The Unicode Standard
Version 8.0 – 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 © 1991–2015 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 8.0
Includes bibliographical references and index.
ISBN 978-1-936213-10-8 (http://www.unicode.org/versions/Unicode8.0.0/)
1. Unicode (Computer character set) I. Allen, Julie D. II. Unicode Consortium.
QA268.U545 2015

ISBN 978-1-936213-10-8
Published in Mountain View, CA
August 2015

Contents

Figures xxi
Tables xxv
Preface xxxi
Why Unicode? xxxi
What's New? xxxi
Organization of This Standard xxxii
Unicode Standard Annexes xxxiii
The Unicode Character Database xxxv
Unicode Code Charts xxxv
Unicode Technical Standards and Unicode Technical Reports xxxv
Updates and Errata xxxvi
Acknowledgements xxxvi
1 Introduction 1
1.1 Coverage 3
Standards Coverage 3
New Characters 3
1.2 Design Goals 4
1.3 Text Handling 6
Characters and Glyphs 6
Text Elements 6
2 General Structure 9
2.1 Architectural Context 10
Basic Text Processes 10
Text Elements, Characters, and Text Processes 10
Text Processes and Encoding 11
2.2 Unicode Design Principles 14
Universality 14
Efficiency 15
Characters, Not Glyphs 15
Semantics 18
Plain Text 18
Logical Order 19
Unification 21
Dynamic Composition 22
Stability 23
Convertibility 23

2.3 Compatibility Characters . . . 25
Compatibility Variants . . . 26
Compatibility Decomposable Characters . . . 26
2.4 Code Points and Characters . . . 29
Types of Code Points . . . 30
2.5 Encoding Forms . . . 33
UTF-32 . . . 35
UTF-16 . . . 36
UTF-8 . . . 36
Comparison of the Advantages of UTF-32, UTF-16, and UTF-8 . . . 38
2.6 Encoding Schemes . . . 40
2.7 Unicode Strings . . . 43
2.8 Unicode Allocation . . . 44
Planes . . . 44
Allocation Areas and Character Blocks . . . 45
Assignment of Code Points . . . 46
2.9 Details of Allocation . . . 48
Plane 0 (BMP) . . . 49
Plane 1 (SMP) . . . 51
Plane 2 (SIP) . . . 52
Other Planes . . . 52
2.10 Writing Direction . . . 53
2.11 Combining Characters . . . 55
Sequence of Base Characters and Diacritics . . . 56
Multiple Combining Characters . . . 57
Ligated Multiple Base Characters . . . 59
Exhibiting Nonspacing Marks in Isolation . . . 60
"Characters" and Grapheme Clusters . . . 60
2.12 Equivalent Sequences . . . 62
Normalization . . . 62
Decompositions . . . 63
Non-decomposition of Certain Diacritics . . . 65
2.13 Special Characters . . . 67
Special Noncharacter Code Points . . . 67
Byte Order Mark (BOM) . . . 67
Layout and Format Control Characters . . . 68
The Replacement Character . . . 68
Control Codes . . . 68
2.14 Conforming to the Unicode Standard . . . 69
Characteristics of Conformant Implementations . . . 69
Unacceptable Behavior . . . 70
Acceptable Behavior . . . 70
Supported Subsets . . . 71

3 Conformance 73

3.1 Versions of the Unicode Standard 74

Stability 74
Version Numbering 75
Errata and Corrigenda 76
References to the Unicode Standard 76
Precision in Version Citation 77
References to Unicode Character Properties 77
References to Unicode Algorithms 78

3.2 Conformance Requirements 79

Code Points Unassigned to Abstract Characters 79
Interpretation 80
Modification 81
Character Encoding Forms 82
Character Encoding Schemes 83
Bidirectional Text 84
Normalization Forms 84
Normative References 84
Unicode Algorithms 84
Default Casing Algorithms 85
Unicode Standard Annexes 85

3.3 Semantics 87

Definitions 87
Character Identity and Semantics 87

3.4 Characters and Encoding 90

3.5 Properties 94

Types of Properties 94
Property Values 95
Default Property Values 96
Classification of Properties by Their Values 97
Property Status 98
Context Dependence 101
Stability of Properties 101
Simple and Derived Properties 103
Property Aliases 103
Private Use 104

