

The Unicode Standard

Version 7.0 – 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 © 1991–2014 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 7.0

Includes bibliographical references and index.

ISBN 978-1-936213-09-2 (<http://www.unicode.org/versions/Unicode7.0.0/>)

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

QA268.U545 2014

ISBN 978-1-936213-09-2

Published in Mountain View, CA

October 2014

I General Index

The General 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.6, Other Unicode Online Resources.*)

A

- abbreviation, Coptic 308
 - abjads 256, 353
 - abstract character sequences
 - definition 90
 - abstract characters 29
 - definition 90
 - abugidas 257, 258, 431, 587
 - accent marks *see* diacritics
 - accented characters
 - encoding 12
 - Latin 287
 - normalization 204
 - accounting numbers, ideographic 178
 - acrophonic numerals 203, 305
 - Aegean numbers 336
 - Africa
 - scripts of 693–713
 - Afrikaans 292
 - Ainu 679
 - Aiton 602
 - Alchemical Symbols 785
 - reference materials 910
 - Algonquian 717
 - Ali Gali 511
 - aliases
 - character name 88, 182, 836
 - property 162
 - property value 162
 - allocation areas 45
 - allocation of encoded characters 44–52
 - Alphabetic (informative property) 188
 - alphabets 256
 - European 285–331
 - mathematical 744–748
 - alternate format characters (deprecated) . . 191, 810–811
 - Americas
 - scripts of 715–721
 - Amharic 694
 - Ancient Symbols 788
 - angle brackets (U+2329 and U+232A)
 - deprecated for technical publication 772
 - Annexes, Unicode Standard (UAX) xxxiv, 853
 - as components of Unicode Standard 79
 - conformance 85
 - list of 85
 - annotation characters 823–825
 - use in plain text discouraged 824
 - ANSI/ISO C
 - wchar_t and Unicode 198
 - apostrophe (U+0027) 272
 - Arabic 361–382
 - digits 751
 - Arabic-Indic digits 365–366
 - signs used with 367
 - ArabicShaping.txt 369, 373, 388
 - Aramaic 404, 431, 511, 537, 542
 - areas of the Unicode Standard 45
 - ARIB 780
 - Armenian 314–315
 - arrows 767–768
 - ASCII
 - characters with multiple semantics 262
 - transparency of UTF-8 36
 - Unicode modeled on 1
 - zero extension 198, 868
 - Assamese 455
 - assigned code points 11, 30
 - Athapascan 717
 - atomic character boundaries 216
 - Avestan 412
 - reference materials 910
- ## B
- Balinese 637–642
 - reference materials 910

- Bamum 709–710
 reference materials 911
- Bangla 455–460
- base characters 322
 definition 105
 multiple 59
 ordered before combining marks 218, 322
- Basic Multilingual Plane (BMP) 1, 44
 allocation areas 49
 representation in UTF-16 36
- Basque 292
- Bassa Vah 711
 reference materials 911
- Batak 648
 reference materials 911
- benefits of Unicode 1
- Bengali 455–460
- Bidi Class (normative property) 173
- Bidi Mirrored (normative property) 180
- Bidi Mirroring Glyph (informative property) ... 180
- BidiMirroring.txt 180
- Bidirectional Algorithm, Unicode 53, 84
- bidirectional ordering 20
 controls 807
- bidirectional text 53, 84
 Middle Eastern scripts 353
 nonspacing marks in 221
 punctuation in 261
- big-endian 40
 definition 83
- Bihari 451
- binary comparison and sort order
 caution for UTF-16 36
 UTF differences 229, 231
 UTF-8 39
- blocks of the Unicode Standard 45, 255
- Blocks.txt 45
- BMP *see* Basic Multilingual Plane
- BNF (Backus-Naur Form) 847
- BOCU-1 *see* UTN #6, BOCU-1
 MIME-Compatible Unicode Compression
- Bodhi 499
- Bodo 450
- BOM (U+FEFF) 40, 67, 130–133, 821–823
- Bopomofo 675–677
- boundaries, text 61, 189, 215–216, 226
 see also UAX #14, Unicode Line Breaking Algorithm
 see also UAX #29, Unicode Text Segmentation
- boustrophedon 53, 345
- box drawing symbols 776
- Brahmi 431, 537, 538–541, 542, 589
 reference materials 911
- Braille 724–725
- Breton 292
- Buginese 635–636
- Buhid 632
- Bulgarian 310
- bullets 275
 numeric 752
- Burmese *see* Myanmar
- Byelorussian 310
- byte order mark (BOM) (U+FEFF) . 40, 67, 130–133,
 821– 823
- byte ordering
 changing 81
 conformance 83
- byte serialization 40, 67
- Byzantine Musical Symbols 731
- ## C
- C language
 wchar_t and Unicode 198
- C0 and C1 control codes 31, 187, 796
- Cambodian *see* Khmer
- camelcase 237
- Canadian Aboriginal Syllabics 717–718
 reference materials 912
- candrabindu 453, 564
- canonical composite characters
 see canonical decomposable characters
- canonical composition algorithm 138
- canonical decomposable characters
 definition 117
- canonical decomposition 63
 definition 116
 mappings 115
- canonical equivalence
 definition 117
 nonspacing marks 223
- canonical equivalent character sequences
 conformance 81
- canonical mappings
 see canonical decomposition mappings
- canonical ordering algorithm 137
- canonical precomposed characters
 see canonical decomposable characters
- Cantonese 659
- capital letters 164, 234, 285
- Carian 339
 reference materials 912
- carriage return (U+000D) (CR) 207, 797
- carriage return and line feed (CRLF) 207
- case 293
 and text processes 12
 beyond ASCII 235
 camelcase 237

