

This PDF file is an excerpt from The Unicode Standard, Version 5.2, issued and published by the Unicode Consortium. The PDF files have not been modified to reflect the corrections found on the Updates and Errata page (<http://www.unicode.org/errata/>). For information about more recent versions of the Unicode Standard see <http://www.unicode.org/versions/enumeratedversions.html>.

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and the publisher was aware of a trademark claim, the designations have been printed with initial capital letters or in all capitals.

The Unicode® Consortium is a registered trademark, and Unicode™ is a trademark of Unicode, Inc. The Unicode logo is a trademark of Unicode, Inc., and may be registered in some jurisdictions.

The authors and publisher have taken care in the preparation of this book, but make no expressed or implied warranty of any kind and assume no responsibility for errors or omissions. No liability is assumed for incidental or consequential damages in connection with or arising out of the use of the information or programs contained herein.

The *Unicode Character Database* and other files are provided as-is by Unicode®, Inc. No claims are made as to fitness for any particular purpose. No warranties of any kind are expressed or implied. The recipient agrees to determine applicability of information provided.

Copyright © 1991–2009 Unicode, Inc.

All rights reserved. This publication is protected by copyright, and permission must be obtained from the publisher prior to any prohibited reproduction, storage in a retrieval system, or transmission in any form or by any means, electronic, mechanical, photocopying, recording, or likewise. For information regarding permissions, inquire at <http://www.unicode.org/reporting.html>. For terms of use, please see <http://www.unicode.org/copyright.html>.

Visit the Unicode Consortium on the Web: <http://www.unicode.org>

The Unicode Standard / the Unicode Consortium ; edited by Julie D. Allen ... [et al.]. — Version 5.2.

Includes bibliographical references and index.

ISBN 978-1-936213-00-9 (<http://www.unicode.org/versions/Unicode5.2.0/>)

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

II. Unicode Consortium.

QA268.U545 2009

ISBN 978-1-936213-00-9

Published in Mountain View, CA

December 2009

Contents

Figures	xxi
Tables	xxv
Foreword	xxix
Preface	xxx
What's New	xxx
Organization of This Standard	xxxiii
Unicode Standard Annexes	xxxiv
The Unicode Character Database	xxxv
Unicode Code Charts	xxxv
Unicode Technical Standards and Unicode Technical Reports	xxxvi
Updates and Errata	xxxvi
1 Introduction	1
1.1 Coverage	2
Standards Coverage	2
New Characters	3
1.2 Design Goals	3
1.3 Text Handling	4
Text Elements	5
2 General Structure	7
2.1 Architectural Context	7
Basic Text Processes	8
Text Elements, Characters, and Text Processes	8
Text Processes and Encoding	9
2.2 Unicode Design Principles	10
Universality	10
Efficiency	11
Characters, Not Glyphs	11
Semantics	13
Plain Text	14
Logical Order	15
Unification	17
Dynamic Composition	18
Stability	18
Convertibility	19
2.3 Compatibility Characters	19
Compatibility Variants	20
Compatibility Decomposable Characters	20
2.4 Code Points and Characters	22
Types of Code Points	23
2.5 Encoding Forms	24
UTF-32	26
UTF-16	27
UTF-8	28
Comparison of the Advantages of UTF-32, UTF-16, and UTF-8	28
2.6 Encoding Schemes	30

2.7	Unicode Strings	32
2.8	Unicode Allocation	32
	Planes	33
	Allocation Areas and Character Blocks	34
	Assignment of Code Points	35
2.9	Details of Allocation	35
	Plane 0 (BMP)	37
	Plane 1 (SMP)	38
	Plane 2 (SIP)	39
	Other Planes	40
2.10	Writing Direction	40
2.11	Combining Characters	41
	Sequence of Base Characters and Diacritics	42
	Multiple Combining Characters	43
	Ligated Multiple Base Characters	45
	Exhibiting Nonspacing Marks in Isolation	45
	“Characters” and Grapheme Clusters	46
2.12	Equivalent Sequences and Normalization	46
	Normalization	47
	Decompositions	48
	Non-decomposition of Overlaid Diacritics	49
2.13	Special Characters and Noncharacters	50
	Special Noncharacter Code Points	50
	Byte Order Mark (BOM)	50
	Layout and Format Control Characters	51
	The Replacement Character	51
	Control Codes	51
2.14	Conforming to the Unicode Standard	51
	Characteristics of Conformant Implementations	52
	Unacceptable Behavior	52
	Acceptable Behavior	53
	Supported Subsets	53
3	Conformance	55
3.1	Versions of the Unicode Standard	55
	Stability	56
	Version Numbering	56
	Errata and Corrigenda	57
	References to the Unicode Standard	57
	Precision in Version Citation	57
	References to Unicode Character Properties	58
	References to Unicode Algorithms	58
3.2	Conformance Requirements	59
	Code Points Unassigned to Abstract Characters	59
	Interpretation	60
	Modification	60
	Character Encoding Forms	61
	Character Encoding Schemes	62
	Bidirectional Text	62
	Normalization Forms	62
	Normative References	63
	Unicode Algorithms	63
	Default Casing Algorithms	63