3.6 Combination 105

Combining Character Sequences 105
Grapheme Clusters 107
Application of Combining Marks 109

3.7 Decomposition 115

Compatibility Decomposition 115
Canonical Decomposition 116

3.8 Surrogates ... 118
3.9 Unicode Encoding Forms ... 119
UTF-32 ... 123
UTF-16 ... 124
UTF-8 ... 124
Encoding Form Conversion ... 126
Constraints on Conversion Processes ... 126
3.10 Unicode Encoding Schemes ... 130
3.11 Normalization Forms ... 134
Normalization Stability ... 134
Combining Classes ... 135
Specification of Unicode Normalization Forms ... 136
Starters ... 136
Canonical Ordering Algorithm ... 137
Canonical Composition Algorithm ... 138
Definition of Normalization Forms ... 140
3.12 Conjoining Jamo Behavior ... 142
Definitions ... 142
Hangul Syllable Decomposition ... 144
Hangul Syllable Composition ... 146
Hangul Syllable Name Generation ... 147
Sample Code for Hangul Algorithms ... 148
3.13 Default Case Algorithms ... 152
Definitions ... 152
Default Case Conversion ... 154
Default Case Folding ... 155
Default Case Detection ... 156
Default Caseless Matching ... 157
4 Character Properties ... 159
4.1 Unicode Character Database ... 161
4.2 Case ... 164
Definitions of Case and Casing ... 164
Case Mapping ... 166
4.3 Combining Classes ... 168
Reordrant, Split, and Subjoined Combining Marks ... 168
4.4 Directionality ... 173
4.5 General Category ... 174
4.6 Numeric Value ... 177
Ideographic Numeric Values ... 178
4.7 Bidi Mirrored ... 180

4.8 Name ... 182
Unicode Name Property ... 185
Code Point Labels ... 186
Use of Character Names in APIs and User Interfaces ... 187
4.9 Unicode 1.0 Names ... 188
4.10 Letters, Alphabetic, and Ideographic ... 189
4.11 Properties Related to Text Boundaries ... 190
4.12 Characters with Unusual Properties ... 191
5 Implementation Guidelines ... 195
5.1 Data Structures for Character Conversion ... 196
Issues ... 196
Multistage Tables ... 196
5.2 Programming Languages and Data Types ... 199
Unicode Data Types for C ... 199
5.3 Unknown and Missing Characters ... 201
5.4 Handling Surrogate Pairs in UTF-16 ... 203
5.5 Handling Numbers ... 205
5.6 Normalization ... 206
5.7 Compression ... 208
5.8 Newline Guidelines ... 209
Definitions ... 209
Line Separator and Paragraph Separator ... 210
Recommendations ... 211
5.9 Regular Expressions ... 214
5.10 Language Information in Plain Text ... 215
Requirements for Language Tagging ... 215
Language Tags and Han Unification ... 215
5.11 Editing and Selection ... 217
Consistent Text Elements ... 217
5.12 Strategies for Handling Nonspacing Marks ... 219
Keyboard Input ... 220
Truncation ... 220
5.13 Rendering Nonspacing Marks ... 222
Canonical Equivalence ... 225
Positioning Methods ... 226
5.14 Locating Text Element Boundaries ... 228
5.15 Identifiers ... 229
5.16 Sorting and Searching ... 230
Culturally Expected Sorting and Searching ... 230
Language-Insensitive Sorting ... 230

