

The Unicode Standard

Version 6.0 – Core Specification

To learn about the latest version of the Unicode Standard, see <http://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.

Copyright © 1991–2011 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 <http://www.unicode.org/reporting.html>. For information about the Unicode terms of use, please see <http://www.unicode.org/copyright.html>.

The Unicode Standard / the Unicode Consortium ; edited by Julie D. Allen ... [et al.]. — Version 6.0.

Includes bibliographical references and index.

ISBN 978-1-936213-01-6 (<http://www.unicode.org/versions/Unicode6.0.0/>)

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

QA268.U545 2011

ISBN 978-1-936213-01-6

Published in Mountain View, CA

February 2011

Appendix D

Changes from Previous Versions

This appendix provides version history of the standard and summarizes updates that have been made to conformance specifications, character content, and data files in the Unicode Character Database since the publication of *The Unicode Standard, Version 5.0*.

D.1 Versions of the Unicode Standard

The Unicode Technical Committee updates the Unicode Standard to respond to the needs of implementers and users while maintaining consistency with ISO/IEC 10646. The relationship between these versions of Unicode and ISO/IEC 10646 is shown in *Table D-1*. For more detail on the relationship of Unicode and ISO/IEC 10646, see *Appendix C, Relationship to ISO/IEC 10646*.

Table D-1. Versions of Unicode and ISO/IEC 10646-1

Year	Version	Published	ISO/IEC 10646-1
1991	Unicode 1.0	Vol. 1, Addison-Wesley	Basis for Committee Draft 2 of 10646-1
1992	Unicode 1.0.1	Vol. 1, 2, Addison-Wesley	Interim merger version
1993	Unicode 1.1	Technical Report #4	Matches ISO 10646-1
1996	Unicode 2.0	Addison-Wesley	Matches ISO 10646-1 plus amendments
1998	Unicode 2.1	Technical Report #8	Matches ISO 10646-1 plus amendments
2000	Unicode 3.0	Addison-Wesley	Matches ISO 10646-1 second edition
2001	Unicode 3.1	Standard Annex #27	Matches ISO 10646-1 second edition plus two characters, 10646-2 first edition
2002	Unicode 3.2	Standard Annex #28	Matches ISO 10646-1 second edition plus amendment, 10646-2 first edition
2003	Unicode 4.0	Addison-Wesley	Matches ISO 10646:2003, third version
2005	Unicode 4.1	Web publication	Matches ISO 10646:2003, third version, plus Amd. 1
2006	Unicode 5.0	Addison-Wesley (2007)	Matches ISO 10646:2003, third version, plus Amd. 1, Amd. 2, and four characters from Amd. 3
2008	Unicode 5.1	Web publication	Matches ISO 10646:2003, third version, plus Amd. 1 through Amd. 4
2009	Unicode 5.2	Web publication	Matches ISO 10646:2003, third version, plus Amd. 1 through Amd. 6
2010	Unicode 6.0	Web publication	Matches ISO 10646:2011, second edition

The Unicode Standard has grown from having 28,294 assigned graphic and format characters in Version 1.0, to having 109,384 characters in Version 6.0. *Table D-2* documents the number of code points allocated in the different versions of the Unicode Standard. The row in *Table D-2* labeled “Graphic + Format” represents the traditional count of Unicode characters and is the typical answer to the question, “How many characters are in the Unicode

Standard?” In *Table D-2* the numbers for Han Compatibility include the 12 unified ideographs encoded in the CJK Compatibility Ideographs block.

Table D-2. Allocation of Code Points by Type

	V3.1	V3.2	V4.0	V4.1	V5.0	V5.1	V5.2	V6.0
Alphabets, Symbols	11,798	12,753	13,973	15,117	16,486	18,101	20,588	22,454
Han (URO)	20,902	20,902	20,902	20,902	20,902	20,902	20,902	20,902
Han (URO Extension)				22	22	30	38	38
Han Extension A	6,582	6,582	6,582	6,582	6,582	6,582	6,582	6,582
Han Extension B	42,711	42,711	42,711	42,711	42,711	42,711	42,711	42,711
Han Extension C							4,149	4,149
Han Extension D								222
Han Compatibility	844	903	903	1,009	1,009	1,009	1,012	1,012
Subtotal Han	71,039	71,098	71,098	71,226	71,226	71,234	75,394	75,616
Hangul Syllables	11,172	11,172	11,172	11,172	11,172	11,172	11,172	11,172
Graphic Characters	94,009	95,023	96,243	97,515	98,884	100,507	107,154	109,242
Format Characters	131	133	139	140	140	141	142	142
Graphic + Format	94,140	95,156	96,382	97,655	99,024	100,648	107,296	109,384
Controls	65	65	65	65	65	65	65	65
Private Use	137,468	137,468	137,468	137,468	137,468	137,468	137,468	137,468
Total Assigned	231,673	232,689	233,915	235,188	236,557	238,181	244,829	246,917
Surrogate Code Points	2,048	2,048	2,048	2,048	2,048	2,048	2,048	2,048
Noncharacters	66	66	66	66	66	66	66	66
Total Designated	233,787	234,803	236,029	237,302	238,671	240,295	246,943	249,031
Reserved Code Points	880,325	879,309	878,083	876,810	875,441	873,817	867,169	865,081

