Appendix D

Changes from Unicode Version 2.0

D.1 Versions of the Unicode Standard

The Unicode Technical Committee periodically updates the Unicode Standard to respond to industry needs and to maintain consistency with ISO/IEC 10646. There have been five formally numbered major and minor versions of the Unicode Standard. 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 the Unicode Standard and ISO/IEC 10646-1

Year	Version	Published	ISO/IEC 10646-1
1991	Unicode 1.0	Vol. 1, Addison-Wesley	Basis of Draft-2 ISO 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

The Unicode Standard has grown from having 28,302 assigned character values in Version 1.0, to having 49,194 assigned character values in Version 3.0. *Table D-2* documents the number of characters assigned in the different versions of the Unicode Standard.

Table D-2. Number of Assigned Characters

	V 1.0	V 1.1	V 2.0	V 2.1	V 3.0
Alphabetics, Symbols	4,748	6,309	6,509	6,511	10,236
Han (URO)	20,902	20,902	20,902	20,902	20,902
Han Extension A					6,582
Han Compatibility	302	302	302	302	302
Hangul Syllables	2,350	6,656	11,172	11,172	11,172
Total assigned characters	28,302	34,169	38,885	38,887	49,194
Private Use	5,632	6,400	6,400	6,400	6,400
Surrogates			2,048	2,048	2,048
Controls	65	65	65	65	65
Not Characters	2	2	2	2	2

Table D-2. Number of Assigned Characters (Continued)

	V 1.0	V 1.1	V 2.0	V 2.1	V 3.0
Total assigned 16-bit code values	34,001	40,636	47,400	47,402	57,709
Unassigned 16-bit code values	31,535	24,900	18,136	18,134	7,827

This appendix enumerates updates to conformance criteria and to character content and semantics made to the Unicode Standard, Version 2.1, and to Version 3.0. For further information on all major, minor, and update versions of the Unicode Standard, see the Unicode Web site at http://www.unicode.org/unicode/standard/versions/.

D.2 Changes from Unicode Version 2.0 to Version 2.1

New Characters Added

Two new characters were added to the Unicode Standard, Version 2.1:

U+20AC EURO SIGN

U+FFFC OBJECT REPLACEMENT CHARACTER

Character Semantics Changes

Significant clarifications or modifications to character semantics include the following:

Apostrophe. Because the character U+0027 APOSTROPHE is very ambiguous, the preferred character for apostrophe was documented as either U+02BC MODIFIER LETTER APOSTROPHE or U+2019 RIGHT SINGLE QUOTATION MARK.

Bidirectional Properties. Certain characters were given new bidirectional properties definitions. U+0026 AMPERSAND and U+0040 COMMERCIAL AT were changed from left to right (L) to other neutral (ON). U+002E FULL STOP and U+2007 FIGURE SPACE were changed from European number separator (ES) to common number separator (CS).

Identifier Syntax. Corrections were made to the implementation guidelines for identifier syntax, adding or removing a few characters from the recommended set for inclusion in identifiers.

Mathematical and Letter Properties. A number of characters were given the mathematical property; see the list found in Unicode Technical Report #8, "The Unicode Standard, Version 2.1," on the CD-ROM or the up-to-date version on the Unicode Web site for details. Two characters, U+02BC MODIFIER LETTER APOSTROPHE and U+055A ARMENIAN APOSTROPHE, were removed from the list of the letter property.

Changes Affecting Conformance

Overall Unicode conformance criteria as described in Chapter 3 of Version 2.0 were unchanged. Aspects of the Unicode bidirectional algorithm were modified, Hangul syllable decompositions were clarified, and some normative character property values were changed. For details of these changes see Unicode Technical Report #8, "The Unicode Standard, Version 2.1," on the CD-ROM or the up-to-date version on the Unicode Web site.

D.3 Changes from Unicode Version 2.1 to Version 3.0

New Characters Added

In Version 3.0 of the Unicode Standard, 10,307 new characters have been added, as shown in *Table D-3*.