- case folding 238
 - case operations (conformance) 85, 152–158
 - case operations and normalization 239
 - case operations, reversibility 237
 - cased (definition) 153
 - case-insensitive comparison 157, 229, 238
 - casing context (definition) 153
 - conversion 154
 - detection 156
 - European alphabets 285
 - exceptional Latin pairs 289, 293
 - Georgian 316
 - lowercase 164, 234, 285
 - mapping tables 194
 - mappings 152, 166, 234–236
 - mappings noted in code charts 837
 - titlecase 164, 234
 - Turkish I 236, 289
 - uppercase 164, 234, 285
 - see also* default case
 - Case (normative property) 164, 234
 - CaseFolding.txt 166, 238
 - caseless letters 293
 - Catalan 291
 - Caucasian Albanian 349
 - reference materials 912
 - cedilla 288
 - CEF *see* character encoding forms
 - CES *see* character encoding schemes
 - CESU-8
 - see* UTR #26, Compatibility Encoding Scheme for UTF-16: 8-Bit (CESU-8)
 - Chakma 530–531
 - reference materials 912
 - Cham 625–626
 - reference materials 912
 - character encoding forms (CEF) 33–39, 868
 - 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 859
 - character literals, Unicode
 - code point notation U+ 848
 - character mapping
 - interchange format *see* UTS #22, Character Mapping Markup Language (CharMapML)
 - character names 88, 181–186, 872
 - aliases 88, 182, 836
 - conventions 845
 - for CJK ideographs 841
 - for control codes 185, 187
 - in code charts 832–836
 - matching 181
- character properties
 - see* properties
 - see also* *individual properties*, e.g. Combining Class
 - character semantics 1, 80, 87–88, 873
 - as Unicode design principle 18
 - ASCII 262
 - definition 88
 - character sequences
 - abstract *see* abstract character sequences
 - canonical equivalent *see* canonical equivalent character sequences
 - compatibility equivalent *see* compatibility equivalent character sequences
 - conformance 81
 - named 182
 - character sequences, combining 105
 - character shaping selectors (deprecated) 810
 - character tabulation (U+0009) 797
 - characters
 - abstract *see* abstract characters
 - arrangement in Unicode 46
 - assigned 11, 30
 - blocks 45, 255
 - boundaries 215
 - canonical decomposable *see* canonical decomposable characters
 - classes 848
 - code charts 831–844, 859
 - coded *see* encoded characters
 - combining *see* combining characters
 - compatibility decomposable *see* compatibility decomposable characters
 - composite *see* decomposable characters
 - concept of 15, 60
 - conformance definitions 90–92
 - confusable 243
 - conversion 194–195
 - 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, 795–830
 - glyphs, relationship to 15
 - graphic 30
 - identity (definition) 87
 - ignored in processing 246–251
 - interpretation 80
 - layout control 68, 799–809