Table D-3 lists the allocation of code points by type for earlier, historic versions of the Unicode Standard prior to Version 3.1. In some cases the values in this table differ slightly from summary statistics published in earlier versions of the standard, primarily due to a refined accounting of the allocations in Unicode 1.0.

Table D-3. Allocation of Code Points by Type (Early Versions)

	V1.0.0	V1.0.1	V1.1	V2.0	V2.1	V3.0
Alphabets, Symbols	4,734	4,728	6,290	6,491	6,493	10,210
Han (URO)		20,902	20,902	20,902	20,902	20,902
Han Extension A						6,582
Han Compatibility		302	302	302	302	302
Subtotal Han		21,204	21,204	21,204	21,204	27,786
Hangul Syllables	2,350	2,350	6,656	11,172	11,172	11,172
Graphic Characters	7,084	28,282	34,150	38,867	38,869	49,168
Format Characters	12	12	18	18	18	26
Graphic + Format	7,096	28,294	34,168	38,885	38,887	49,194
Controls	65	65	65	65	65	65
Private Use	5,632	6,144	6,400	137,468	137,468	137,468
Total Assigned	12,793	34,503	40,633	176,418	176,420	186,727
Surrogate Code Points				2,048	2,048	2,048
Noncharacters	2	2	2	34	34	34
Total Designated	12,795	34,505	40,635	178,500	178,502	188,809
Reserved Code Points	52,741	31,031	24,901	935,612	935,610	925,303

D.2 Clause and Definition Updates

Several updates were made to definitions and conformance clauses in Version 5.1 primarily to address potential security exploits. The updates also reflect updated Consortium policies to increase property stability, and include a few other textual clarifications.

Table D-4 provides a list of all clauses and definitions that were clarified, changed, or newly added in Version 5.1.

Table D-4. Version 5.1 Clause and Definition Updates

Number	Clause or Definition	Type of Update
C7	Modification	Clarification
D40	Stable property	Clarification
D51a	Extended base	New
D56a	Extended combining character sequence	New
D60	Grapheme cluster	Changed
D61	Extended grapheme cluster	Changed
D84a	Ill-formed code unit subsequence	New
D85	Well-formed	Changed
D85a	Minimal well-formed code unit subsequence	New
D86	Well-formed UTF-8 code unit sequence	Changed
D121	Case-ignorable	Changed

For Version 5.2, a number of updates were made to incorporate the specification of the normalization algorithm into *Chapter 3, Conformance*, including definitions formerly specified in Unicode Standard Annex #15, “Unicode Normalization Forms.” Other changes include those to tighten security for the handling of noncharacters, and new or changed definitions for deprecated character, code point type, and contributory property. Due to the creation of a new section on normalization, many definitions were renumbered, and a few were moved into other sections.

Table D-5 provides a list of all clauses and definitions that were clarified, changed, newly added, renumbered, or moved in Version 5.2.