Searching . . . 231
Sublinear Searching . . . 231
5.17 Binary Order . . . 233
UTF-8 in UTF-16 Order . . . 233
UTF-16 in UTF-8 Order . . . 234
5.18 Case Mappings . . . 236
Titlecasing . . . 236
Complications for Case Mapping . . . 237
Reversibility . . . 239
Caseless Matching . . . 240
Normalization and Casing . . . 242
5.19 Mapping Compatibility Variants . . . 244
5.20 Unicode Security . . . 246
5.21 Ignoring Characters in Processing . . . 249
Characters Ignored in Text Segmentation . . . 249
Characters Ignored in Line Breaking . . . 250
Characters Ignored in Cursive Joining . . . 250
Characters Ignored in Identifiers . . . 250
Characters Ignored in Searching and Sorting . . . 251
Characters Ignored for Display . . . 251
5.22 Best Practice for U+FFFD Substitution . . . 255
6 Writing Systems and Punctuation . . . 257
6.1 Writing Systems . . . 258
6.2 General Punctuation . . . 263
Blocks Devoted to Punctuation . . . 264
Format Control Characters . . . 265
Space Characters . . . 266
Dashes and Hyphens . . . 267
Paired Punctuation . . . 269
Language-Based Usage of Quotation Marks . . . 270
Apostrophes . . . 274
Other Punctuation . . . 274
Archaic Punctuation and Editorial Marks . . . 279
Indic Punctuation . . . 282
CJK Punctuation . . . 284
Unknown or Unavailable Ideographs . . . 285
CJK Compatibility Forms . . . 286
7 Europe-I . . . 289
Modern and Liturgical Scripts
7.1 Latin . . . 291
Letters of Basic Latin: U+0041–U+007A . . . 295
Letters of the Latin-1 Supplement: U+00C0–U+00FF . . . 295
Latin Extended-A: U+0100–U+017F . . . 295

Latin Extended-B: U+0180–U+024F 296
IPA Extensions: U+0250–U+02AF 298
Phonetic Extensions: U+1D00–U+1DBF 300
Latin Extended Additional: U+1E00–U+1EFF 301
Latin Extended-C: U+2C60–U+2C7F 301
Latin Extended-D: U+A720–U+A7FF 302
Latin Extended-E: U+AB30–U+AB6F 303
Latin Ligatures: U+FB00–U+FB06 303
7.2 Greek 304
Greek: U+0370–U+03FF 304
Greek Extended: U+1F00–U+1FFF 308
Ancient Greek Numbers: U+10140–U+1018F 309
7.3 Coptic 311
7.4 Cyrillic 314
Cyrillic: U+0400–U+04FF 314
Cyrillic Supplement: U+0500–U+052F 315
Cyrillic Extended-A: U+2DE0–U+2DFF 315
Cyrillic Extended-B: U+A640–U+A69F 316
7.5 Glagolitic 318
7.6 Armenian 319
7.7 Georgian 321
7.8 Modifier Letters 323
Spacing Modifier Letters: U+02B0–U+02FF 324
Modifier Tone Letters: U+A700–U+A71F 326
7.9 Combining Marks 327
Combining Diacritical Marks: U+0300–U+036F 332
Combining Diacritical Marks Extended: U+1AB0–U+1AFF 333
Combining Diacritical Marks Supplement: U+1DC0–U+1DFF 333
Combining Marks for Symbols: U+20D0–U+20FF 334
Combining Half Marks: U+FE20–U+FE2F 335
Combining Marks in Other Blocks 335
8 Europe-II 337
Ancient and Other Scripts
8.1 Linear A 339
8.2 Linear B 340
Linear B Syllabary: U+10000–U+1007F 340
Linear B Ideograms: U+10080–U+100FF 340
Aegean Numbers: U+10100–U+1013F 340
8.3 Cypriot Syllabary 342