- modification 81
- names list 832–836
- names *see* character names
- not encoded in Unicode 3
- number encoded in Version 7.0 3
- precomposed *see* decomposable characters
- properties *see* properties
- semantics *see* character semantics
- special 67, 795–830
- supplementary *see* supplementary characters
- transcoding 194–195
- unsupported 199
- characters, not glyphs
 - in spoofing 244
 - Unicode principle 15
- CharMapML
 - see* UTS #22, Character Mapping Markup Language (CharMapML)
- charsets
 - IANA registered names 41
- charts, character code *see* code charts
- Cherokee 716
 - reference materials 912
- Chinese 658–659
 - Cantonese 659
 - Hakka 676
 - Mandarin 659
 - Minnan (Hokkien/Fujian, incl. Taiwanese) . . . 676
 - simplified and traditional 658
- Chu hán 657
- Chu Nôm 884
- citations for
 - properties 77
 - Unicode algorithms 77
 - Unicode Standard 76
- CJK ideographs 258, 653–668
 - accounting numbers 178
 - CJK Compatibility Ideographs 667–668
 - CJK Compatibility Supplement 668
 - CJK Strokes 670, 891
 - CJK Unified Ideographs 653–666
 - CJK Unified Ideographs Extension A 655
 - CJK Unified Ideographs Extension B 666
 - CJK Unified Ideographs Extension C 667
 - CJK Unified Ideographs Extension D 667
- code charts 841
- compatibility ideographs in Plane 2 52
- component structure 662
- encoding blocks 654
- ideographic description sequences 671–674
- ideographic variation mark (U+303E) 673
- KangXi radicals 665, 668–669
- names 841
- numbers 751
- numeric values 178, 203
- order of encoding 664
- radicals 668–669
- source standards 886–890
- unknown or unavailable 283
- Vietnamese 652
- CJK Miscellaneous Area 50
- CJK punctuation and symbols 281
 - compatibility forms 283
 - overscores and underscores 283
 - quotation marks 270
 - sesame dots 282
 - vertical forms 283
- CJK-JRG (Chinese/Japanese/Korean Joint Research Group) 882
- CJKV Ideographs Area 50
- CLDR (Unicode Common Locale Data Repository) . . . 860
- cluster boundaries 215
- code charts 831–844, 859
 - representative glyphs 832
- code point sequences
 - notation 846
- code points 7, 29
 - assigned 11, 30
 - assignment 46
 - categories 30
 - default ignorable 199, 250
 - definition 90
 - designated 30
 - notation 845
 - number in Unicode Standard 1
 - private-use *see* private-use code points
 - reserved *see* reserved code points
 - semantics 31
 - surrogate *see* surrogates
 - unassigned *see* unassigned code points
 - undesignated 30
- code positions *see* code points
- code set independence 18
- code unit sequences
 - definition 119
 - ill-formed (definition) 121
 - notation 846
 - well-formed (definition) 121
- code units
 - definition 119
 - isolated 118
- code values *see* code units
- coded character representations
 - see* coded character sequences
- coded character sequences
 - definition 91
- coded characters *see* encoded characters

- codespace *see* Unicode codespace
 coeng 603, 606
 Collation Algorithm, Unicode (UCA) 12
 collation *see* sorting
 collation tables 194
 combining character sequences 56, 105
 defective 221
 definition 107
 Latin 287
 line breaking 217
 matching 217
 order of base character and marks 218, 322
 rendering 217
 selection 215
 truncation 218–219
 combining characters 55–60, 109–114, 217–225
 blocking reordering 806
 canonical ordering 62, 137, 168
 class zero 169
 combining marks 322–323
 definition 105
 dependence 322
 display order 58
 keyboard input 218
 ligatures 59
 multiple 57
 multiple base characters 59
 normalization of 204
 ordering conventions 56
 rendering of marks 220–225
 reordrant 169
 script-specific 56
 split 170
 strikethrough 172
 subjoined 172
 typographical interaction 58, 168
 vertical stacking 58
 see also diacritics
 Combining Class (normative property) 168
 combining classes 135, 168, 223–224
 class zero characters 168
 definition 135
 combining grapheme joiner (U+034F) 805
 combining half marks 190, 330
 combining marks *see* combining characters
 comma below 288
 Compatibility and Specials Area 26, 50
 compatibility characters 22
 compatibility composite characters 27
 see compatibility decomposable characters
 compatibility decomposable characters 26
 definition 115
 compatibility decomposition 63
 definition 115
 compatibility decomposition mappings 115
 Compatibility Encoding Scheme for UTF-16
 see UTR #26, Compatibility Encoding Scheme for
 UTF-16: 8-Bit (CESU-8)
 compatibility equivalence
 definition 116
 compatibility equivalent character sequences
 conformance 81
 compatibility mappings
 see compatibility decomposition mappings
 compatibility precomposed characters
 see compatibility decomposable characters
 compatibility variants 26
 mapping 241
 composite characters
 see decomposable characters
 compatibility *see* compatibility decomposable
 characters
 Composition Exclusion (normative property) ... 99
 compression 206
 see also UTS #6, A Standard Compression Scheme
 for Unicode (SCSU)
 conferences 859
 conformance 73–158
 clause and definition updates 880
 definitions 87–92
 examples 69
 ISO/IEC 10646 implementations 873
 requirements 79–84
 confusables 243
 conjunct consonants
 Indic 215, 437
 Myanmar 597
 selection of clusters 215
 contextual shaping
 apostrophe 272
 Arabic 361
 not used for Hebrew final forms 356
 quotation marks 268
 Syriac 387
 contour tones 320
 control codes 31, 68, 796
 graphics for 771
 names 187
 properties 797
 semantics 31, 797
 specified in Unicode 797
 control sequences 796
 conversion of characters 127, 194–195, 252
 convertibility
 as Unicode design principle 23
 Coptic 303, 307–309
 reference materials 913
 Coptic Epat numbers 755