Table D-5. Version 5.2 Clause and Definition Updates

Number	Clause or Definition	Type of Update
C7	Modification	Clarification
C13	Normalization	Changed
C14	Normalization	Changed
D10a	Code point type	New
D13	Deprecated character	Clarification
D35a	Contributory property	New
D61a	Dependence	Renumbered (was D102)
D61b	Graphical application	Renumbered (was D103)
D107	Starter	New
D108	Reorderable pair	New
D109	Canonical Ordering Algorithm	New
D110	Singleton decomposition	New
D111	Non-starter decomposition	New
D112	Composition exclusion	New
D113	Full composition exclusion	New
D114	Primary composite	New

Table D-5. Version 5.2 Clause and Definition Updates (Continued)

Number	Clause or Definition	Type of Update
D115	Blocked	New
D116	Non-blocked pair	New
D117	Canonical Composition Algorithm	New
D118	Normalization Form D	New
D119	Normalization Form KD	New
D120	Normalization Form C	New
D121	Normalization Form KC	New
D122 to D133	Hangul syllables	Renumbered (were D107 to D118)
D134	Standard Korean syllable	Renumbered (was D119)
D135 to D138	Case	Renumbered (were D120 to D123)
D139 to D143	Case detection	Renumbered (were D124 to D128)
D144 to D146	Caseless matching	Renumbered (were D129 to D131)

Version 6.0 of the Unicode Standard updated the explanatory text of a few conformance clauses to highlight security considerations. Extensive new text was added to clarify the best practices for using U+FFFD, two new definitions were added in support of the Canonical Composition Algorithm, and new definitions, rules, and explanatory text were added to default case algorithms.

Table D-6 provides a list of all clauses and definitions that were clarified, changed, or newly added in Version 6.0.

Table D-6. Version 6.0 Clause and Definition Updates

Number	Clause or Definition	Type of Update
C7	Modification	Clarification
C10	Character encoding forms	Clarification
D92	UTF-8	Clarification
D93a to D93b	Constraints on Conversion Process	New
D110a to D110b	Canonical Composition Algorithm	New
D111	Canonical Composition Algorithm	Clarification
D138	Default Case Algorithms	Clarification
D144 to D146	Default Case Algorithms	Clarification
D147	Default Case Algorithms	New

D.3 Changes from Version 5.2 to Version 6.0

For a summary of major changes in Version 6.0, see:

<http://www.unicode.org/versions/Unicode6.0.0/>

For a detailed listing of changes to the Unicode Character Database for Version 6.0, see:

http://www.unicode.org/reports/tr44/tr44-6.html#Change_History

D.4 Changes from Version 5.1 to Version 5.2

For a summary of major changes in Version 5.2, see:

<http://www.unicode.org/versions/Unicode5.2.0/>

For a detailed listing of changes to the Unicode Character Database for Version 5.2, see:

http://www.unicode.org/reports/tr44/tr44-6.html#Change_History

D.5 Changes from Version 5.0 to Version 5.1

New Characters Added

In total, 1,624 new characters were encoded in the Unicode Standard, Version 5.1. These additions include characters required for Malayalam and Myanmar and important individual characters such as U+1E9E LATIN CAPITAL LETTER SHARP S for German. Version 5.1 extends support for languages in Africa, India, Indonesia, Myanmar, and Vietnam, with the addition of eight minority scripts: Cham, Kayah Li, Lepcha, Ol Chiki, Rejang, Saurashtra, Sundanese, and Vai. Scholarly support includes important editorial punctuation marks, as well as the Carian, Lycian, and Lydian scripts, and the Phaistos Disc Symbols. Other new symbol sets include dominoes, Mahjong, dictionary punctuation marks, and significant math additions.

The new character additions were to both the BMP and the SMP (Plane 1). See the file `DerivedAge.txt` in the Unicode Character Database (UCD) for a complete listing of ranges where characters were added in Unicode 5.1.

Unicode Character Database Changes

The Unicode Character Database (UCD) was extended to cover the character repertoire additions, and new block definitions and script values were added. A number of other updates were made; the most significant are listed in this section. For more details on the enhancements to the UCD, see Unicode Standard Annex #44, “Unicode Character Database (UCD).”

New with Unicode Version 5.1 is an XML version of the complete Unicode Character Database. For details see Unicode Standard Annex #42, “An XML Representation of the UCD.”

Arabic Shaping. A new joining group, BURUSHASKI YEH BARREE, was added.

Bidirectional Behavior. Directional quotation marks changed to Bidi_Mirrored=N and the glyph mappings for the 11 characters that are now no longer mirrored were removed. The Bidi_Mirrored property change implemented Corrigendum #6: Bidi Mirroring, which mandated a partial reversion of the change for Version 5.0 related to Public Review Issue #91. The glyph mappings for U+2278 and U+2279 were updated to (BEST FIT).

Five Arabic characters that surround numeral sequences (U+0600..U+0603, U+06DD) were changed from Bidirectional_Class=AL to AN. This change in bidi class has the effect of putting the surrounding sign and its associated numeral sequence in the same directional run, thus simplifying implementations.