8.4 Ancient Anatolian Alphabets . . . 343
Lycian: U+10280–U+1029F . . . 343
Carian: U+102A0–U+102DF . . . 343
Lydian: U+10920–U+1093F . . . 343
8.5 Old Italic . . . 345
8.6 Runic . . . 348
8.7 Old Hungarian . . . 351
8.8 Gothic . . . 352
8.9 Elbasan . . . 353
8.10 Caucasian Albanian . . . 354
8.11 Old Permic . . . 355
8.12 Ogham . . . 356
8.13 Shavian . . . 357
9 Middle East-I . . . 359
Modern and Liturgical Scripts
9.1 Hebrew . . . 361
Hebrew: U+0590–U+05FF . . . 361
Alphabetic Presentation Forms: U+FB1D–U+FB4F . . . 365
9.2 Arabic . . . 367
Arabic: U+0600–U+06FF . . . 367
Arabic Cursive Joining . . . 374
Arabic Ligatures . . . 377
Arabic Joining Groups . . . 379
Arabic Supplement: U+0750–U+077F . . . 386
Arabic Extended-A: U+08A0–U+08FF . . . 386
Arabic Presentation Forms-A: U+FB50–U+FDFF . . . 386
Arabic Presentation Forms-B: U+FE70–U+FEFF . . . 388
9.3 Syriac . . . 389
Syriac: U+0700–U+074F . . . 389
Syriac Shaping . . . 393
9.4 Samaritan . . . 398
9.5 Mandaic . . . 400
10 Middle East-II . . . 403
Ancient Scripts
10.1 Old North Arabian . . . 405
10.2 Old South Arabian . . . 406
10.3 Phoenician . . . 408
10.4 Imperial Aramaic . . . 410
10.5 Manichaean . . . 412

10.6 Pahlavi and Parthian 416
Inscriptional Parthian: U+10B40–U+10B5F 416
Inscriptional Pahlavi: U+10B60–U+10B7F 416
Psalter Pahlavi: U+10B80–U+10BAF 417
10.7 Avestan 418
10.8 Nabataean 420
10.9 Palmyrene 421
10.10 Hatran 422
11 Cuneiform and Hieroglyphs 423
11.1 Sumero-Akkadian 424
Cuneiform: U+12000–U+123FF 424
Cuneiform Numbers and Punctuation: U+12400–U+1247F 427
Early Dynastic Cuneiform: U+12480–U+1254F 427
11.2 Ugaritic 428
11.3 Old Persian 429
11.4 Egyptian Hieroglyphs 430
11.5 Meroitic 434
11.6 Anatolian Hieroglyphs 436
12 South and Central Asia-I 439
Official Scripts of India
12.1 Devanagari 441
Devanagari: U+0900–U+097F 441
Principles of the Devanagari Script 442
Rendering Devanagari 448
Devanagari Digits, Punctuation, and Symbols 457
Extensions in the Main Devanagari Block 458
Devanagari Extended: U+A8E0–U+A8FF 460
Vedic Extensions: U+1CD0–U+1CFF 461
12.2 Bengali (Bangla) 463
12.3 Gurmukhi 469
12.4 Gujarati 474
12.5 Oriya (Odia) 476
12.6 Tamil 479
Tamil: U+0B80–U+0BFF 479
Tamil Vowels 480
Tamil Ligatures 481
Tamil Named Character Sequences 485
12.7 Telugu 488
12.8 Kannada 491
Kannada: U+0C80–U+0CFF 491

Principles of the Kannada Script 491
Rendering Kannada 493
12.9 Malayalam 495
Malayalam Orthographic Reform 495
Rendering Malayalam 496
Malayalam Numbers and Punctuation 501
13 South and Central Asia-II 503
Other Modern Scripts
13.1 Thaana 505
13.2 Sinhala 507
Sinhala Archaic Numbers: U+111E0–U+111FF 508
13.3 Tibetan 509
13.4 Mongolian 521
13.5 Limbu 530
13.6 Meetei Mayek 534
13.7 Mro 536
13.8 Warang Citi 537
13.9 Ol Chiki 538
13.10 Chakma 540
13.11 Lepcha 541
13.12 Saurashtra 544
14 South and Central Asia-III 545
Ancient Scripts
14.1 Brahmi 546
14.2 Kharoshthi 550
Kharoshthi: U+10A00–U+10A5F 550
Rendering Kharoshthi 552
14.3 Phags-pa 556
14.4 Old Turkic 563
15 South and Central Asia-IV 565
Other Historic Scripts
15.1 Syloti Nagri 567
15.2 Kaithi 569
15.3 Sharada 572
15.4 Takri 574
15.5 Siddham 576
15.6 Mahajani 578
15.7 Khojki 580
15.8 Khudawadi 582