	Unicode Standard Annexes	63
3.3	Semantics	64
	Definitions	64
	Character Identity and Semantics	64
3.4	Characters and Encoding	66
3.5	Properties	69
	Types of Properties	69
	Property Values	70
	Classification of Properties by Their Values	71
	Property Status	72
	Context Dependence	75
	Stability of Properties	75
	Simple and Derived Properties	76
	Property Aliases	76
	Private Use	77
3.6	Combination	77
	Combining Character Sequences	77
	Grapheme Clusters	80
	Application of Combining Marks	81
3.7	Decomposition	85
	Compatibility Decomposition	85
	Canonical Decomposition	86
3.8	Surrogates	87
3.9	Unicode Encoding Forms	88
	UTF-32	92
	UTF-16	92
	UTF-8	92
	Encoding Form Conversion	94
	Constraints on Conversion Processes	94
3.10	Unicode Encoding Schemes	95
3.11	Normalization Forms	99
	Normalization Stability	99
	Combining Classes	100
	Specification of Unicode Normalization Forms	101
	Starters	101
	Canonical Ordering Algorithm	102
	Canonical Composition Algorithm	102
	Definition of Normalization Forms	104
3.12	Conjoining Jamo Behavior	104
	Definitions	105
	Hangul Syllable Boundary Determination	106
	Standard Korean Syllables	106
	Hangul Syllable Composition	108
	Hangul Syllable Decomposition	109
	Hangul Syllable Name Generation	109
3.13	Default Case Algorithms	110
	Definitions	110
	Default Case Conversion	111
	Default Case Detection	112
	Default Caseless Matching	113
4	Character Properties	115
4.1	Unicode Character Database	116

4.2	Case—Normative	118
	Definitions of Case and Casing	118
	Case Mapping	120
4.3	Combining Classes—Normative	120
	Reordrant, Split, and Subjoined Combining Marks	121
4.4	Directionality—Normative	124
4.5	General Category—Normative	124
4.6	Numeric Value—Normative	126
	Ideographic Numeric Values	126
4.7	Bidi Mirrored—Normative	127
4.8	Name—Normative	128
	Unicode Name Property	129
	Code Point Labels	131
	Use of Character Names in APIs and User Interfaces	131
4.9	Unicode 1.0 Names	132
4.10	Letters, Alphabetic, and Ideographic	132
4.11	Properties Related to Text Boundaries	133
4.12	Characters with Unusual Properties	133
5	Implementation Guidelines	137
5.1	Transcoding to Other Standards	137
	Issues	137
	Multistage Tables	138
5.2	Programming Languages and Data Types	139
	Unicode Data Types for C	139
5.3	Unknown and Missing Characters	140
	Reserved and Private-Use Character Codes	140
	Interpretable but Unrenderable Characters	141
	Default Property Values	141
	Default Ignorable Code Points	141
	Interacting with Downlevel Systems	141
5.4	Handling Surrogate Pairs in UTF-16	142
5.5	Handling Numbers	143
5.6	Normalization	144
5.7	Compression	145
5.8	Newline Guidelines	146
	Definitions	146
	Line Separator and Paragraph Separator	147
	Recommendations	148
5.9	Regular Expressions	149
5.10	Language Information in Plain Text	149
	Requirements for Language Tagging	149
	Language Tags and Han Unification	150
5.11	Editing and Selection	151
	Consistent Text Elements	151
5.12	Strategies for Handling Nonspacing Marks	152
	Keyboard Input	153
	Truncation	153
5.13	Rendering Nonspacing Marks	155
	Canonical Equivalence	157
	Positioning Methods	158

