

# The Unicode® Standard

## Version 11.0 – 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.

© 2018 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 the Unicode Consortium. — Version 11.0.

Includes index.

ISBN 978-1-936213-19-1 (<http://www.unicode.org/versions/Unicode11.0.0/>)

1. Unicode (Computer character set) I. Unicode Consortium.

QA268.U545 2018

ISBN 978-1-936213-19-1

Published in Mountain View, CA

June 2018

# I Index

The index covers the contents of this core specification. To find topics in the Unicode Standard Annexes, Unicode Technical Standards, and Unicode Technical Reports, use the search feature on the Unicode website.

For definitions of terms used, see the glossary on the Unicode website. To find the code points for specific characters or the code ranges for particular scripts, use the Character Index on the Unicode website. (See *Section B.3, Other Unicode Online Resources*.)

## A

- abbreviation, Coptic ..... 310
- abjads ..... 256, 359
- abstract character sequences
  - definition ..... 90
- abstract characters ..... 29
  - definition ..... 90
- abugidas ..... 257, 258, 441, 621
- accent marks *see* diacritics
- accented characters
  - encoding ..... 12
  - Latin ..... 289
  - normalization ..... 206
- accounting numbers, ideographic ..... 176
- acrophonic numerals ..... 205, 307
- Adlam ..... 762–763
- Aegean numbers ..... 340
- Africa
  - scripts of ..... 741–764
- Afrikaans ..... 294
- Ahom ..... 616–617
- Ainu ..... 721
- Aiton ..... 636
- Alchemical Symbols ..... 844
- Algonquian ..... 768
- Ali Gali ..... 528
- aliases
  - character name ..... 88, 181
  - informative ..... 896
  - normative ..... 897
  - property ..... 162
  - property value ..... 162
- allocation areas ..... 45
- allocation of encoded characters ..... 44–52
- Alphabetic (informative property) ..... 188
- alphabets ..... 256
  - European ..... 287–335
  - mathematical ..... 801–805
- alternate format characters (deprecated) ... 192, 870–871
- Americas
  - scripts of ..... 765–773
- Amharic ..... 742
- Anatolian hieroglyphs ..... 439–440
- Ancient Symbols ..... 848
- angle brackets (U+2329 and U+232A)
  - deprecated for technical publication ..... 831
- Annexes, Unicode Standard (UAX) ..... xxiv, 917
  - as components of Unicode Standard ..... 79
  - conformance ..... 85
  - list of ..... 85
- annotation characters ..... 883–885
  - use in plain text discouraged ..... 884
- ANSI/ISO C
  - wchar\_t and Unicode ..... 200
- apostrophe (U+0027) ..... 272
- Arabic ..... 367–390
  - digits ..... 808
- Arabic-Indic digits ..... 371–372
  - signs used with ..... 373
- ArabicShaping.txt ..... 375, 380, 396
- Aramaic ..... 412, 441, 528, 557, 563
- areas of the Unicode Standard ..... 45
- ARIB ..... 840
- Armenian ..... 318–319
- arrows ..... 827–828
- ASCII
  - characters with multiple semantics ..... 262
  - transparency of UTF-8 ..... 36
  - Unicode modeled on ..... 1
  - zero extension ..... 200, 929
- Assamese ..... 468
- assigned code points ..... 11, 30
- Athapascan ..... 768
- atomic character boundaries ..... 218
- Avestan ..... 420

**B**

Balinese ..... 673–678  
 Bamum ..... 757–758  
 Bangla ..... 468–474  
 base characters ..... 326  
   definition ..... 106  
   multiple ..... 59  
   ordered before combining marks ..... 220, 326  
 Basic Multilingual Plane (BMP) ..... 1, 44  
   allocation areas ..... 49  
   representation in UTF-16 ..... 36  
 Basque ..... 294  
 Bassa Vah ..... 759  
 Batak ..... 684–685  
 benefits of Unicode ..... 1  
 Bengali ..... 468–474  
 Bhaiksuki ..... 569–570  
 Bidi Class (normative property) ..... 171  
 Bidi Mirrored (normative property) ..... 178  
 Bidi Mirroring Glyph (informative property) ..... 179  
 BidiMirroring.txt ..... 179  
 Bidirectional Algorithm, Unicode ..... 53, 84  
 bidirectional ordering ..... 20  
   controls ..... 867  
 bidirectional text ..... 53, 84  
   Middle Eastern scripts ..... 359  
   nonspacing marks in ..... 223  
   punctuation in ..... 261  
 big-endian ..... 40  
   definition ..... 83  
 Bihari ..... 464  
 binary comparison and sort order  
   caution for UTF-16 ..... 36  
   UTF differences ..... 231, 233  
   UTF-8 ..... 39  
 block ..... 45, 90, 255, 891  
   headers ..... 903  
 BMP *see* Basic Multilingual Plane  
 BNF (Backus-Naur Form) ..... 911  
 BOCU-1 *see* UTN #6, BOCU-1  
   MIME-Compatible Unicode Compression  
 Bodhi ..... 517  
 Bodo ..... 463  
 BOM (U+FEFF) ..... 40, 67, 130–133, 881–883  
 Bopomofo ..... 717–719  
 boundaries, text ..... 61, 189, 217–218, 228  
   *see also* UAX #14, Unicode Line Breaking Algorithm  
   *see also* UAX #29, Unicode Text Segmentation  
 boustrophedon ..... 53, 349  
 box drawing symbols ..... 835  
 Brahmi ..... 441, 557, 559–562, 563, 623  
 Braille ..... 776–777  
 Breton ..... 294

Buginese ..... 671–672  
 Buhid ..... 668  
 Bulgarian ..... 312  
 bullets ..... 275  
   numeric ..... 809  
 Burmese *see* Myanmar  
 Byelorussian ..... 312  
 byte order mark (BOM) (U+FEFF) .. 40, 67, 130–133,  
   881–883  
 byte ordering  
   changing ..... 81  
   conformance ..... 83  
 byte serialization ..... 40, 67  
 Byzantine Musical Symbols ..... 784

**C**

C language  
   wchar\_t and Unicode ..... 200  
 C0 and C1 control codes ..... 31, 187, 856  
 Cambodian *see* Khmer  
 Canadian Aboriginal Syllabics ..... 768–769  
 candrabindu ..... 466, 594  
 canonical composite characters  
   *see* canonical decomposable characters  
 canonical composition algorithm ..... 138  
 canonical decomposable characters  
   definition ..... 118  
 canonical decomposition ..... 63  
   definition ..... 117  
   mappings ..... 116  
 canonical equivalence  
   definition ..... 118  
   nonspacing marks ..... 225  
 canonical equivalent character sequences  
   conformance ..... 81  
 canonical mappings  
   *see* canonical decomposition mappings  
 canonical ordering algorithm ..... 137  
 canonical precomposed characters  
   *see* canonical decomposable characters  
 Cantonese ..... 700  
 capital letters ..... 164, 236, 287  
 Carian ..... 343  
 carriage return (U+000D) (CR) ..... 209, 857  
 carriage return and line feed (CRLF) ..... 209  
 case ..... 295  
   and text processes ..... 12  
   beyond ASCII ..... 237  
   camelcase ..... 239  
   case folding ..... 240  
   case operations (conformance) ..... 85, 152–158  
   case operations and normalization ..... 242  
   case operations, reversibility ..... 239