15.9 Multani 584
15.10 Tirhuta 585
15.11 Modi 588
15.12 Grantha 591
Rendering Behavior 591
15.13 Ahom 594
15.14 Sora Sompeng 596
16 Southeast Asia 597
16.1 Thai 599
16.2 Lao 603
16.3 Myanmar 606
Myanmar: U+1000–U+109F 606
Myanmar Extended-A: U+AA60–U+AA7F 610
Khamti Shan 610
Aiton and Phake 612
Myanmar Extended-B: U+A9E0–U+A9FF 612
16.4 Khmer 613
Khmer: U+1780–U+17FF 613
Principles of the Khmer Script 613
Khmer Symbols: U+19E0–U+19FF 623
16.5 Tai Le 624
16.6 New Tai Lue 626
16.7 Tai Tham 629
16.8 Tai Viet 632
16.9 Kayah Li 635
16.10 Cham 636
16.11 Pahawh Hmong 638
16.12 Pau Cin Hau 640
17 Indonesia and Oceania 641
17.1 Philippine Scripts 642
Tagalog: U+1700–U+171F 642
Hanunóo: U+1720–U+173F 642
Buhid: U+1740–U+175F 642
Tagbanwa: U+1760–U+177F 642
Principles of the Philippine Scripts 642
17.2 Buginese 645
17.3 Balinese 647
17.4 Javanese 653
17.5 Rejang 657
17.6 Batak 658

17.7 Sundanese . . . 659
Sundanese: U+1B80–U+1BBF . . . 659
Sundanese Supplement: U+1CC0–U+1CCF . . . 660
18 East Asia . . . 661
18.1 Han . . . 663
CJK Unified Ideographs . . . 663
Blocks Containing Han Ideographs . . . 664
General Characteristics of Han Ideographs . . . 666
Principles of Han Unification . . . 670
Unification Rules . . . 671
Abstract Shape . . . 672
Han Ideograph Arrangement . . . 674
Radical-Stroke Indices . . . 675
Mappings for Han Ideographs . . . 676
CJK Unified Ideographs Extension B: U+20000–U+2A6D6 . . . 677
CJK Unified Ideographs Extension C: U+2A700–U+2B734 . . . 677
CJK Unified Ideographs Extension D: U+2B740–U+2B81D . . . 677
CJK Unified Ideographs Extension E: U+2B820–U+2CEA1 . . . 677
CJK Compatibility Ideographs: U+F900–U+FAFF . . . 677
CJK Compatibility Supplement: U+2F800–U+2FA1D . . . 678
Kanbun: U+3190–U+319F . . . 678
Symbols Derived from Han Ideographs . . . 679
CJK and KangXi Radicals: U+2E80–U+2FD5 . . . 679
CJK Additions from HKSCS and GB 18030 . . . 680
CJK Strokes: U+31C0–U+31EF . . . 680
18.2 Ideographic Description Characters . . . 681
18.3 Bopomofo . . . 685
18.4 Hiragana and Katakana . . . 688
Hiragana: U+3040–U+309F . . . 688
Katakana: U+30A0–U+30FF . . . 688
Katakana Phonetic Extensions: U+31F0–U+31FF . . . 689
Kana Supplement U+1B000–U+1B0FF . . . 689
18.5 Halfwidth and Fullwidth Forms . . . 690
18.6 Hangul . . . 691
Hangul Jamo: U+1100–U+11FF . . . 691
Hangul Jamo Extended-A: U+A960–U+A97F . . . 692
Hangul Jamo Extended-B: U+D7B0–U+D7FF . . . 692
Hangul Compatibility Jamo: U+3130–U+318F . . . 692
Hangul Syllables: U+AC00–U+D7A3 . . . 693
18.7 Yi . . . 695
18.8 Lisu . . . 698
18.9 Miao . . . 701