5.14	Locating Text Element Boundaries	159
5.15	Identifiers	160
5.16	Sorting and Searching	160
	Culturally Expected Sorting and Searching	160
	Language-Insensitive Sorting	161
	Searching	161
	Sublinear Searching	162
5.17	Binary Order	162
	UTF-8 in UTF-16 Order	163
	UTF-16 in UTF-8 Order	163
5.18	Case Mappings	164
	Titlecasing	165
	Complications for Case Mapping	165
	Reversibility	167
	Caseless Matching	168
	Normalization and Casing	169
5.19	Mapping Compatibility Variants	170
5.20	Unicode Security	171
5.21	Default Ignorable Code Points	173
6	Writing Systems and Punctuation	177
6.1	Writing Systems	178
6.2	General Punctuation	181
	Blocks Devoted to Punctuation	183
	Format Control Characters	183
	Space Characters	184
	Dashes and Hyphens	185
	Paired Punctuation	187
	Language-Based Usage of Quotation Marks	187
	Apostrophes	189
	Other Punctuation	190
	Archaic Punctuation and Editorial Marks	193
	Indic Punctuation	196
	CJK Punctuation	196
	Unknown or Unavailable Ideographs	197
	CJK Compatibility Forms	198
7	European Alphabetic Scripts	199
7.1	Latin	200
	Letters of Basic Latin: U+0041–U+007A	203
	Letters of the Latin-1 Supplement: U+00C0–U+00FF	203
	Latin Extended-A: U+0100–U+017F	203
	Latin Extended-B: U+0180–U+024F	204
	IPA Extensions: U+0250–U+02AF	205
	Phonetic Extensions: U+1D00–U+1DBF	207
	Latin Extended Additional: U+1E00–U+1EFF	208
	Latin Extended-C: U+2C60–U+2C7F	208
	Latin Extended-D: U+A720–U+A7FF	208
	Latin Ligatures: U+FB00–U+FB06	209
7.2	Greek	209
	Greek: U+0370–U+03FF	209
	Greek Extended: U+1F00–U+1FFF	213
	Ancient Greek Numbers: U+10140–U+1018F	214

7.3	Coptic	215
7.4	Cyrillic	217
	Cyrillic: U+0400–U+04FF	218
	Cyrillic Supplement: U+0500–U+052F	218
	Cyrillic Extended-A: U+2DE0–U+2DFF	219
	Cyrillic Extended-B: U+A640–U+A69F	219
7.5	Glagolitic	219
7.6	Armenian	220
7.7	Georgian	221
7.8	Modifier Letters	222
	Spacing Modifier Letters: U+02B0–U+02FF	224
	Modifier Tone Letters: U+A700–U+A71F	225
7.9	Combining Marks	225
	Combining Diacritical Marks: U+0300–U+036F	229
	Combining Diacritical Marks Supplement: U+1DC0–U+1DFF	229
	Combining Marks for Symbols: U+20D0–U+20FF	230
	Combining Half Marks: U+FE20–U+FE2F	230
	Combining Marks in Other Blocks	231
8	Middle Eastern Scripts	233
8.1	Hebrew	234
	Hebrew: U+0590–U+05FF	234
	Alphabetic Presentation Forms: U+FB1D–U+FB4F	238
8.2	Arabic	238
	Arabic: U+0600–U+06FF	238
	Arabic Cursive Joining	243
	Arabic Ligatures	245
	Arabic Supplement: U+0750–U+077F	249
	Arabic Presentation Forms-A: U+FB50–U+FDFF	249
	Arabic Presentation Forms-B: U+FE70–U+FEFF	250
8.3	Syriac	250
	Syriac: U+0700–U+074F	250
	Syriac Shaping	254
8.4	Samaritan	257
8.5	Thaana	259
9	South Asian Scripts-I	261
9.1	Devanagari	262
	Devanagari: U+0900–U+097F	262
	Principles of the Devanagari Script	263
	Rendering Devanagari	268
	Devanagari Extended: U+A8E0–U+A8FF	277
	Vedic Extensions: U+1CD0–U+1CFF	278
9.2	Bengali (Bangla)	279
9.3	Gurmukhi	283
9.4	Gujarati	286
9.5	Oriya	287
9.6	Tamil	289
	Tamil: U+0B80–U+0BFF	289
	Tamil Vowels	290
	Tamil Ligatures	291
	Tamil Named Character Sequences	294