General Category. U+05BE HEBREW PUNCTUATION MAQAF was changed from gc=Po to gc=Pd. U+02EC MODIFIER LETTER VOICING and U+0374 GREEK NUMERAL SIGN were changed from gc=Sk to gc=Lm. U+A802 SYLOTI NAGRI SIGN DVISVARA was changed from gc=Mc to gc=Mn.

Named Character Sequences. Thirty-four new standardized named character sequences were added for Lithuanian as well as detailed documentation of provisional named character sequences for Tamil.

New Property Definitions and Values. New property values were added in support of text segmentation: Sentence_Break property values CR, LF, Extend, and SContinue,

Word_Break property values CR, LF, Newline, Extend, and MidNumLet, Grapheme_Cluster_Break property values Prepend and SpacingMark. The definition of the Default_Ignorable_Code_Point property was revised.

Other Updates. The middle dots (U+00B7, U+0387) were added to identifiers by changing them to Other_ID_Continue=Y. Ten compatibility ideographs were given numeric values. For consistency, the halfwidth Katakana sound marks (U+FF9E, U+FF9F) were added to Grapheme_Extend by including them in Other_Grapheme_Extend. The tag characters (U+E0001, U+E0020..U+E007F) were added to Deprecated. Other_Math values were adjusted for a number of mathematical symbols, and U+05BE was changed to Dash=Y, consistent with the change in its General Category.

Unihan. Documentation of U-source ideographs was added. Two existing unified ideographs, U+6F06 and U+9621, were given numeric values. One new provisional property, kXHC1983, was added. Corrections and additions to other properties were also made.

Conformance and Stability

Stability Policies. The Unicode Character Encoding Stability Policy has been updated. This update strengthens normalization stability, adds stability policy for case pairs, and extends constraints on property values. For the current statement of these policies, see the page detailing the Unicode Character Encoding Stability Policy on the Unicode Web site at: http://www.unicode.org/policies/stability_policy.html.

Important Clarification of UTF-8 Conformance. In Version 5.1, the text regarding ill-formed UTF-8 code unit sequences was extended with new definitions that clarify well-formed and ill-formed code unit subsequences in strings. Conversion of ill-formed UTF-8 code unit sequences was also clarified to prevent over-consumption of ill-formed sequences. For more details on potential security issues because of over-consumption of ill-formed sequences, see Unicode Technical Report #36, “Unicode Security Considerations.”

Updates to Definitions of Character Sequences. In order to take into account the normalization behavior of Hangul syllables and conjoining jamo sequences, additional definitions for extended base and extended combining character sequence were added to the standard. See D51a and D56a, respectively. In addition, the existing definitions of grapheme cluster and extended grapheme cluster were slightly modified to bring them into line with Unicode Standard Annex #29, “Unicode Text Segmentation,” where they are defined algorithmically. See D60 and D61.

Updates to Table of Named Unicode Algorithms. Table 3-1, *Named Unicode Algorithms* and the associated explanatory text was updated to account for some slight changes in naming conventions for Unicode algorithms in Version 5.1.

Updates to Default Algorithms. In Section 3.13, *Default Case Algorithms*, D121 was updated to include the value MidNumLet in the definition of case-ignorable. This change was occasioned by the split of the Word_Break property value MidLetter into MidLetter and MidNumLet.

In Section 3.13, *Default Case Algorithms*, the first paragraph under “Default Case Conversion” was updated and a new paragraph following rules R1-R4 was added.

Updates to Stability of Properties. As a result of the strengthening of the Normalization Stability Policy, the Canonical_Combining_Class property became an immutable property, rather than merely a stable property. To adjust for this change, the first bullet after D40 was removed from the text.

Updates for Security. Because of the potential security issues involved in the deletion of characters, an explanatory bullet was added to the text after the first two bullet items following Conformance Clause C7.

Other Updates to the Text of the Standard

The text of the standard was revised to document the new and changed character properties including the new Arabic joining group `BURUSHASKI YEH BARREE`. Explanation of character properties and behavior was expanded for overlaid diacritics, modifier letters, `General_Category` value assignments, Hangul jamo handling, tailored casing operations, and the concepts of lowercase and uppercase. The text was revised to clarify canonical equivalence issues for Greek punctuation, Coptic font styles, rendering default ignorable code points, and stateful format controls. Other updates covered handling noncharacters, tag characters, and the Ideographic Variation Database.