- cased (definition) ..... 153
- case-insensitive comparison ..... 157, 231, 240
- casing context (definition) ..... 153
- conversion ..... 154
- detection ..... 156
- European alphabets ..... 287
- exceptional Latin pairs ..... 291, 295
- Georgian ..... 321
- lowercase ..... 164, 236, 287
- mapping tables ..... 196
- mappings ..... 152, 166, 236–238
- mappings noted in code charts ..... 900
- titlecase ..... 164, 236
- Turkish I ..... 238, 291
- uppercase ..... 164, 236, 287
- see also* default case
- Case (normative property) ..... 164, 236
- CaseFolding.txt ..... 166, 240
- caseless letters ..... 295
- Catalan ..... 293
- Caucasian Albanian ..... 354
- cedilla ..... 290
- CEF *see* character encoding forms
- CES *see* character encoding schemes
- Chakma ..... 547
- Cham ..... 661–662
- character encoding forms (CEF) ..... 33–39, 929
- see also* Unicode encoding forms
- character encoding model ..... 33, 42
- see also* UTR #17, Unicode Character Encoding Model
- character encoding schemes (CES) ..... 40–43
- see also* Unicode encoding schemes
- character encoding standards
- coverage by Unicode ..... 3
- Character Index ..... 918
- character literals, Unicode
- code point notation U+ ..... 912
- character names ..... 88, 180–186, 933
- aliases ..... 88, 181
- conventions ..... 909
- for CJK ideographs ..... 905
- for control codes ..... 185, 187
- in code charts ..... 896
- matching ..... 181
- character properties
- see* properties
- see also* individual properties, e.g. Combining Class
- character semantics ..... 1, 80, 87–88, 934
- as Unicode design principle ..... 18
- ASCII ..... 262
- definition ..... 87
- character sequences
- abstract *see* abstract character sequences
- canonical equivalent *see* canonical equivalent
- character sequences
- compatibility equivalent *see* compatibility equivalent
- character sequences
- conformance ..... 81
- named ..... 181
- character sequences, combining ..... 106
- character shaping selectors (deprecated) ..... 870
- character tabulation (U+0009) ..... 857
- characters
- abstract *see* abstract characters
- arrangement in Unicode ..... 46
- assigned ..... 11, 30
- boundaries ..... 217
- canonical decomposable *see* canonical decomposable
- characters
- classes ..... 912
- code charts ..... 891–908
- coded *see* encoded characters
- combining *see* combining characters
- compatibility decomposable *see* compatibility
- decomposable characters
- composite *see* decomposable characters
- concept of ..... 15, 60
- conformance definitions ..... 90–93
- confusable ..... 245
- conversion ..... 196–197
- decomposable *see* decomposable characters
- deprecated *see* deprecated characters
- encoded *see* encoded characters
- encoding forms *see* encoding forms
- encoding schemes *see* encoding schemes
- end-user perceived ..... 60
- format control ..... 30, 68, 263, 855–871
- glyphs, relationship to ..... 15
- graphic ..... 30
- identity (definition) ..... 87
- ignored in processing ..... 248–253
- interpretation ..... 80
- layout control ..... 68, 859–869
- modification ..... 81
- names list ..... 892–904
- names *see* character names
- not encoded in Unicode ..... 3
- number encoded in Version 10.0 ..... 3
- precomposed *see* decomposable characters
- properties *see* properties
- semantics *see* character semantics
- special ..... 67, 855–890
- supplementary *see* supplementary characters
- transcoding ..... 196–197
- unsupported ..... 201

- characters, not glyphs
  - in spoofing ..... 246
  - Unicode principle ..... 15
- charsets
  - IANA registered names ..... 41
- Cherokee ..... 766
- Chinese ..... 699–701
  - Cantonese ..... 700
  - Hakka ..... 718
  - Mandarin ..... 700
  - Minnan (Hokkien/Fujian, incl. Taiwanese) .. 718
  - simplified and traditional ..... 699
- Chu hán ..... 698
- Chu Nôm ..... 944
- citations for
  - properties ..... 77
  - Unicode algorithms ..... 78
  - Unicode Standard ..... 76
- CJK ideographs ..... 258, 694–710
  - accounting numbers ..... 176
  - CJK Compatibility Ideographs ..... 709–710
  - CJK Compatibility Supplement ..... 710
  - CJK Strokes ..... 712, 947
  - CJK Unified Ideographs ..... 694–708
  - CJK Unified Ideographs Extension A ..... 696
  - CJK Unified Ideographs Extension B ..... 708
  - CJK Unified Ideographs Extension C ..... 709
  - CJK Unified Ideographs Extension D ..... 709
  - CJK Unified Ideographs Extension E ..... 709
  - CJK Unified Ideographs Extension F ..... 709
- code charts ..... 905
  - compatibility ideographs in Plane 2 ..... 52
  - component structure ..... 704
  - encoding blocks ..... 695
  - ideographic description sequences ..... 713–716
  - ideographic variation mark (U+303E) ..... 715
  - KangXi radicals ..... 707, 711–712
  - names ..... 905
  - numbers ..... 808
  - numeric values ..... 176, 205
  - order of encoding ..... 706
  - radicals ..... 711–712
  - source standards ..... 946
  - unknown or unavailable ..... 284
  - Vietnamese ..... 692
- CJK Miscellaneous Area ..... 50
- CJK punctuation and symbols ..... 282
  - compatibility forms ..... 284
  - overscores and underscores ..... 284
  - quotation marks ..... 270
  - sesame dots ..... 283
  - vertical forms ..... 284
- CJK-JRG (Chinese/Japanese/Korean Joint Research Group) ..... 942
- CJKV Ideographs Area ..... 50
- cluster boundaries ..... 217
- code charts ..... 891–908
  - representative glyphs ..... 892
- code point sequences
  - notation ..... 910
- code points ..... 7, 29
  - assigned ..... 11, 30
  - assignment ..... 46
  - categories ..... 30
  - default ignorable ..... 201, 252
  - definition ..... 90
  - designated ..... 30
  - notation ..... 909
  - number in Unicode Standard ..... 1
  - private-use *see* private-use code points
  - reserved *see* reserved code points
  - semantics ..... 32
  - surrogate *see* surrogates
  - unassigned *see* unassigned code points
  - undesignated ..... 30
- code positions *see* code points
- code set independence ..... 18
- code unit sequences
  - definition ..... 120
  - ill-formed (definition) ..... 122
  - notation ..... 910
  - well-formed (definition) ..... 122
- code units
  - definition ..... 120
  - isolated ..... 119
- code values *see* code units
- coded character representations
  - see* coded character sequences
- coded character sequences
  - definition ..... 92
- coded characters *see* encoded characters
- codespace *see* Unicode codespace
- coeng ..... 637, 640
- Collation Algorithm, Unicode (UCA) ..... 12
- collation *see* sorting
- collation tables ..... 196
- combining character sequences ..... 56, 106
  - defective ..... 223
  - definition ..... 108
  - Latin ..... 289
  - line breaking ..... 219
  - matching ..... 219
  - order of base character and marks ..... 220, 326
  - rendering ..... 219
  - selection ..... 217
  - truncation ..... 220–221
- combining characters ..... 55–60, 110–115, 219–227
  - blocking reordering ..... 866

- canonical ordering ..... 62, 137, 168
- combining marks ..... 326–327
- definition ..... 106
- dependence ..... 326
- display order ..... 58
- keyboard input ..... 220
- ligatures ..... 59
- multiple ..... 57
- multiple base characters ..... 59
- normalization of ..... 206
- ordering conventions ..... 56
- rendering of marks ..... 222–227
- reordrant ..... 169
- script-specific ..... 56
- split ..... 169
- strikethrough ..... 170
- subjoined ..... 170
- typographical interaction ..... 58, 168
- vertical stacking ..... 58
- see also* diacritics
- Combining Class (normative property) ..... 168
- combining classes ..... 135, 168, 225–226
  - class zero characters ..... 168
  - definition ..... 135
- combining grapheme joiner (U+034F) ..... 865
- combining half marks ..... 190, 334
- combining marks *see* combining characters
- comma below ..... 290
- Compatibility and Specials Area ..... 26, 50
- compatibility characters ..... 22
- compatibility composite characters ..... 27
  - see* compatibility decomposable characters
- compatibility decomposable characters ..... 26
  - definition ..... 116
- compatibility decomposition ..... 63
  - definition ..... 116
- compatibility decomposition mappings ..... 116
- compatibility equivalence
  - definition ..... 117
- compatibility equivalent character sequences
  - conformance ..... 81
  - see* compatibility decomposition mappings
- compatibility precomposed characters
  - see* compatibility decomposable characters
- compatibility variants ..... 26
  - mapping ..... 243
- composite characters
  - see* decomposable characters
- Composition Exclusion (normative property) ... 100
- compression ..... 208
  - see also* UTS #6, A Standard Compression Scheme for Unicode (SCSU)
- conferences ..... 918
- conformance ..... 73–158
  - definitions ..... 87–93
  - examples ..... 69
  - ISO/IEC 10646 implementations ..... 934
  - requirements ..... 79–84
- confusables ..... 245
- conjunct consonants
  - Indic ..... 217, 449
  - Myanmar ..... 631
  - selection of clusters ..... 217
- contextual shaping
  - apostrophe ..... 272
  - Arabic ..... 367
    - not used for Hebrew final forms ..... 362
    - quotation marks ..... 268
  - Syriac ..... 395
- contour tones ..... 324
- control codes ..... 31, 68, 856
  - graphics for ..... 830
  - names ..... 187
  - properties ..... 857
  - semantics ..... 32, 857
  - specified in Unicode ..... 857
- control sequences ..... 856
- conversion of characters ..... 196–197
- convertibility
  - as Unicode design principle ..... 24
- Coptic ..... 306, 309–311
- Coptic Epact numbers ..... 813
- corporate use subarea ..... 876
- corrigenda ..... 76
- CR (U+000D carriage return) ..... 209, 857
- CRLF (carriage return and line feed) ..... 209
- Croatian ..... 294
  - digraphs ..... 294
- culturally expected sorting ..... 12, 230
- Cuneiform
  - Old Persian ..... 431
  - Sumero-Akkadian ..... 426–429
  - Ugaritic ..... 430
- Cuneiform and Hieroglyphic Area ..... 51
- Cuneiform and Hieroglyphs ..... 425–440
- currency symbols block ..... 795–798
  - currency symbols encoded in other blocks .. 796
  - currency symbols, other ..... 797
  - dollar sign, form and usage ..... 796
  - euro sign ..... 797
  - lari sign ..... 797
  - lira sign, compatibility usage ..... 796
  - lira sign, Turkish ..... 797
  - peso signs, usage ..... 796
  - ruble sign ..... 797
  - rupee signs, Indian, usage ..... 797
  - yen and yuan signs, usage ..... 796