9.7	Telugu	296
9.8	Kannada	297
	Kannada: U+0C80–U+0CFF	297
	Principles of the Kannada Script	297
	Rendering Kannada	299
9.9	Malayalam	299
10	South Asian Scripts-II	305
10.1	Sinhala	306
10.2	Tibetan	307
10.3	Lepcha	316
10.4	Phags-pa	318
10.5	Limbu	323
10.6	Syloti Nagri	326
10.7	Kaithi	327
10.8	Saurashtra	329
10.9	Meetei Mayek	330
10.10	Ol Chiki	331
10.11	Kharoshthi	332
	Kharoshthi: U+10A00–U+10A5F	332
	Rendering Kharoshthi	334
11	Southeast Asian Scripts	337
11.1	Thai	338
11.2	Lao	340
11.3	Myanmar	342
	Myanmar: U+1000–U+109F	342
	Myanmar Extended-A: U+AA60–U+AA7F	345
	Khamti Shan	346
	Aiton and Phake	347
11.4	Khmer	348
	Khmer: U+1780–U+17FF	348
	Principles of the Khmer Script	348
	Khmer Symbols: U+19E0–U+19FF	357
11.5	Tai Le	357
11.6	New Tai Lue	358
11.7	Tai Tham	359
11.8	Tai Viet	361
11.9	Kayah Li	363
11.10	Cham	364
11.11	Philippine Scripts	366
	Tagalog: U+1700–U+171F	366
	Hanunóo: U+1720–U+173F	366
	Buhid: U+1740–U+175F	366
	Tagbanwa: U+1760–U+177F	366
	Principles of the Philippine Scripts	366
11.12	Buginese	367
11.13	Balinese	368
11.14	Javanese	373
11.15	Rejang	375

11.16 Sundanese	376
12 East Asian Scripts	379
12.1 Han	380
CJK Unified Ideographs.....	380
CJK Standards.....	381
Blocks Containing Han Ideographs.....	383
General Characteristics of Han Ideographs.....	384
Principles of Han Unification.....	388
Unification Rules.....	389
Abstract Shape.....	390
Han Ideograph Arrangement.....	392
Mappings for Han Ideographs.....	393
CJK Unified Ideographs Extension B: U+20000–U+2A6D6.....	393
CJK Unified Ideographs Extension C: U+2A700–U+2B734.....	393
CJK Compatibility Ideographs: U+F900–U+FAFF.....	393
CJK Compatibility Supplement: U+2F800–U+2FA1D.....	394
Kanbun: U+3190–U+319F.....	394
Symbols Derived from Han Ideographs.....	394
CJK and KangXi Radicals: U+2E80–U+2FD5.....	394
CJK Additions from HKSCS and GB 18030.....	395
CJK Strokes: U+31C0–U+31EF.....	396
12.2 Ideographic Description Characters	396
12.3 Bopomofo	399
12.4 Hiragana and Katakana	401
Hiragana: U+3040–U+309F.....	401
Katakana: U+30A0–U+30FF.....	402
Katakana Phonetic Extensions: U+31F0–U+31FF.....	402
12.5 Halfwidth and Fullwidth Forms	402
12.6 Hangul	403
Hangul Jamo: U+1100–U+11FF.....	403
Hangul Jamo Extended-A: U+A960–U+A97F.....	404
Hangul Jamo Extended-B: U+D7B0–U+D7FF.....	404
Hangul Compatibility Jamo: U+3130–U+318F.....	404
Hangul Syllables: U+AC00–U+D7A3.....	404
12.7 Yi	406
13 Additional Modern Scripts	409
13.1 Ethiopic	410
Ethiopic: U+1200–U+137F.....	410
Ethiopic Extensions: U+1380–U+139F, U+2D80–U+2DDF.....	412
13.2 Mongolian	412
13.3 Osmanya	419
13.4 Tifinagh	420
13.5 N’Ko	421
13.6 Vai	425
13.7 Bamum	427
13.8 Cherokee	427
13.9 Canadian Aboriginal Syllabics	428
Canadian Aboriginal Syllabics: U+1400–U+167F.....	428
Canadian Aboriginal Syllabics Extended: U+18B0–U+18FF.....	429
13.10 Deseret	429

