.

Acknowledgments

Thanks to ~~Ken Whistler and~~
Julie Allen for comments on this annex, including earlier versions. Asmus Freytag added significant sections to the text for Revisions 7 and 9. Eric Muller added the new Section 2.4 for Revision 10 and suggested modifications for Section 2.2.

References

For references for this annex, see Unicode Standard Annex #41, "Common References for Unicode Standard Annexes."

Modifications

The following summarizes modifications from the previous revision of this annex.

Revision 10

- Prepared for Unicode 5.1.0 release and updated title. [KW]
- Added surrogates to list of code points which get **Unknown** script value. [KW]
- Added new Section 2.4 regarding use of the script property in rendering systems. [EM]
- Added clarification in Section 2.2 regarding script inheritance in combining character sequences. [EM, KW]
- Added new Section 4.1 noting script anomalies for some East Asian compatibility symbols. [KW]

Revision 9

- Prepared for Unicode 5.0.0 release [AF].
- Added **Unknown**, and made it default value instead of **Common** [AF].

Revision 8 being a proposed update, only changes between revisions 9 and 7 are noted here.

Revision 7

- Prepared for Unicode 4.1 release [AF].
- Split section 3.2 and added section 3.3 [AF].
- Major rewrite of Introduction and usage model. [AF].
- Added section on Maintenance and table of classifications types [AF].

Revision 6 being a proposed update, only changes between revisions 7 and 5 are noted here.

Revision 5

- Changed to Unicode Standard Annex.
- Added note on the stability of Q names
- Abbreviated the list of values, so that people would not get the mistaken impression that it was complete
- Added note on Braille
- Added note on Mn, Me characters
- Added note on use of scripts with regard to spoofing

- Minor edits

Revision 4

- Updated references, including reference to Property Value Aliases
- Clarified that the list is for illustration only; the definitive values are in the UCD
- Minor edits

Revision 3

- Minor link editing only