19 Africa 703
19.1 Ethiopic 704
Ethiopic: U+1200–U+137F 704
Ethiopic Extensions 706
19.2 Osmanya 708
19.3 Tifinagh 709
19.4 N'Ko 712
19.5 Vai 717
19.6 Bamum 719
Bamum: U+A6A0–U+A6FF 719
Bamum Supplement: U+16800–U+16A3F 719
19.7 Bassa Vah 721
19.8 Mende Kikakui 722
20 Americas 725
20.1 Cherokee 726
20.2 Canadian Aboriginal Syllabics 728
Canadian Aboriginal Syllabics: U+1400–U+167F 728
Canadian Aboriginal Syllabics Extended: U+18B0–U+18FF 729
20.3 Deseret 730
21 Notational Systems 733
21.1 Braille 734
21.2 Western Musical Symbols 736
21.3 Byzantine Musical Symbols 741
21.4 Ancient Greek Musical Notation 742
21.5 Duployan 745
21.6 Sutton SignWriting 747
22 Symbols 749
22.1 Currency Symbols 751
22.2 Letterlike Symbols 754
Letterlike Symbols: U+2100–U+214F 754
Mathematical Alphanumeric Symbols: U+1D400–U+1D7FF 755
Mathematical Alphabets 756
Fonts Used for Mathematical Alphabets 758
Arabic Mathematical Alphabetic Symbols: U+1EE00–U+1EEFF 760
22.3 Numerals 761
Decimal Digits 761
Other Digits 763
Non-Decimal Radix Systems 765
Acrophonic Systems and Other Letter-based Numbers 766
Coptic Epact Numbers: U+102E0–U+102FF 767
Rumi Numeral Forms: U+10E60–U+10E7E 768

CJK Numerals 768
Fractions 770
Common Indic Number Forms: U+A830–U+A83F 770
22.4 Superscript and Subscript Symbols 772
Superscripts and Subscripts: U+2070–U+209F 772
22.5 Mathematical Symbols 774
Mathematical Operators: U+2200–U+22FF 775
Supplements to Mathematical Symbols and Arrows 778
Supplemental Mathematical Operators: U+2A00–U+2AFF 778
Miscellaneous Mathematical Symbols-A: U+27C0–U+27EF 778
Miscellaneous Mathematical Symbols-B: U+2980–U+29FF 779
Miscellaneous Symbols and Arrows: U+2B00–U+2B7F 779
Arrows: U+2190–U+21FF 779
Supplemental Arrows 780
Standardized Variants of Mathematical Symbols 781
22.6 Invisible Mathematical Operators 782
22.7 Technical Symbols 783
Control Pictures: U+2400–U+243F 783
Miscellaneous Technical: U+2300–U+23FF 783
Optical Character Recognition: U+2440–U+245F 787
22.8 Geometrical Symbols 788
Box Drawing and Block Elements 788
Geometric Shapes: U+25A0–U+25FF 789
Geometric Shapes Extended: U+1F780–U+1F7FF 790
22.9 Miscellaneous Symbols 791
Miscellaneous Symbols and Pictographs 792
Emoticons: U+1F600–U+1F64F 795
Transport and Map Symbols: U+1F680–U+1F6FF 795
Dingbats: U+2700–U+27BF 795
Ornamental Dingbats: U+1F650–U+1F67F 797
Alchemical Symbols: U+1F700–U+1F77F 797
Mahjong Tiles: U+1F000–U+1F02F 798
Domino Tiles: U+1F030–U+1F09F 798
Playing Cards: U+1F0A0–U+1F0FF 798
Yijing Hexagram Symbols: U+4DC0–U+4DFF 799
Tai Xuan Jing Symbols: U+1D300–U+1D356 800
Ancient Symbols: U+10190–U+101CF 800
Phaistos Disc Symbols: U+101D0–U+101FF 801
22.10 Enclosed and Square 802
Enclosed Alphanumerics: U+2460–U+24FF 804
Enclosed CJK Letters and Months: U+3200–U+32FF 804
CJK Compatibility: U+3300–U+33FF 805
Enclosed Alphanumeric Supplement: U+1F100–U+1F1FF 805
Enclosed Ideographic Supplement: U+1F200–U+1F2FF 806