Table D-3. New Characters Added

Allocation	Count	Character Name
01F6..01F9	4	LATIN CAPITAL LETTER HWAIR LATIN CAPITAL LETTER WYNN LATIN CAPITAL LETTER N WITH GRAVE LATIN SMALL LETTER N WITH GRAVE
0218..021F	8	LATIN CAPITAL LETTER S WITH COMMA BELOW LATIN SMALL LETTER S WITH COMMA BELOW LATIN CAPITAL LETTER T WITH COMMA BELOW LATIN SMALL LETTER T WITH COMMA BELOW LATIN CAPITAL LETTER YOGH LATIN SMALL LETTER YOGH LATIN CAPITAL LETTER H WITH CARON LATIN SMALL LETTER H WITH CARON
0222..0233	18	LATIN CAPITAL LETTER OU LATIN SMALL LETTER OU LATIN CAPITAL LETTER Z WITH HOOK LATIN SMALL LETTER Z WITH HOOK LATIN CAPITAL LETTER A WITH DOT ABOVE LATIN SMALL LETTER A WITH DOT ABOVE LATIN CAPITAL LETTER E WITH CEDILLA LATIN SMALL LETTER E WITH CEDILLA LATIN CAPITAL LETTER O WITH DIAERESIS AND MACRON LATIN SMALL LETTER O WITH DIAERESIS AND MACRON LATIN CAPITAL LETTER O WITH TILDE AND MACRON LATIN SMALL LETTER O WITH TILDE AND MACRON LATIN CAPITAL LETTER O WITH DOT ABOVE LATIN SMALL LETTER O WITH DOT ABOVE LATIN CAPITAL LETTER O WITH DOT ABOVE AND MACRON LATIN SMALL LETTER O WITH DOT ABOVE AND MACRON LATIN CAPITAL LETTER Y WITH MACRON LATIN SMALL LETTER Y WITH MACRON
02A9..02AD	5	LATIN SMALL LETTER FENG DIGRAPH LATIN SMALL LETTER LS DIGRAPH LATIN SMALL LETTER LZ DIGRAPH LATIN LETTER BILABIAL PERCUSSIVE LATIN LETTER BIDENTAL PERCUSSIVE
02DF, 02EA..02EE	6	MODIFIER LETTER CROSS ACCENT MODIFIER LETTER YIN DEPARTING TONE MARK MODIFIER LETTER YANG DEPARTING TONE MARK MODIFIER LETTER VOICING MODIFIER LETTER UNASPIRATED MODIFIER LETTER DOUBLE APOSTROPHE

Table D-3. New Characters Added (Continued)

Allocation	Count	Character Name
0346..034E, 0362	10	COMBINING BRIDGE ABOVE COMBINING EQUALS SIGN BELOW COMBINING DOUBLE VERTICAL LINE BELOW COMBINING LEFT ANGLE BELOW COMBINING NOT TILDE ABOVE COMBINING HOMOTHETIC ABOVE COMBINING ALMOST EQUAL TO ABOVE COMBINING LEFT RIGHT ARROW BELOW COMBINING UPWARDS ARROW BELOW COMBINING DOUBLE RIGHTWARDS ARROW BELOW
03D7	1	GREEK KAI SYMBOL
03DB, 03DD, 03DF, 03E1	4	GREEK SMALL LETTER STIGMA GREEK SMALL LETTER DIGAMMA GREEK SMALL LETTER KOPPA GREEK SMALL LETTER SAMPI
0400, 040D, 0450, 045D	4	CYRILLIC CAPITAL LETTER IE WITH GRAVE CYRILLIC SMALL LETTER IE WITH GRAVE CYRILLIC CAPITAL LETTER II WITH GRAVE CYRILLIC SMALL LETTER II WITH GRAVE
0488..0489	2	CYRILLIC HUNDRED THOUSANDS SIGN CYRILLIC MILLIONS SIGN
048E..048F, 04EC..04ED	4	CYRILLIC CAPITAL LETTER ER WITH TICK CYRILLIC SMALL LETTER ER WITH TICK CYRILLIC CAPITAL LETTER E WITH DIAERESIS CYRILLIC SMALL LETTER E WITH DIAERESIS
058A	1	ARMENIAN HYPHEN
0653..0655, 06B8..06B9, 06BF, 06CF, 06FA..06FE	12	ARABIC MADDAH ABOVE ARABIC HAMZA ABOVE ARABIC HAMZA BELOW ARABIC LETTER LAM WITH THREE DOTS BELOW ARABIC LETTER NOON WITH DOT BELOW ARABIC LETTER TCHEH WITH DOT ABOVE ARABIC LETTER WAW WITH DOT ABOVE ARABIC LETTER SHEEN WITH DOT BELOW ARABIC LETTER DAD WITH DOT BELOW ARABIC LETTER GHAIN WITH DOT BELOW ARABIC SIGN SINDHI AMPERSAND ARABIC SIGN SINDHI POSTPOSITION MEN
0700..074A	71	Syriac
0780..07B1	50	Thaana
0D80..0DFF	80	Sinhala
0F6A, 0F96, 0FAE..0FCF	25	Tibetan Extensions
1000..1059	78	Myanmar
1200..137F	346	Ethiopic
13A0..13FF	85	Cherokee
1401..1676	630	Canadian Syllabics
1680..169F	29	Ogham
16A0..16FF	81	Runic
1780..17E9	103	Khmer
1800..18A9	155	Mongolian