- cursive joining ..... 861–865
  - Arabic ..... 375–382
  - control characters for .... 191, 369–370, 531, 860
  - Mandaic ..... 403
  - Mongolian ..... 530–532
  - N’Ko ..... 753
  - Phags-pa ..... 576
  - Syriac ..... 395–398
  - transparency ..... 864
- cursive scripts ..... 359
- Cypriot ..... 342
  - see also* Linear B
- Cyrillic ..... 312–315
- Czech ..... 294
- D**
- danda, in Devanagari block ..... 462
- Danish ..... 293
- dashes ..... 265
- Database, Unicode Character
  - see* Unicode Character Database (UCD)
- dead consonants, Indic ..... 446
- dead keys ..... 220
- decomposable characters ..... 63
  - definition ..... 116
  - normalization of ..... 206
- decomposition ..... 63, 116–118
  - canonical *see* canonical decomposition
  - compatibility *see* compatibility decomposition
  - definition ..... 116
  - in normalization ..... 206
  - mapping, definition ..... 116
  - mappings noted in code charts ..... 900
- default case
  - algorithms ..... 85, 152–158
  - conversion ..... 154
  - detection ..... 156
  - folding ..... 155
- default caseless matching ..... 157
- default grapheme clusters ..... 217
  - see also* UAX #29, Unicode Text Segmentation
- Default Ignorable Code Point (property) ..... 252
- default ignorable code points ..... 201, 252
- default property values
  - definition ..... 97
- defective combining character sequences ..... 223
  - definition ..... 108
- dependent vowel signs
  - Indic ..... 445
  - Khmer ..... 642
  - Philippine scripts ..... 668
- deprecated characters ..... 74, 895
  - alternate format ..... 192, 870–871
  - definition ..... 92
- Derived Age (property) ..... 201
- derived properties
  - definition ..... 104
- DerivedCoreProperties.txt ..... 153, 164, 253
- DerivedNormalizationProps.txt ..... 242
- Deseret ..... 771–773
- design goals of Unicode ..... 4
- design principles of Unicode ..... 14–24
- designated code points ..... 30
- Devanagari ..... 443–467
- Dhivehi ..... 511
- diacritics ..... 55, 326
  - alternative glyphs ..... 289, 326
  - Czech ..... 290
  - display in isolation ..... 60, 265, 327
  - double ..... 114, 190, 328
  - German dialectology ..... 332
  - Greek ..... 302–303, 306
  - Latin ..... 289–292
  - Latvian ..... 290
  - mathematical ..... 804
  - on i and j ..... 291
  - rendering ..... 222–227
  - Slovak ..... 290
  - spacing clones of ..... 324, 328
  - symbol ..... 55, 333
  - see also* combining characters
- dictionary symbols ..... 840
- digit form names ..... 371
- digits ..... 205
  - Arabic ..... 808
  - Arabic-Indic ..... 371–372
  - compatibility ..... 808
  - decimal ..... 175
  - glyph variants ..... 810
  - hexadecimal ..... 808
  - Myanmar ..... 808
  - national shapes ..... 871
  - Shan ..... 808
  - superscript and subscript ..... 809
  - Tai Laing ..... 808
  - Tai Tham ..... 808
- digraphs ..... 294, 297, 299
- dingbats ..... 843–844
- directionality ..... 20, 53
  - East Asian scripts ..... 692
  - Middle Eastern scripts ..... 359
  - Mongolian ..... 529
  - musical symbols ..... 779
  - normative property ..... 171
  - Ogham ..... 356
  - Old Italic ..... 346
  - Philippine scripts ..... 669
  - Runic ..... 349

discussion list for Unicode ..... 918  
 Dogra ..... 619–620  
 Dogri ..... 463  
 Domino Tiles ..... 845  
 dotless i ..... 238, 291  
 dotted circle  
   in code charts ..... 107, 327  
   in fallback rendering ..... 222  
   to indicate diacritic ..... 55  
   to indicate vowel sign placement ..... 56  
 double diacritics ..... 114, 190, 328  
 Duployan ..... 788–789  
 Dutch ..... 293, 294  
 dynamic composition  
   as Unicode design principle ..... 23  
 Dzongkha ..... 517

## E

East Asian scripts ..... 691–740  
   writing direction ..... 53  
   *see also* CJK ideographs  
 Eastern Arabic-Indic digits ..... 371  
 EBCDIC  
   newline function ..... 210  
 editing, text boundaries for ..... 217–218  
 efficiency  
   as Unicode design principle ..... 15  
 Egyptian hieroglyphs ..... 432–436  
 Elbasan ..... 353  
 ellipsis ..... 273–274  
 e-mail discussion list for Unicode ..... 918  
 emoji ..... 838, 918  
   animal symbols ..... 842  
   charts ..... 918  
   cultural symbols ..... 842  
   zodiacal symbols ..... 842  
 emoji modifiers ..... 842  
 emoticons ..... 842  
 Enclosed Alphanumerics ..... 852  
 enclosing marks ..... 334  
   definition ..... 107  
 encoded characters ..... 7, 29  
   allocation ..... 44–52  
   definition ..... 92  
 encoding form conversion  
   definition ..... 127  
 encoding forms ..... 33–39  
   ISO/IEC 10646 definitions ..... 929  
 encoding forms, Unicode  
   *see* Unicode encoding forms  
 encoding model for Unicode characters ..... 33, 42  
   *see also* UTR #17, Unicode Character Encoding Model

encoding schemes ..... 40–43  
 encoding schemes, Unicode  
   *see* Unicode encoding schemes  
 endian ordering  
   *see* byte order mark (BOM) (U+FEFF)  
 end-user subarea ..... 877  
 English ..... 293  
 equivalent sequences ..... 206  
   as Unicode design principle ..... 23  
   case-insensitivity ..... 231, 240  
   combining characters in matching ..... 219  
   conformance ..... 82  
   Hangul syllables ..... 727  
   in sorting and searching ..... 230  
   language-specific ..... 118  
   security implications ..... 245  
   *see also* canonical equivalence  
   *see also* compatibility equivalence  
   *see also* encoding forms, encoding schemes  
 errata ..... xxvii, 76, 919  
 escape sequences ..... 856  
   not used in Unicode ..... 1, 4  
 Esperanto ..... 294  
 Estonian ..... 294  
 Ethiopic ..... 742–745  
 Etruscan ..... 345  
 European scripts ..... 287–335  
   ancient ..... 337–357  
 eyelash-RA ..... 455

## F

fallback rendering ..... 252  
   of nonspacing marks ..... 222  
 FAQ (Frequently Asked Questions) ..... 918  
 Faroese ..... 293  
 Farsi ..... 367, 370  
 featural syllabaries ..... 257  
 FF (U+000C form feed) ..... 209, 857  
 file separator (U+001C) ..... 857  
 Finnish ..... 293  
 Finno-Ugric Transcription (FUT)  
   *see* Uralic Phonetic Alphabet (UPA)  
 fixed-width Unicode encoding form (UTF-32) ... 35,  
   124  
 flat tables ..... 196  
 Flemish ..... 293  
 fleurons ..... 844  
 fonts  
   and Unicode characters ..... 16  
   for mathematical alphabets ..... 803–805  
   style variation for symbols ..... 793  
 form feed (U+000C) (FF) ..... 209, 857



format control characters ..... 30, 68, 263, 855–871  
     deprecated ..... 870–871  
     prefixed ..... 192, 330  
     stateful ..... 868  
 fraction characters ..... 821  
 fraction slash (U+2044) ..... 273, 817  
 French ..... 294  
 Frisian ..... 294  
 fullwidth forms in East Asian encodings ..... 724  
 futhark ..... 348