13.11 Shavian	431
13.12 Lisu	432
14 Ancient and Historic Scripts	435
14.1 Ogham	436
14.2 Old Italic	437
14.3 Runic	439
14.4 Gothic	441
14.5 Old Turkic	442
14.6 Linear B	442
Linear B Syllabary: U+10000–U+1007F	442
Linear B Ideograms: U+10080–U+100FF	443
Aegean Numbers: U+10100–U+1013F	443
14.7 Cypriot Syllabary	443
14.8 Ancient Anatolian Alphabets	444
Lycian: U+10280–U+1029F	444
Carian: U+102A0–U+102DF	444
Lydian: U+10920–U+1093F	444
14.9 Old South Arabian	445
14.10 Phoenician	447
14.11 Imperial Aramaic	448
14.12 Inscriptional Parthian and Inscriptional Pahlavi	449
14.13 Avestan	450
14.14 Ugaritic	451
14.15 Old Persian	452
14.16 Sumero-Akkadian	452
Cuneiform: U+12000–U+123FF	452
Cuneiform Numbers and Punctuation: U+12400–U+1247F	455
14.17 Egyptian Hieroglyphs	455
15 Symbols	459
15.1 Currency Symbols	460
15.2 Letterlike Symbols	462
Letterlike Symbols: U+2100–U+214F	462
Mathematical Alphanumeric Symbols: U+1D400–U+1D7FF	463
Mathematical Alphabets	463
Fonts Used for Mathematical Alphabets	466
15.3 Number Forms	467
Number Forms: U+2150–U+218F	467
Common Indic Number Forms: U+A830–U+A83F	468
Rumi Numeral Forms: U+10E60–U+10E7E	468
CJK Number Forms	469
Superscripts and Subscripts: U+2070–U+209F	470
15.4 Mathematical Symbols	471
Mathematical Operators: U+2200–U+22FF	471
Supplements to Mathematical Symbols and Arrows	473
Supplemental Mathematical Operators: U+2A00–U+2AFF	473
Miscellaneous Mathematical Symbols-A: U+27C0–U+27EF	473
Miscellaneous Mathematical Symbols-B: U+2980–U+29FF	474
Miscellaneous Symbols and Arrows: U+2B00–U+2B7F	474
Arrows: U+2190–U+21FF	474
Supplemental Arrows	475

	Standardized Variants of Mathematical Symbols	475
15.5	Invisible Mathematical Operators	476
15.6	Technical Symbols	476
	Control Pictures: U+2400–U+243F	476
	Miscellaneous Technical: U+2300–U+23FF	477
	Optical Character Recognition: U+2440–U+245F	480
15.7	Geometrical Symbols	480
	Box Drawing and Box Elements	480
	Geometric Shapes: U+25A0–U+25FF	481
15.8	Miscellaneous Symbols and Dingbats	482
	Miscellaneous Symbols: U+2600–U+26FF	482
	Dingbats: U+2700–U+27BF	483
	Mahjong Tiles: U+1F000–U+1F02F	484
	Domino Tiles: U+1F030–U+1F09F	485
	Yijing Hexagram Symbols: U+4DC0–U+4DFF	485
	Tai Xuan Jing Symbols: U+1D300–U+1D356	485
	Ancient Symbols: U+10190–U+101CF	486
	Phaistos Disc Symbols: U+101D0–U+101FF	486
15.9	Enclosed and Square	487
	Enclosed Alphanumerics: U+2460–U+24FF	489
	Enclosed CJK Letters and Months: U+3200–U+32FF	489
	CJK Compatibility: U+3300–U+33FF	489
	Enclosed Alphanumeric Supplement: U+1F100–U+1F1FF	490
	Enclosed Ideographic Supplement: U+1F200–U+1F2FF	490
15.10	Braille	490
15.11	Western Musical Symbols	492
15.12	Byzantine Musical Symbols	496
15.13	Ancient Greek Musical Notation	496
16	Special Areas and Format Characters	499
16.1	Control Codes	500
	Representing Control Sequences	500
	Specification of Control Code Semantics	501
16.2	Layout Controls	501
	Line and Word Breaking	502
	Cursive Connection and Ligatures	503
	Combining Grapheme Joiner	507
	Bidirectional Ordering Controls	508
16.3	Deprecated Format Characters	509
16.4	Variation Selectors	510
16.5	Private-Use Characters	511
	Private Use Area: U+E000–U+F8FF	512
	Supplementary Private Use Areas	513
16.6	Surrogates Area	513
16.7	Noncharacters	514
16.8	Specials	515
	Byte Order Mark (BOM): U+FEFF	515
	Specials: U+FFF0–U+FFF8	517
	Annotation Characters: U+FFF9–U+FFFB	517
	Replacement Characters: U+FFFC–U+FFFD	518
16.9	Deprecated Tag Characters	519
	Deprecated Tag Characters: U+E0000–U+E007F	519