23 Special Areas and Format Characters 807
23.1 Control Codes 808
Representing Control Sequences 808
Specification of Control Code Semantics 809
23.2 Layout Controls 811
Line and Word Breaking 811
Cursive Connection and Ligatures 813
Combining Grapheme Joiner 817
Bidirectional Ordering Controls 819
Stateful Format Controls 820
23.3 Deprecated Format Characters 822
23.4 Variation Selectors 824
23.5 Private-Use Characters 827
Private Use Area: U+E000–U+F8FF 828
Supplementary Private Use Areas 829
23.6 Surrogates Area 830
23.7 Noncharacters 831
23.8 Specials 833
Byte Order Mark (BOM): U+FEFF 833
Specials: U+FFF0–U+FFF8 835
Annotation Characters: U+FFF9–U+FFFB 835
Replacement Characters: U+FFFC–U+FFFD 837
23.9 Tag Characters 838
Tag Characters: U+E0000–U+E007F 838
Deprecated Use for Language Tagging 838
Syntax for Embedding Tags 838
Working with Language Tags 840
Unicode Conformance Issues 842
Formal Tag Syntax 842
24 About the Code Charts 845
24.1 Character Names List 846
Images in the Code Charts and Character Lists 846
Special Characters and Code Points 848
Character Names 850
Informative Aliases 850
Normative Aliases 850
Cross References 850
Information About Languages 851
Case Mappings 851
Decompositions 852
Standardized Variation Sequences 853
Subheads 854

24.2 CJK Ideographs . . . 855
CJK Unified Ideographs . . . 855
Compatibility Ideographs . . . 856
24.3 Hangul Syllables . . . 858
A Notational Conventions . . . 859
Code Points . . . 859
Character Names . . . 859
Character Blocks . . . 859
Sequences . . . 860
Rendering . . . 860
Properties and Property Values . . . 861
Miscellaneous . . . 861
Extended BNF . . . 861
Operators . . . 863
B Unicode Publications and Resources . . . 865
B.1 The Unicode Consortium . . . 866
The Unicode Technical Committee . . . 866
Other Activities . . . 866
B.2 Unicode Publications . . . 867
B.3 Unicode Technical Standards . . . 868
UTS #6: A Standard Compression Scheme for Unicode . . . 868
UTS #10: Unicode Collation Algorithm . . . 868
UTS #18: Unicode Regular Expressions . . . 868
UTS #22: Character Mapping Markup Language (CharMapML) . . . 868
UTS #35: Unicode Locale Data Markup Language (LDML) . . . 868
UTS #37: Unicode Ideographic Variation Database . . . 868
UTS #39: Unicode Security Mechanisms . . . 868
UTS #46: Unicode IDNA Compatibility Processing . . . 869
B.4 Unicode Technical Reports . . . 870
UTR #16: UTF-EBCDIC . . . 870
UTR #17: Unicode Character Encoding Model . . . 870
UTR #20: Unicode in XML and Other Markup Languages . . . 870
UTR #23: The Unicode Character Property Model . . . 870
UTR #25: Unicode Support for Mathematics . . . 870
UTR #26: Compatibility Encoding Scheme for UTF-16: 8-Bit (CESU-8) . . . 870
UTR #33: Unicode Conformance Model . . . 871
UTR #36: Unicode Security Considerations . . . 871
UTR #50: Unicode Vertical Text Layout . . . 871
UTR #51: Unicode Emoji . . . 871
B.5 Unicode Technical Notes . . . 872
B.6 Other Unicode Online Resources . . . 873
Unicode Online Resources . . . 873
How to Contact the Unicode Consortium . . . 874

C Relationship to ISO/IEC 10646 875
C.1 History 876
C.2 Encoding Forms in ISO/IEC 10646 882
Zero Extending 882
C.3 UTF-8 and UTF-16 883
UTF-8 883
UTF-16 883
C.4 Synchronization of the Standards 884
C.5 Identification of Features for Unicode 885
C.6 Character Names 886
C.7 Character Functional Specifications 887
D Changes from Previous Versions 889
D.1 Versions of the Unicode Standard 890
D.2 Clause and Definition Updates 895
E Han Unification History 897
E.1 Development of the URO 898
E.2 Ideographic Rapporteur Group 900
E.3 CJK Sources 902
F Documentation of CJK Strokes 907
R References 915
R.1 Source Standards and Specifications 916
R.2 Source Dictionaries for Han 925
R.3 Other Script Sources 926
R.4 Selected Resources: Technical 953
R.5 Selected Resources: Other 956
I General Index 961