## G

Garshuni ..... 391  
 Ge'ez ..... 742  
 General Category (normative property) ..... 172  
     list of values ..... 172  
 general punctuation ..... 261–285  
 General Scripts Area ..... 50  
 geometrical symbols ..... 835–837  
 Georgian ..... 320–321  
 German ..... 293  
 geta mark (U+3013) ..... 284  
 Glagolitic ..... 317  
 Glossary ..... 918  
 glyph selection tables ..... 196  
 glyphs ..... 6, 15  
     characters, relationship to ..... 15  
     diacritics alternative ..... 289, 326  
     Greek alternative ..... 303–305  
     Latin alternative ..... 289  
     mathematical alternative ..... 823  
     missing ..... 252  
     representative in code charts ..... 892  
     standardized variants ..... 872  
     symbols alternative ..... 793  
 golden numbers ..... 350  
 Gothic ..... 352  
 Grantha ..... 613–615  
 grapheme base ..... 326  
     definition ..... 108  
 grapheme clusters ..... 11, 60–61  
     *see also* UAX #29, Unicode Text Segmentation  
     default ..... 217  
     definition ..... 109  
 grapheme extender  
     definition ..... 109  
 grapheme joiner, combining (U+034F) ..... 865  
 graphic characters ..... 30  
 Greek ..... 302–307  
     acrophonic numerals ..... 205, 307  
     alternative glyphs ..... 303–305  
     ancient musical notation ..... 785–787  
     editorial marks ..... 279

    letters as symbols ..... 303–305, 824  
     *see also* Cypriot, Linear B  
 Greenlandic ..... 294  
 group separator (U+001D) ..... 857  
 guillemets ..... 268  
 Gujarati ..... 480–481  
 Gunjala Gondi ..... 554–555  
 Gurmukhi ..... 475–479

## H

Hakka ..... 718  
 halant ..... 441  
     *see also* virama  
 half marks, combining ..... 190, 334  
 half-consonants, Indic ..... 450  
 halfwidth forms in East Asian encodings ..... 724  
 Han ideographs *see* CJK ideographs  
 Han unification ..... 701–708  
     and language tags ..... 215  
     history ..... 941–946  
     language usage ..... 699  
     source separation rule ..... 696, 702  
     source standards ..... 946  
 hand symbols ..... 842  
 Hangul Area ..... 50  
 Hangul syllables ..... 691, 725–728  
     and combining marks ..... 114  
     as grapheme clusters ..... 61  
     canonical decomposition ..... 144  
     collation ..... 727  
     composition ..... 146  
     conjoining jamo ..... 142–151  
     equivalent sequences ..... 727  
     Hangul Compatibility Jamo ..... 726  
     Hangul Jamo ..... 725–728  
     Hangul Syllables block ..... 727–728  
     Johab set ..... 727  
     name generation ..... 147  
     normalization ..... 726  
     standard ..... 143  
 Hangzhou numerals ..... 816  
 Hanifi Rohingya ..... 666  
 Hanja *see* CJK ideographs  
 Hanunóo ..... 668  
 Hanzi *see* CJK ideographs  
 harakat ..... 368  
 hasant ..... 468  
 hash tables ..... 197  
 Hatran ..... 424  
 Hebrew ..... 361–366  
 hentaigana ..... 721–722

- hieroglyphs
    - Anatolian ..... 439–440
    - Egyptian ..... 432–436
    - Meroitic ..... 437–438
  - high surrogate
    - definition ..... 119
    - high-surrogate code points ..... 79, 878
    - high-surrogate code units ..... 119
  - higher-level protocols
    - definition ..... 93
  - Hindi ..... 443
  - Hiragana ..... 720
  - horizontal tab (U+0009) ..... 857
  - HTML newline function ..... 210
  - Hungarian ..... 294
  - hyphenation ..... 860
    - as a text process ..... 10
  - hyphens ..... 265, 860
- I**
- I Ching symbols ..... 847
  - IANA charset names ..... 41
  - Icelandic ..... 293
  - identifiers ..... 229
    - see also* UAX #31, Unicode Identifier and Pattern Syntax
  - Ideographic (informative property) ..... 188
  - ideographic description sequences ..... 714
  - Ideographic Rapporteur Group (IRG) ..... 944
  - ideographs *see also* CJK ideographs
  - IICore ..... 697, 944
  - ill-formed
    - definition ..... 122
  - Imperial Aramaic ..... 412–413
  - implementation guidelines ..... 195–254
  - in a Unicode encoding form
    - definition ..... 123
  - in-band mechanisms ..... 890
  - India
    - Official scripts ..... 441–507
  - Indian rupee signs, usage ..... 797
  - Indic scripts ..... 441–507
    - principles, in terms of Devanagari ..... 444–454
    - relation to ISCII standard ..... 443
  - Indic Siyaq ..... 814
  - Indonesia and Oceania
    - scripts of ..... 667–689
  - Indonesian ..... 293
  - industry character sets
    - covered in Unicode ..... 3
  - information separators (U+001C..U+001F) ..... 857
  - informative properties
    - definition ..... 101
  - Inscriptional Pahlavi ..... 418
  - Inscriptional Parthian ..... 418
  - inside-out rule ..... 222
  - interchange restrictions ..... 31
  - International Phonetic Alphabet (IPA) 256, 296–297
    - Spacing Modifier Letters ..... 323
    - see also* phonetic alphabets
  - internationalization ..... 18
  - Internationalization & Unicode Conference .... 918
  - Internet protocols
    - UTF-8 as preferred encoding ..... 37
  - Inuktitut ..... 768
  - invisible operators ..... 829
  - iota subscript ..... 303
  - IPA *see* International Phonetic Alphabet
  - IRG (Ideographic Rapporteur Group) ..... 944
  - Irish ..... 293, 356
  - ISCII standard and Unicode ..... 443
  - ISO/IEC 10646 ..... 921–934
    - conformance of Unicode implementations .. 934
    - encoding forms ..... 929
    - synchrony with Unicode Standard ..... 931
    - timeline compared to Unicode versions ..... 923
  - Italian ..... 293
  - ITC Zapf Dingbats ..... 843
  - IUC *see* Internationalization & Unicode Conference
- J**
- jamos *see* Hangul syllables
  - Japanese ..... 691
  - Javanese ..... 679–682
  - Jawi ..... 387
  - jihvamuliya ..... 467, 594
  - Johab ..... 727
  - joiners ..... 369
    - combining grapheme joiner (U+034F) ..... 865
    - word joiner (U+2060) ..... 859
    - zero width joiner (U+200D) ..... 369–370, 862
  - justification ..... 224
- K**
- Kaithi ..... 591–593
  - Kana (Hiragana and Katakana) ..... 720–721
  - Kanbun ..... 710
  - KangXi radicals ..... 707, 711–712
  - Kanji *see* CJK ideographs
  - Kannada ..... 497–500
  - Kashmiri ..... 464
  - Katakana ..... 720–721
  - Kawi ..... 673, 675
  - Kayah Li ..... 660
  - KC (normalization form)
    - see* Normalization Form KC

KD (normalization form)  
*see* Normalization Form KD

keytop labels ..... 830

Khamti Shan ..... 634

Kharoshthi ..... 563–564

Khmer ..... 637–648  
 characters not recommended ..... 645  
 syllable components, order of ..... 646

Khojki ..... 602–603

Khudawadi ..... 604–605

killer ..... 258  
 Batak ..... 684  
 Brahmi ..... 559  
 Meetei Mayek ..... 541  
 Myanmar (asat) ..... 632  
*see also* virama

Konkani ..... 463

Korean Hangul *see* Hangul

Kurdish ..... 387

**L**

Ladino ..... 361

language tags ..... 215, 886–890  
 and Han unification ..... 215  
 use strongly discouraged ..... 886, 889

Lanna ..... 651

Lao ..... 627–629

last-resort glyphs ..... 252

Latin ..... 289–301  
 alternative glyphs ..... 289  
 Basic Latin ..... 293  
 encoding blocks ..... 45  
 IPA Extensions ..... 296–297  
 Latin Extended Additional ..... 299–301  
 Latin Extended-A ..... 293  
 Latin Extended-B ..... 294–296  
 Latin Extended-C ..... 299  
 Latin Extended-D ..... 300  
 Latin Extended-E ..... 301  
 Latin Ligatures ..... 299  
 Latin-1 Supplement ..... 293  
 Phonetic Extensions ..... 298–301