Table D-3. New Characters Added (Continued)

Allocation	Count	Character Name
202F, 2048..204D	7	NARROW NO-BREAK SPACE QUESTION EXCLAMATION MARK EXCLAMATION QUESTION MARK TIRONIAN SIGH ET REVERSED PILCROW SIGN BLACK LEFTWARDS BULLET BLACK RIGHTWARDS BULLET
20AD..20AF	3	KIP SIGN TUGRIK SIGN DRACHMA SIGN
20E2..20E3	2	COMBINING ENCLOSING SCREEN COMBINING ENCLOSING KEYCAP
2139..213A	2	INFORMATION SOURCE ROTATED CAPITAL Q
2183	1	ROMAN NUMERAL REVERSED ONE HUNDRED
21EB..21F3	9	UPWARDS WHITE ARROW ON PEDESTAL UPWARDS WHITE ARROW ON PEDESTAL WITH HORIZONTAL BAR UPWARDS WHITE ARROW ON PEDESTAL WITH VERTICAL BAR UPWARDS WHITE DOUBLE ARROW UPWARDS WHITE DOUBLE ARROW ON PEDESTAL RIGHTWARDS WHITE ARROW FROM WALL NORTH WEST ARROW TO CORNER SOUTH EAST ARROW TO CORNER UP DOWN WHITE ARROW
2301, 237B, 237D, 237E..237F	5	ELECTRIC ARROW NOT CHECK MARK SHOULDERED OPEN BOX BELL SYMBOL VERTICAL LINE WITH MIDDLE DOT
2380..238C	13	INSERTION SYMBOL CONTINUOUS UNDERLINE SYMBOL DISCONTINUOUS UNDERLINE SYMBOL EMPHASIS SYMBOL COMPOSITION SYMBOL WHITE SQUARE WITH CENTRE VERTICAL LINE ENTER SYMBOL ALTERNATIVE KEY SYMBOL HELM SYMBOL CIRCLED HORIZONTAL BAR WITH NOTCH CIRCLED TRIANGLE DOWN BROKEN CIRCLE WITH NORTHWEST ARROW UNDO SYMBOL
238D..239A	14	MONOSTABLE SYMBOL HYSTERESIS SYMBOL OPEN-CIRCUIT-OUTPUT H-TYPE SYMBOL OPEN-CIRCUIT-OUTPUT L-TYPE SYMBOL PASSIVE-PULL-DOWN-OUTPUT SYMBOL PASSIVE-PULL-UP-OUTPUT SYMBOL DIRECT CURRENT SYMBOL FORM TWO SOFTWARE-FUNCTION SYMBOL APL FUNCTIONAL SYMBOL QUAD DECIMAL SEPARATOR KEY SYMBOL PREVIOUS PAGE NEXT PAGE PRINT SCREEN SYMBOL CLEAR SCREEN SYMBOL
2425..2426	2	SYMBOL FOR DELETE FORM TWO SYMBOL FOR SUBSTITUTE FORM TWO

Table D-3. New Characters Added (Continued)

Allocation	Count	Character Name
25F0..25F7	8	WHITE SQUARE WITH UPPER LEFT QUADRANT WHITE SQUARE WITH LOWER LEFT QUADRANT WHITE SQUARE WITH LOWER RIGHT QUADRANT WHITE SQUARE WITH UPPER RIGHT QUADRANT WHITE CIRCLE WITH UPPER LEFT QUADRANT WHITE CIRCLE WITH LOWER LEFT QUADRANT WHITE CIRCLE WITH LOWER RIGHT QUADRANT WHITE CIRCLE WITH UPPER RIGHT QUADRANT
2619, 2670..2671	3	REVERSED ROTATED FLORAL HEART BULLET WEST SYRIAC CROSS EAST SYRIAC CROSS
2800..28FF	256	Braille Pattern Symbols
2E80..2EF3	115	CJK Radicals Supplement
2F00..2FD5	214	KangXi radicals
2FF0..2FFB	12	IDEOGRAPHIC DESCRIPTION CHARACTER LEFT TO RIGHT IDEOGRAPHIC DESCRIPTION CHARACTER ABOVE TO BELOW IDEOGRAPHIC DESCRIPTION CHARACTER LEFT TO MIDDLE AND RIGHT IDEOGRAPHIC DESCRIPTION CHARACTER ABOVE TO MIDDLE AND BELOW IDEOGRAPHIC DESCRIPTION CHARACTER FULL SURROUND IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM ABOVE IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM BELOW IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM LEFT IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM UPPER LEFT IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM UPPER RIGHT IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM LOWER LEFT IDEOGRAPHIC DESCRIPTION CHARACTER OVERLAID
3038..303A, 303E	4	HANGZHOU NUMERAL TEN HANGZHOU NUMERAL TWENTY HANGZHOU NUMERAL THIRTY IDEOGRAPHIC VARIATION INDICATOR
31A0..31B7	24	Extended Bopomofo
3400..4DFF	6,582	CJK Unified Ideographs, Extension A
A000..A48C	1,165	Yi
A490..A4C1	50	Yi radicals
FB1D	1	HEBREW LETTER YOD WITH HIRIQ
FFF9..FFFB	3	INTERLINEAR ANNOTATION ANCHOR INTERLINEAR ANNOTATION SEPARATOR INTERLINEAR ANNOTATION TERMINATOR