Guidance on the use of the Myanmar script was significantly improved with a new block introduction. For the Malayalam script, the representation of the newly-encoded Malayalam chillu characters was described in detail. In addition, new scripts and significant symbol additions were documented.

Unicode Standard Annexes

The following Unicode Standard annexes were added in Version 5.1:

- UAX #38, “Unicode Han Database (Unihan),” provides detailed information about CJK ideograph data in Unicode.
- UAX #42: “Unicode Character Database in XML,” provides the UCD in XML.
- UAX #44, “Unicode Character Database (UCD),” provides general information about the UCD.

Some of the Unicode Standard Annexes had minor changes to their titles for consistency with revised technical report naming practices. The following summarizes the more significant changes in the Unicode Standard Annexes. Detailed notes can be found in the modifications section of each document.

- In UAX #9, “Unicode Bidirectional Algorithm,” clarified one conformance clause and added new informative text.
- In UAX #14, “Unicode Line Breaking Algorithm,” modified conformance to allow for additional tailoring, made a number of property changes, split the rule LB12 to accommodate Polish and Portuguese hyphenation, added new informative text and new test files, and edited the text for clarity.
- In UAX #15, “Unicode Normalization Forms,” added new sections on stabilized strings (including conformance clauses), buffering guidelines for normalization, additional new informative text. and edited the text for clarity.
- In UAX #24, “Unicode Script Property,” added surrogates to the list of code points which get the “Unknown” script value, added new informative text on rendering systems and script inheritance in combining character sequences, and edited the text for clarity.
- In UAX #29, “Unicode Text Segmentation,” added extended combining character sequences and improved word and sentence segmentation through the addition of new properties and modified property values and rules.
- In UAX #31, “Unicode Identifier and Pattern Syntax,” added filtered identifiers and rules which affect the use of joiners in specific contexts for Indic languages,

new informative text on the canonical equivalence of identifiers and the handling of format characters, and edited the text for clarity.

Errata

The following list summarizes errata incorporated into the standard in the release of Unicode Version 5.1. For a detailed description of these corrections see “Errata Fixed in Unicode 5.1” (<http://www.unicode.org/versions/Unicode5.1.0/erratafixed.html>). This page contains the definitive listing of all errata of record since the publication of *The Unicode Standard, Version 5.0* and considered resolved by the release of Unicode Version 5.1.

- The representative glyph for U+1D81 LATIN SMALL LETTER D WITH PALATAL HOOK was corrected.
- The representative glyph for U+1E9A LATIN SMALL LETTER A WITH RIGHT HALF RING was corrected.
- The representative glyphs for U+0333 COMBINING DOUBLE LOW LINE and U+0347 COMBINING EQUALS SIGN BELOW were corrected.
- The representative glyphs for U+0460 CYRILLIC CAPITAL LETTER OMEGA and U+047E CYRILLIC CAPITAL LETTER OT were corrected.
- The representative glyphs for U+075E ARABIC LETTER AIN WITH THREE DOTS POINTING DOWNWARDS ABOVE and U+075F ARABIC LETTER AIN WITH TWO DOTS VERTICALLY ABOVE were swapped.
- The representative glyphs for U+0478 CYRILLIC CAPITAL LETTER UK and U+0479 CYRILLIC SMALL LETTER UK were corrected.
- The representative glyph for U+2626 ORTHODOX CROSS was corrected.
- In Unicode Standard Annex #15, “Unicode Normalization Forms,” an erroneous statement in the last paragraph of Section 14, Detecting Normalization Forms was corrected.
- The representative glyphs for U+047C CAPITAL LETTER OMEGA WITH TITLO and U+047D SMALL LETTER OMEGA WITH TITLO were corrected.
- The sample code in Section 7 of Unicode Standard Annex #14, “Unicode Line Breaking Algorithm,” was updated to correctly handle leading spaces.
- In the file `DerivedCoreProperties.txt` in the Version 5.0 Unicode Character Database, the stated rule in the comments for the generation of the `Default_Ignorable_Code_Point` property was incomplete. It was corrected to also include all characters with the `Variation_Selector` property.
- U+00A0 NO-BREAK SPACE was corrected to have the `Sentence_Break` property value `Sp` in a future version of the standard.
- The representative glyph for U+1031 MYANMAR VOWEL SIGN E was corrected.
- The `Index.txt` file in version 5.0.0 of the Unicode Character Database is not valid UTF-8. A fix was documented.