Latvian ..... 294, 301  
 cedilla ..... 290

layout control characters ..... 68, 859–869

leading surrogates  
*see* high-surrogate code units

legibility criterion for plain text ..... 19

Lepcha ..... 548–550

letter spacing ..... 860

letterlike symbols ..... 799–805

LF (U+000A line feed) ..... 209, 857

ligatures ..... 861–865  
 Arabic ..... 378–379  
 combining characters on ..... 59  
 control characters for ..... 191  
 for nonspacing marks ..... 226  
 Latin ..... 299  
 selection ..... 218  
 Syriac ..... 398

Limbu ..... 537–540

line breaking ..... 209–213, 859–861  
 control characters ..... 190  
 in South Asian scripts ..... 625, 633, 648  
 recommendations ..... 211  
*see also* UAX#14, Unicode Line Breaking Algorithm

line feed (U+000A) (LF) ..... 209, 857

line separator (U+2028) (LS) ..... 209, 861

line tabulation (U+000B) (VT) ..... 857

Linear A ..... 339

Linear B ..... 340–341  
*see also* Cypriot

linear boundaries ..... 218

Lisu ..... 733–735

Lithuanian ..... 294

little-endian ..... 40  
 definition ..... 83

logical order  
 as Unicode design principle ..... 19  
 exceptions to ..... 169

logograph ..... 258

logosyllabaries ..... 258

low surrogate  
 definition ..... 119  
 low-surrogate code points ..... 79, 878  
 low-surrogate code units ..... 119

lowercase ..... 164, 236, 287

LS (U+2028 line separator) ..... 209, 861

Lycian ..... 343

Lydian ..... 343

**M**

MacOS newline function ..... 210

Mahajani ..... 600–601

Mahjong Tiles ..... 845

mail discussion list for Unicode ..... 918

Maithili ..... 463

major version ..... 75

Makasar ..... 688–689

Malay ..... 293

Malay, Patani ..... 626

Malayalam ..... 501–507  
 Suriyani ..... 399, 502

Maltese ..... 294

Manchu ..... 529

Mandaic ..... 402–404  
Mandarin ..... 700  
Manden ..... 750  
Manichaean ..... 414–417  
map symbols ..... 840  
mapping tables *see* tables of character data  
Marathi ..... 443, 455, 462  
Marchen ..... 578  
markup languages  
    and Unicode conformance ..... 890  
    line breaking ..... 209  
Masaram Gondi ..... 552–553  
Mathematical (informative property) ..... 821  
mathematical expression format characters ..... 192  
    *see also* UTR #25, Unicode Support for Mathematics  
mathematical symbols ..... 821–828  
    alphabets ..... 801–805  
    alphanumeric ..... 800–805  
    fonts ..... 803–805  
    format characters ..... 829  
    fragments for typesetting ..... 831  
    invisible operators ..... 829  
    operators ..... 822–825  
    standardized variants ..... 828  
MathML ..... 825  
matras ..... 168, 445  
Medefaidrin ..... 764  
Meetei Mayek ..... 541–542  
Mende Kikakui ..... 760–761  
Meroitic  
    cursive ..... 437–438  
    hieroglyphs ..... 437–438  
Miao ..... 736–737  
Middle Eastern scripts ..... 359–512  
    ancient ..... 405–424  
Min ..... 700  
Minnan (Hokkien/Fujian, incl. Taiwanese) ..... 718  
minor version ..... 75  
minus sign ..... 824  
    commercial (U+2052) ..... 276  
mirrored property  
    *see* Bidi Mirrored (normative property)  
mirroring of paired punctuation ..... 267  
Miscellaneous Symbols ..... 839  
missing glyphs ..... 252  
Modi ..... 610–612  
modifier letters ..... 322–325  
Modifier Letters, Spacing ..... 299  
Mongolian ..... 528–536, 571  
    writing direction ..... 529  
moon symbols ..... 840  
Mro ..... 543  
Multani ..... 606

multibyte encodings  
    compared to UTF-8 ..... 37  
multistage tables ..... 196  
musical symbols ..... 778–787  
    ancient Greek ..... 785–787  
    Balinese ..... 677  
    Byzantine ..... 784  
    directionality ..... 779  
    Gregorian ..... 783  
    Kievan ..... 783  
    Western ..... 778–783  
Myanmar ..... 630–636  
    digits ..... 808  
    Myanmar Extended-A ..... 634  
    Myanmar Extended-B ..... 634

## N

N’Ko ..... 750–754  
Nabataean ..... 422  
named character sequences ..... 181  
names, character *see* character names  
namespace ..... 89  
NEL (U+0085 next line) ..... 209, 857  
Nepali ..... 443  
neutral directional characters ..... 171  
New Tai Lue ..... 651–653  
Newa ..... 515–516  
newline function (NLF) ..... 210, 858  
newline guidelines ..... 209–213  
next line (U+0085) (NEL) ..... 209, 857  
NFC (Normalization Form C) ..... 62  
NFD (Normalization Form D) ..... 62  
NFKC (Normalization Form KC) ..... 62  
NFKD (Normalization Form KD) ..... 62  
NLF (newline function) ..... 210, 858  
no-break space (U+00A0) ..... 859  
    base for diacritic in isolation ..... 60, 265, 327  
no-break space, narrow (U+202F) ..... 534  
noncharacter code points *see* noncharacters  
noncharacters ..... 31, 879  
    conformance ..... 79  
    definition ..... 93  
    handling ..... 82  
    in code charts ..... 895  
    interchange restrictions ..... 31  
    semantics ..... 32  
    U+10FFFF (not a character code) ..... 879  
    U+FDD0..U+FDEF ..... 31, 879  
    U+FFFE (not a character code) ..... 67, 880  
    U+FFFF (not a character code) ..... 31, 879  
nondecomposable characters ..... 64  
non-joiner, zero width (U+200C) ..... 369–370, 863  
nonlinear boundaries ..... 218

- non-overlap principle in Unicode encoding forms 33
- nonspacing marks ..... 326  
 definition ..... 107  
 display in isolation ..... 60, 265, 327  
 positioning ..... 226  
 rendering ..... 222–227  
*see also* combining characters  
*see also* diacritics
- normalization ..... 62, 206–207  
 and case operations ..... 242  
 canonical ordering algorithm ..... 62, 137, 168  
 conformance ..... 84  
 of private-use characters ..... 876  
*see also* UAX #15, Unicode Normalization Forms  
 stability ..... 134
- Normalization Form C (NFC) ..... 62
- Normalization Form D (NFD) ..... 62
- Normalization Form KC (NFKC) ..... 62
- Normalization Form KD (NFKD) ..... 62
- normalization forms ..... 134–141  
 definition ..... 140  
 specification ..... 136
- normative behaviors  
 definition ..... 87
- normative properties  
 definition ..... 99  
 list ..... 100  
 may change ..... 99
- Norwegian ..... 293
- notational conventions ..... 909–913
- notational systems ..... 260, 775–791
- nukta ..... 368, 388, 456
- null (U+0000)  
 as Unicode string terminator ..... 858
- number forms  
 CJK ideographs ..... 205
- numbers  
 Coptic Epact ..... 813  
 handling ..... 205  
 ideographic accounting ..... 176
- numerals ..... 806–818  
 acrophonic ..... 307  
 Chinese counting rods ..... 819  
 Coptic ..... 311  
 Cuneiform ..... 429  
 Ethiopic ..... 744  
 Greek acrophonic ..... 205  
 Hangzhou ..... 816  
 Meroitic cursive ..... 438  
 old-style ..... 273  
 Roman ..... 205, 821  
 Rumi ..... 814  
 Suzhou-style ..... 816
- numeric separators ..... 276
- numeric shape selectors (deprecated) ..... 871
- Numeric Type (normative property) ..... 175
- Numeric Value (normative property) ..... 175
- numero sign (U+2116) ..... 799
- Nūshu ..... 732
- ## O
- object replacement character (U+FFFC) ..... 885
- octet ..... 911
- Ogham ..... 356
- OI Chiki ..... 545–546
- Old Church Slavonic ..... 312
- Old Hungarian ..... 351
- Old Italic ..... 345–347
- Old North Arabian ..... 407
- Old Permic ..... 355
- Old Persian ..... 431
- Old Sogdian ..... 585
- Old South Arabian ..... 408–409
- Old Turkic ..... 584
- old-style numerals ..... 273
- Oriya ..... 482–484
- ornamental dingbats ..... 844
- Oromo ..... 742
- Osage ..... 770
- Osmanya ..... 746
- out-of-band mechanisms ..... 890
- overlapping encodings ..... 33
- overscores ..... 273
- ## P
- Pahawh Hmong ..... 663–664
- Pahlavi, Inscriptional ..... 418
- Pahlavi, Psalter ..... 419
- Palmyrene ..... 423
- Panjabi ..... 475
- paragraph or section marks ..... 276
- paragraph separator (U+2029) (PS) ..... 209, 861
- Parthian, Inscriptional ..... 418
- Pashto ..... 367
- Patani Malay ..... 626
- Pau Cin Hau ..... 665
- Persian ..... 367, 370
- Phags-pa ..... 571–577
- Phaistos Disc symbols ..... 848
- Phake ..... 636
- Philippine scripts ..... 668–670
- Phoenician ..... 410
- phonemes ..... 259
- phonetic alphabets ..... 256  
 IPA Extensions ..... 296–297  
 Phonetic Extensions ..... 298–301  
 Spacing Modifier Letters ..... 323–325