	Syntax for Embedding Tags	519
	Working with Language Tags	521
	Unicode Conformance Issues	522
	Formal Tag Syntax	522
17	Code Charts	525
17.1	Character Names List	525
	Images in the Code Charts and Character Lists	526
	Special Characters and Code Points	527
	Character Names	528
	Informative Aliases	528
	Normative Aliases	529
	Cross References	529
	Information About Languages	529
	Case Mappings	529
	Decompositions	530
	Subheads	531
17.2	CJK Unified Ideographs	531
17.3	Hangul Syllables	533
18	Han Radical-Stroke Index	535
A	Notational Conventions	537
	Code Points	537
	Character Names	537
	Character Blocks	537
	Sequences	538
	Rendering	538
	Properties and Property Values	538
	Miscellaneous	538
	Extended BNF	539
	Operators	540
B	Unicode Publications and Resources	541
B.1	The Unicode Consortium	541
	The Unicode Technical Committee	541
	Other Activities	542
B.2	Unicode Publications	542
B.3	Unicode Technical Standards	542
	UTS #6: A Standard Compression Scheme for Unicode	542
	UTS #10: Unicode Collation Algorithm	543
	UTS #18: Unicode Regular Expressions	543
	UTS #22: Character Mapping Markup Language (CharMapML)	543
	UTS #35: Unicode Locale Data Markup Language (LDML)	543
	UTS #37: Unicode Ideographic Variation Database	543
	UTS #39: Unicode Security Mechanisms	543
B.4	Unicode Technical Reports	543
	UTR #16: UTF-EBCDIC	543
	UTR #17: Unicode Character Encoding Model	543
	UTR #20: Unicode in XML and Other Markup Languages	543
	UTR #23: The Unicode Character Property Model	544
	UTR #25: Unicode Support for Mathematics	544
	UTR #26: Compatibility Encoding Scheme for UTF-16: 8-Bit (CESU-8) ..	544
	UTR #33: Unicode Conformance Model	544
	UTR #36: Unicode Security Considerations	544
	UTR #45: U-Source Ideographs	544

B.5	Unicode Technical Notes	544
B.6	Other Unicode Online Resources	545
	Unicode Online Resources	545
	How to Contact the Unicode Consortium	546
C	Relationship to ISO/IEC 10646	549
C.1	History	549
C.2	Encoding Forms in ISO/IEC 10646	552
	Zero Extending	553
C.3	UCS Transformation Formats	553
	UTF-8	553
	UTF-16	554
C.4	Synchronization of the Standards	554
C.5	Identification of Features for the Unicode Standard	554
C.6	Character Names	555
C.7	Character Functional Specifications	555
D	Changes from Previous Versions	557
D.1	Versions of the Unicode Standard	557
D.2	Clause and Definition Updates	559
D.3	Changes from Version 5.1 to Version 5.2	560
D.4	Changes from Version 5.0 to Version 5.1	560
E	Han Unification History	565
E.1	Development of the URO	565
E.2	Ideographic Rapporteur Group	566
R	References	569
R.1	Source Standards and Specifications	569
R.2	Source Dictionaries for Han Unification	576
R.3	Other Sources for the Unicode Standard	576
R.4	Selected Resources: Technical	589
R.5	Selected Resources: Scripts and Languages	591
I	General Index	595