Character Semantics Changes

Significant clarifications or modifications to character semantics include the following:

Bidirectional Properties. Bidirectional properties were made more consistent with the general category property, and new bidirectional properties were created. See *Section 3.12, Bidirectional Behavior.*

Byte Order Mark. The use of the byte order mark with transformation formats was clarified. See *Section 3.8, Transformations.*

Capital Letters with Iota Adscript. The representative glyphs, semantics, case mappings, and decompositions have been revised to make their handling more consistent.

Case. Case properties have been extended for those situations where there is a mapping to multiple characters and where case is locale-dependent. See the SpecialCasing.txt file on the CD-ROM.

Combining Classes. These classes were updated significantly to resolve problems of normalization and decomposition for Indic scripts in particular. See *Table 4-3, Combining Classes*.

Decompositions. Unicode character decompositions have been significantly updated to fix errors in the original assignments, to allow correct collation weighting, and to make decompositions consistent for normalization.

Eyelash Ra. Consonant RA rules have been updated and expanded. See rules R5 and R5a in *Section 9.1, Devanagari*.

Figure Space. U+2007 FIGURE SPACE is no longer treated like a numeric separator for purposes of bidirectional layout. See *Section 6.1, General Punctuation*, for a description of its tabular width characteristics.

General Category. A series of General Category changes were made to assist the convergence of the Unicode definition of identifier with ISO TR 10176.

Identifier Syntax. The list of recommended characters for computer language identifiers was corrected again, and the syntax for identifiers was further simplified. (See *Section 5.16, Identifiers*.)

Layout Controls. The description of layout controls was enhanced to include the behavior of U+00A0 NO-BREAK SPACE, U+00AD SOFT HYPHEN, and zero width spaces. See *Section 13.2, Layout Controls.*

Line and Paragraph Separators. Use of line and paragraph separators is clarified in *Section 3.12, Bidirectional Behavior*, and in *Section 13.1, Control Codes*.

Newlines. Line-handling characteristics have been documented more fully for Unicode environments. Discussions on CR and LF can be found in *Section 3.12, Bidirectional Behavior*; *Section 5.9, Line Handling*; *Section 13.1, Control Codes*; and in Unicode Technical Report #13, "Unicode Newline Guidelines," on the CD-ROM or the up-to-date version on the Unicode Web site.

Quotation Marks. Two new punctuation categories, Pi and Pf, were created for initial and final quotes. The use of these categories and language-based usage of quotation marks is clarified in *Section 6.1, General Punctuation*.

Script Capital p. The Weierstrass elliptic function symbol, U+2118 SCRIPT CAPITAL P, actually has the form of a *lowercase* calligraphic p despite the use of *capital* in its name.

Symmetric Swapping. The symmetric swapping value for guillemets was corrected.

Tibetan. Semantics of many Tibetan characters have been clarified or revised. See *Section 9.13, Tibetan*.

Tilde. The use of U+007E TILDE as a spacing clone of combining tilde and as a regular character is described in *Section 6.1, General Punctuation*.

Changes Affecting Conformance

Conformance clauses, definitions, and explanatory text were added for handling Unicode Transformation Formats. The Unicode bidirectional algorithm rules were clarified and expanded, and new bidirectional character properties were documented. Other normative

character property values were changed; see the Unicode Character Database found on the CD-ROM for details.

Unicode Technical Reports

- UTR #11: East Asian Width, Version 5.0
- UTR #13: Unicode Newline Guidelines, Version 5.0
- UTR #14: Line Breaking Properties, Version 6.0
- UTR #15: Unicode Normalization Forms, Version 18.0

This PDF file is an excerpt from *The Unicode Standard, Version 3.0*, issued by the Unicode Consortium and published by Addison-Wesley. The material has been modified slightly for this online edition, however the PDF files have not been modified to reflect the corrections found on the Updates and Errata page (see http://www.unicode.org/unicode/uni2errata/UnicodeErrata.html). More recent versions of the Unicode standard exist (see http://www.unicode.org/unicode/standard/versions/).

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

ISBN 0-201-61633-5

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

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

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

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

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

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

First printing, January 2000.