- Uralic Phonetic Alphabet (UPA) . . . . . 276, 298
    - see also* International Phonetic Alphabet (IPA)
  - Pinyin . . . . . 293
  - pipeline table
    - proposed new characters . . . . . 919
  - pivot code, Unicode as . . . . . 196
  - plain text
    - as Unicode design principle . . . . . 18
    - legibility criterion . . . . . 19
  - planes of Unicode codespace . . . . . 44
    - Plane 0 (BMP) . . . . . 44
    - Plane 1 (SMP) . . . . . 44, 51
    - Plane 14 (SSP) . . . . . 45
    - Plane 2 (SIP) . . . . . 44, 52
    - Planes 15-16 (Private Use) . . . . . 52, 877
  - Playing Cards . . . . . 846
  - points, Hebrew pronunciation marks . . . . . 361
  - policies of the Unicode Consortium . . . . . 919
  - Polish . . . . . 294
  - Portuguese . . . . . 293
  - precomposed characters
    - see* decomposable characters
    - compatibility *see* compatibility decomposable characters
  - prefixed format control characters . . . . . 192
  - prepended concatenation marks . . . . . 253, 330
  - Private Use Area (PUA) . . . . . 50, 876
  - Private Use planes . . . . . 45, 52, 877
  - private-use characters
    - properties . . . . . 875
    - semantics . . . . . 32
  - private-use code points . . . . . 31, 201
    - conformance . . . . . 80
    - definition . . . . . 105
    - high surrogates . . . . . 878
  - processing code, Unicode as . . . . . 38
  - properties . . . . . 18, 95–105, 159–193
    - aliases . . . . . 162
    - aliases (definition) . . . . . 104
    - and Unicode algorithms . . . . . 99
    - data tables . . . . . 196
    - derived *see* derived properties
    - in Unicode Character Database (UCD) . . . . . 46
    - informative *see* informative properties
    - normative references to . . . . . 77, 84
    - normative *see* normative properties
    - of control codes . . . . . 857
    - provisional *see* provisional properties
    - simple *see* simple properties
    - see also individual properties, e.g.* combining classes
  - property values
    - aliases . . . . . 162
    - aliases (definition) . . . . . 105
    - default . . . . . 97
    - default (definition) . . . . . 97
    - normative references to . . . . . 84
  - PropertyAliases.txt . . . . . 104, 912
  - PropertyValueAliases.txt . . . . . 105, 912
  - PropList.txt . . . . . 166
  - Provençal . . . . . 294
  - provisional properties
    - definition . . . . . 101
  - PS (U+2029 paragraph separator) . . . . . 209, 861
  - Psalter Pahlavi . . . . . 419
  - PUA (Private Use Area) . . . . . 50, 876
  - punctuation . . . . . 261–285
    - blocks containing . . . . . 255
    - CJK . . . . . 282
    - doubled . . . . . 273
    - in bidirectional text . . . . . 261
    - paired . . . . . 267
    - small form variants . . . . . 285
    - typographic forms . . . . . 261
    - vertical forms . . . . . 284
  - Punctuation and Symbols Area . . . . . 50
  - Punjabi . . . . . 475
- ## Q
- quotation marks . . . . . 268–271
    - East Asian . . . . . 270
    - European . . . . . 268
- ## R
- radicals, KangXi and other CJK . . . . . 711–712
  - radical-stroke index . . . . . 707
  - record separator (U+001E) . . . . . 857
  - recycling symbols . . . . . 841
  - references . . . . . 919
  - referencing . . . . . 84
    - properties . . . . . 77
    - Unicode algorithms . . . . . 78
    - Unicode Standard . . . . . 76
  - regional indicator symbols . . . . . 853
  - regular expressions . . . . . 214
    - and line breaking . . . . . 209
    - see also* UTS #18, Unicode Regular Expressions
  - Rejang . . . . . 683
  - rendering of text . . . . . 6, 10, 17
    - fallback . . . . . 252
    - unsupported characters . . . . . 201
  - repertoire of abstract characters . . . . . 29
  - reph . . . . . 454, 458, 495
  - replacement character (U+FFFD) . . . . . 43, 68, 83, 885
  - reserved code points . . . . . 30, 201
    - definition . . . . . 93
    - in code charts . . . . . 895

- preservation in interchange ..... 31
    - see also* unassigned code points
  - Rhaeto-Romanic ..... 294
  - rich text ..... 18
  - right single quotation mark (U+2019)
    - preferred for apostrophe ..... 272
  - right-to-left text ..... 53
    - East Asian scripts ..... 692
    - Middle Eastern scripts ..... 359
  - roadmap for script additions ..... 46, 919
  - Roman numerals ..... 205, 821
  - Romanian ..... 294
    - comma below ..... 291
  - Romany ..... 294
  - Rong ..... 548–550
  - Rumi numeral symbols ..... 814
  - Runic ..... 348–350
  - Russian ..... 312
- S**
- Samaritan ..... 400–401
  - Sami ..... 294
  - Sanskrit ..... 443
  - Saurashtra ..... 551
  - scalar values, Unicode
    - see* Unicode scalar values
  - scripts
    - in Unicode Standard ..... 3
    - roadmap for future additions ..... 46, 919
    - types of ..... 260
    - see also* UAX #24, Unicode Script Property
  - SCSU
    - see* UTS #6, A Standard Compression Scheme for Unicode
  - searching ..... 230–232
    - as a text process ..... 10
    - case-insensitive ..... 231, 240
  - section or paragraph marks ..... 276
  - security issues ..... 245
  - self-synchronization of encoding forms ..... 34
  - semantics
    - see* character semantics
  - sequences
    - notation ..... 910
  - Serbian
    - corresponding digraphs in Croatian ..... 294
  - Shan ..... 649
    - digits ..... 808
  - Sharada ..... 594–595
  - Shavian ..... 357, 733
  - Show Hidden ..... 81, 222, 252, 873
  - SHY (U+00AD soft hyphen) ..... 860
  - Sibe ..... 529
  - Siddham ..... 598–599
  - signature for Unicode data ..... 67, 881–883
  - simple properties
    - definition ..... 104
  - simplified Chinese ..... 699
  - Sindhi ..... 367, 463
  - Sinhala ..... 513–514
  - Sinological dot ..... 301
  - SIP (Supplementary Ideographic Plane) ..... 44, 52
  - Siyaq Numbers ..... 814
    - Indic ..... 814
  - slash, fraction (U+2044) ..... 273
  - Slovak ..... 294
  - Slovenian ..... 294
  - small letters ..... 164, 236, 287
  - SMP (Supplementary Multilingual Plane) ..... 44, 51
  - soft hyphen (U+00AD) (SHY) ..... 860
  - Sogdian ..... 586
  - Somali ..... 746
  - Sora Sompeng ..... 618
  - Sorbian ..... 294
  - sorting ..... 12, 230
    - and combining grapheme joiner ..... 866
    - as a text process ..... 10
    - case-insensitive ..... 231
    - culturally expected ..... 12, 230
    - language-insensitive ..... 230
    - see also* Unicode Collation Algorithm (UCA)
  - source separation rule ..... 696, 702
  - South and Central Asian scripts
    - Ancient ..... 557–586
    - Other historic ..... 587–620
    - Other modern ..... 509–555
  - South Asian scripts ..... 441–540
  - Southeast Asian scripts ..... 621–666
  - Soyombo ..... 582–583
  - space (U+0020)
    - base for diacritic in isolation ..... 60, 265, 327
  - space characters ..... 264, 859–861
    - graphics for ..... 830
  - space, zero width (U+200B) ..... 264
  - spacing clones of diacritics ..... 324, 328
  - spacing marks ..... 326
    - definition ..... 108
  - Spacing Modifier Letters ..... 323–325
  - Spanish ..... 293
  - special characters ..... 67, 855–890
  - SpecialCasing.txt ..... 152, 166
  - Specials ..... 881–885
  - spell-checking
    - as a text process ..... 11
  - spellings, alternative
    - see* equivalent sequences
  - spoofing ..... 245

SSP (Supplementary Special-purpose Plane) . . . . . 45

stability . . . . . 102, 161

    as Unicode design principle . . . . . 23

stacked boundaries . . . . . 217

stacking sequences . . . . . 57

    nondefault . . . . . 58

standardized variants . . . . . 532, 872

    in the code charts . . . . . 902

    mathematical symbols . . . . . 828

StandardizedVariants.txt . . . . . 532, 828

standards coverage . . . . . 3

starters . . . . . 136

stateful encoding

    not used in Unicode . . . . . 4

    paired format controls . . . . . 868

string comparison . . . . . 12

string literals, Unicode

    code point notation `\u1234` . . . . . 912

strings, Unicode . . . . . 43, 121

    null termination . . . . . 858

strong directional characters . . . . . 171

styled text . . . . . 18

sublinear searching . . . . . 232

subsets, supported . . . . . 71

    conformance . . . . . 80

    ISO/IEC 10646 specification for . . . . . 932

substitution character

*see* replacement character

Sumero-Akkadian . . . . . 426–429

Sundanese . . . . . 686–687

superscripts . . . . . 324

    and subscripts . . . . . 819

supplementary characters

    in UTF-16 strings . . . . . 43

    tables for . . . . . 197

Supplementary General Scripts Area . . . . . 50

Supplementary Ideographic Plane (SIP) . . . . . 44, 52

Supplementary Multilingual Plane (SMP) . . . . . 44, 51

supplementary planes

    representation in UTF-16 . . . . . 36

    representation in UTF-8 . . . . . 37

Supplementary Private Use Areas . . . . . 52, 877

Supplementary Special-purpose Plane (SSP) . . . . . 45

supported subsets . . . . . 71

    conformance . . . . . 80

supralineation . . . . . 310

surrogate code points

*see* surrogates

surrogate pairs . . . . . 36, 125

    definition . . . . . 119

    processing . . . . . 38, 203–204

surrogates . . . . . 31, 119, 878

    interchange restrictions . . . . . 31

    isolated surrogates, handling . . . . . 43

    isolated surrogates, ill-formed . . . . . 125

    isolated surrogates, uninterpreted . . . . . 119

    support levels . . . . . 203

Surrogates Area . . . . . 50, 878

Sutton SignWriting . . . . . 790–791

Suzhou-style numerals . . . . . 816

svasti signs . . . . . 524

Swahili . . . . . 293

Swedish . . . . . 293

syllabaries . . . . . 257

    alphabetic property . . . . . 188

    featural . . . . . 257

Syloti Nagri . . . . . 589–590

symbols . . . . . 793–854

    animal . . . . . 842

    appearance variation . . . . . 793

    arrows . . . . . 827–828

    box drawing . . . . . 835

    cultural . . . . . 842

    currency symbols block . . . . . 795–798

    dictionary . . . . . 840

    dingbats . . . . . 843–844

    emoji . . . . . 838, 853

    Enclosed Alphanumerics . . . . . 852

    fragments for mathematical typesetting . . . . . 831

    game . . . . . 841

    gender . . . . . 841

    genealogical . . . . . 841

    geometrical . . . . . 835–837

    hand . . . . . 842

    Khmer lunar calendar . . . . . 648

    letterlike . . . . . 799–805

    map . . . . . 840

    mathematical . . . . . 821–828

    mathematical alphanumeric . . . . . 800–805

    miscellaneous . . . . . 839

    musical . . . . . 778–787

    numerals . . . . . 806–818

    recycling . . . . . 841

    regional indicator . . . . . 853

    technical . . . . . 830–834

    weather . . . . . 840

    zodiacal . . . . . 842

symmetric swapping format characters . . . . . 870

Syriac . . . . . 391–399

## T

tab (U+0009 character tabulation) . . . . . 857

tab, vertical (U+000B) . . . . . 209, 857

tables of character data . . . . . 196–197

    optimization . . . . . 197

    supplementary characters . . . . . 197

tag characters . . . . . 886–890



- Tagalog ..... 668
  - Tagbanwa ..... 668
  - tags, language ..... 215, 886–890
    - use strongly discouraged ..... 889
  - Tai Laing
    - digits ..... 808
  - Tai Le ..... 649–650
  - Tai Tham ..... 654–656
    - digits ..... 808
  - Tai Viet ..... 657–659
  - Tai Xuan Jing symbols ..... 847
  - Takri ..... 596–597
  - Tamil ..... 485–493
  - Tangut ..... 738–740
    - components ..... 739–740
    - radicals ..... 739
  - tashkil ..... 368
  - tashkil, harakat, points ..... 370
  - TCHAR in Win32 API ..... 200
  - Technical Reports (UTR) ..... 917
  - Technical Standards (UTS) ..... xxvi, 917
    - abstracts ..... 918
  - technical symbols ..... 830–834
  - Telugu ..... 494–496
  - terminal emulation ..... 794
  - text boundaries ..... 61, 189, 217–218, 228
    - see also UAX #14, Unicode Line Breaking Algorithm
    - see also UAX #29, Unicode Text Boundaries
  - text elements ..... 6, 10, 217
    - boundaries ..... 228
    - for sorting ..... 230
    - variable-width nature ..... 38
  - text processes ..... 6, 10–13
  - text rendering ..... 6, 10, 17
  - text selection, boundaries for ..... 217–218
  - Thaana ..... 511–512
  - Thai ..... 623–626
  - Tibetan ..... 517–527
  - Tifinagh ..... 747
  - Tigre ..... 742
  - tilde (U+007E) ..... 276
  - Tirhuta ..... 607–609
  - titlecase ..... 164, 236
  - Todo ..... 529
  - tone letters ..... 324–325
  - tone marks
    - Bopomofo spacing ..... 717, 718
    - Chinantec ..... 325
    - Chinese ..... 325
    - Tai Le ..... 649
    - Thai ..... 623
    - Vietnamese ..... 292
  - traditional Chinese ..... 699
  - traffic signs ..... 840
  - trailing surrogates
    - see low-surrogate code units
  - transcoding ..... 196–197
    - tables ..... 196
  - Transport and Map Symbols ..... 843
  - triangulation in transcoding ..... 196
  - tries ..... 196
  - truncation
    - combining character sequences ..... 220–221
    - surrogates and ..... 204
  - Turkish ..... 294
    - case mapping of İ ..... 238, 291
    - cedilla ..... 291
    - lira sign ..... 797
  - two-stage tables ..... 197
- ## U
- U+ notation ..... 912
  - U+10FFFF (not a character code) ..... 879
  - U+FEFF (BOM) ..... 881–883
  - U+FFFE (not a character code) ..... 880
  - U+FFFF (not a character code) ..... 879
  - UAX (Unicode Standard Annex) ..... xxiv, 917
    - as component of Unicode Standard ..... 79
    - conformance ..... 85
    - list of ..... 85
  - UCA see Unicode Collation Algorithm and see also
    - UTS #10, Unicode Collation Algorithm
  - UCD see Unicode Character Database
  - UCS (Universal Character Set)
    - see ISO/IEC 10646
  - UCS-2 ..... 929
  - UCS-4 ..... 929
  - Ugaritic ..... 430
  - Ukrainian ..... 312
  - unassigned code points ..... 30, 79, 201
    - defined as reserved code points ..... 93
    - handling ..... 74
    - properties of ..... 97
    - semantics ..... 79
    - see also reserved code points
  - underscores ..... 273
  - undesignated code points ..... 30
  - Unicode 1.0 Name (informative property) ..... 187
  - Unicode algorithms
    - and properties ..... 99
    - conformance ..... 84
    - definition ..... 93
    - normative references to ..... 78, 84
  - Unicode Bidirectional Algorithm ..... 21, 53
    - see also UAX #9, Unicode Bidirectional Algorithm
  - Unicode Character Database (UCD) .. xxiv, 161, 919
    - as component of Unicode Standard ..... 79

- changes ..... 74
- properties in ..... 46
- Unicode character encoding model ..... 33, 42
  - see also* UTR #17, Unicode Character Encoding Model
- Unicode character literals
  - code point notation U+ ..... 912
- Unicode codespace
  - allocation numbers ..... 936
  - definition ..... 90
  - planes ..... 44
  - size ..... 1, 29
- Unicode Collation Algorithm (UCA) ..... 12
- Unicode conferences ..... 918
- Unicode Consortium ..... 916
  - addresses ..... 920
  - Consortium membership in standards bodies 916
  - e-mail discussion list ..... 918
  - membership ..... 916
  - policies ..... 919
  - website ..... 918
- Unicode data signature ..... 67, 881–883
- Unicode data types ..... 199–200
  - for C ..... 199–200
- Unicode encoding forms ..... 120–127
  - advantages of each ..... 38
  - conformance ..... 34, 82
  - definition ..... 121
  - fixed-width (UTF-32) ..... 35, 124
  - signatures ..... 882, 883
  - variable-width ..... 36, 125
  - see also* encoding forms
- Unicode encoding schemes
  - conformance ..... 130–133
  - definition ..... 130
  - endian ordering ..... 40
  - see also* encoding schemes
- Unicode escape sequence notation `\u1234` ..... 912
- Unicode scalar values
  - definition ..... 120
- Unicode security ..... 245
  - see also* UTS #39, Unicode Security Mechanisms
- Unicode Standard
  - allocation of encoded characters ..... 44–52
  - architecture ..... 10–13
  - areas ..... 45
  - benefits ..... 1
  - blocks ..... 255
  - code charts ..... 891–908
  - components ..... 79
  - conformance ..... 73–158
  - conformance of ISO/IEC 10646 implementations ..... 934
  - corrections ..... 76
  - definitions for conformance ..... 87–93
  - design goals ..... 4
  - design principles ..... 14–24
  - errata ..... 76, 919
  - normative references to ..... 76, 84
  - number of characters ..... 3
  - number of code points ..... 1, 29
  - script coverage ..... 3
  - security issues ..... 245
  - synchrony with ISO/IEC 10646 ..... 931
  - updates ..... 919
  - versions *see* versions of the Unicode Standard
  - see also* Version 11.0
- Unicode Standard Annexes (UAX) ..... xxiv, 917
  - as components of Unicode Standard ..... 79
  - conformance ..... 85
  - list of ..... 85
- Unicode string literals
  - code point notation `\u1234` ..... 912
- Unicode strings ..... 43
  - definition ..... 121
- Unicode Technical Committee (UTC) ..... 916
- Unicode Technical Reports (UTR) ..... 917
- Unicode Technical Standards (UTS) ..... xxvi, 917
  - abstracts ..... 918
- UnicodeData.txt ..... 152, 166
- unification
  - as Unicode design principle ..... 21
  - see also* Han unification
- Unified Repertoire and Ordering (URO) ... 703, 943
  - see also* Han unification
- Unihan Database ..... 161, 707, 708, 919, 944
- Unihan.zip ..... 102, 161
- unit separator (U+001F) ..... 857
- Universal Character Set (UCS)
  - see* ISO/IEC 10646
- universality
  - as Unicode design principle ..... 14
- Unix
  - and UTFs ..... 38
  - newline function ..... 210
  - UTF-32 in ..... 35
  - UTF-8 in ..... 18
- unsupported characters ..... 201
- upadhmaniya ..... 467, 594
- update version ..... 75
- uppercase ..... 164, 236, 287
- Uralic Phonetic Alphabet (UPA) ..... 276, 298
- Urdu ..... 367
- URO (Unified Repertoire and Ordering) ... 703, 943
  - see also* Han unification
- UTF, Unicode Transformation Formats ..... 33, 121
  - advantages of each ..... 38
  - as encoding form or scheme ..... 133

binary comparison and sort order differences  
 ..... 231, 233  
 in APIs ..... 200  
 UTF-16 ..... 36, 125, 930  
 binary comparison and sort order caution .... 36  
 bit distribution (table) ..... 125  
 BOM in ..... 131, 881  
 encoding form (definition) ..... 125  
 encoding scheme (definition) ..... 131  
 encoding schemes ..... 40  
 in ISO/IEC 10646 ..... 930  
 in UTF-8 order ..... 234  
 surrogates and string handling ..... 43, 203  
 UTF-16BE (Big-endian) ..... 882  
 encoding scheme ..... 41  
 encoding scheme (definition) ..... 130  
 UTF-16LE (Little-endian) ..... 882  
 encoding scheme ..... 41  
 encoding scheme (definition) ..... 130  
 UTF-32 ..... 35, 124  
 as processing code ..... 38  
 BOM in ..... 132  
 encoding form (definition) ..... 124  
 encoding scheme (definition) ..... 132  
 encoding schemes ..... 40  
 in Unix ..... 35  
 UTF-32BE (Big-endian)  
 encoding scheme ..... 41  
 encoding scheme (definition) ..... 132  
 UTF-32LE (Little-endian)  
 encoding scheme ..... 41  
 encoding scheme (definition) ..... 132  
 UTF-8 ..... 36, 125, 930  
 ASCII transparency ..... 36  
 binary comparison and sort order ..... 39  
 bit distribution (table) ..... 126  
 BOM in ..... 130, 133, 882  
 byte ranges ..... 126  
 compared to multibyte encodings ..... 37  
 encoding form (definition) ..... 125  
 encoding scheme ..... 40  
 encoding scheme (definition) ..... 130  
 in Unix ..... 18  
 in UTF-16 order ..... 233  
 non-shortest form is invalid ..... 125, 245  
 preferred encoding for Internet protocols .... 37  
 security and ..... 245  
 signature ..... 130, 133, 882  
 UTR (Unicode Technical Report) ..... 917  
 UTS (Unicode Technical Standard) ..... xxvi, 917  
 abstracts ..... 918  
 Uyghur ..... 367, 571

## V

Vai ..... 755–756  
 valid (synonym for well-formed) ..... 123  
 variable-width Unicode encoding form ..... 36, 125  
 variants  
 compatibility ..... 26  
 fullwidth and halfwidth ..... 285  
 mathematical symbols ..... 828  
 small form ..... 285  
 standardized ..... 872  
 variation selectors ..... 193, 872  
 ideographic variation mark (U+303E) ..... 715  
 Mongolian free variation selectors ..... 532  
 variation sequences ..... 872  
 for Phags-pa ..... 575–577  
 Version 11.0 ..... 79  
 number of characters ..... 3  
 versions of the Unicode Standard .. xxiv, 74, 919, 935–937  
 backward compatibility ..... 74  
 compared to ISO/IEC 10646 editions ..... 935  
 content ..... 75  
 interaction in implementations ..... 201  
 numbering ..... 75  
 property changes ..... 74  
 stability ..... 74  
 updates ..... 919  
 vertical tab (U+000B) ..... 209, 857  
 vertical text ..... 53, 262, 284  
 East Asian scripts ..... 692  
 Mongolian ..... 529  
 Vietnamese ..... 292, 299  
 ideographs ..... 692  
 virama ..... 258, 441  
 definition ..... 446  
 Kharoshthi ..... 567  
 Khmer ..... 640  
 Myanmar ..... 631  
 Philippine scripts ..... 668  
 virama-like characters ..... 191  
 visual order used for Thai and Lao ..... 21  
 vowel harmony  
 Mongolian ..... 533  
 vowel marks, Middle Eastern scripts ..... 359  
 vowel separator  
 Mongolian ..... 535  
 vowel signs  
 Indic ..... 56, 445  
 Khmer ..... 642  
 Philippine scripts ..... 668

**W**

- Warang Citi ..... 544
- wchar\_t
  - and Unicode encoding forms ..... 38
  - in C language ..... 200
- weak directional characters ..... 171
- weather symbols ..... 840
- website, Unicode Consortium ..... 918
- Weierstrass elliptic function symbol ..... 800
- well-formed
  - definition ..... 122
- Welsh ..... 294
- Where Is My Character? ..... 920
- wide characters
  - data type in C ..... 200
- wiggly fence (U+29DB) ..... 826
- Windows newline function ..... 210
- word breaks ..... 219, 859–861
  - in South Asian scripts ..... 625, 633, 648
- word joiner (U+2060) ..... 859
- writing direction *see* directionality
- writing systems ..... 256–260
- Wu (Shanghainese) ..... 700

**X**

- Xibe ..... 529
- Xishuangbanna Dai ..... 651

**Y**

- Yi ..... 729–731
- Yiddish ..... 361
- Yijing Hexagram Symbols ..... 847
- ypogegrammeni ..... 303

**Z**

- Zanabazar Square ..... 579–581
- Zapf Dingbats ..... 843
- zero extension relation among encodings ..... 929
- zero width joiner (U+200D) ..... 369–370, 862
- zero width no-break space (U+FEFF) ... 67, 83, 859
  - initial ..... 133, 882
- zero width non-joiner (U+200C) ..... 369–370, 863
- zero width space (U+200B) ..... 860
  - for word breaks in South Asian scripts . 625, 633, 648
- zero-width space characters ..... 860
- ZWJ *see* zero width joiner (U+200D)
- ZWNBSpace *see* zero width no-break space (U+FEFF)
- ZWNJ *see* zero width non-joiner (U+200C)
- ZWSP *see* zero width space (U+200B)