- corporate use subarea 816
 - corrigenda 76
 - CR (U+000D carriage return) 207, 797
 - CRLF (carriage return and line feed) 207
 - Croatian 292
 - digraphs 292
 - culturally expected sorting 12, 228
 - Cuneiform
 - Old Persian 423
 - Sumero-Akkadian 418–421
 - Ugaritic 422
 - Cuneiform and Hieroglyphic Area 51
 - Cuneiform and Hieroglyphs 417–429
 - currency symbols 739–741
 - encoded in script blocks 740
 - cursive joining 801–805
 - Arabic 368–375
 - control characters for 190, 363–364, 514, 800
 - Mandaic 395
 - Mongolian 513–515
 - N’Ko 705
 - Phags-pa 553
 - Syriac 387–390
 - transparency 804
 - cursive scripts 353
 - Cypriot 338
 - reference materials 919
 - see also* Linear B
 - Cyrillic 310–312
 - Czech 292
- D**
- dammatan 365
 - danda, in Devanagari block 449
 - Danish 291
 - dashes 265
 - Database, Unicode Character
 - see* Unicode Character Database (UCD)
 - dead consonants, Indic 436
 - dead keys 218
 - decomposable characters 63
 - definition 115
 - normalization of 204
 - decomposition 63, 115–117
 - canonical *see* canonical decomposition
 - compatibility *see* compatibility decomposition
 - definition 115
 - in normalization 204
 - mapping, definition 115
 - mappings noted in code charts 838
 - default case
 - algorithms 85, 152–158
 - conversion 154
 - detection 156
 - folding 155
 - default caseless matching 157
 - default grapheme clusters 215
 - see also* UAX #29, Unicode Text Segmentation
 - Default Ignorable Code Point (property) 250
 - default ignorable code points 199, 250
 - default property values 96
 - definition 96
 - defective combining character sequences 221
 - definition 107
 - dependent vowel signs
 - Indic 435
 - Khmer 608
 - Philippine scripts 632
 - deprecated characters 74, 835
 - alternate format 191, 810–811
 - definition 91
 - Derived Age (property) 199
 - derived properties
 - definition 103
 - DerivedCoreProperties.txt 153, 164, 250
 - DerivedNormalizationProps.txt 240
 - Deseret 719–721
 - reference materials 913
 - design goals of Unicode 4
 - design principles of Unicode 14–24
 - designated code points 30
 - Devanagari 433–454
 - Dhivehi 495
 - diacritics 55, 322
 - alternative glyphs 287, 322
 - Czech 288
 - display in isolation 60, 265, 323
 - double 112, 190, 324
 - German dialectology 328
 - Greek 300–301, 304
 - Latin 287–290
 - Latvian 288
 - mathematical 747
 - on i and j 289
 - rendering 220–225
 - Slovak 288
 - spacing clones of 320, 324
 - symbol 55, 329
 - see also* combining characters
 - dictionary symbols 781
 - digit form names 365
 - digits 203
 - Arabic 751
 - Arabic-Indic 365–366
 - compatibility 751
 - decimal 177
 - glyph variants 753

- hexadecimal 751
 - Myanmar 751
 - national shapes 811
 - Shan 751
 - superscript and subscript 752
 - Tai Laing 751
 - Tai Tham 751
 - digraphs 292, 295, 297
 - dingbats 783–784
 - dingbats, ornamental 784
 - directionality 20, 53
 - East Asian scripts 652
 - Middle Eastern scripts 353
 - Mongolian 512
 - musical symbols 727
 - normative property 173
 - Ogham 351
 - Old Italic 342
 - Philippine scripts 633
 - Runic 345
 - discussion list for Unicode 859
 - Dogri 450
 - dollar 740
 - Domino Tiles 786
 - dotless i 236, 289
 - dotted circle
 - in code charts 106, 323
 - in fallback rendering 220
 - to indicate diacritic 55
 - to indicate vowel sign placement 56
 - double diacritics 112, 190, 324
 - Duployan 735
 - reference materials 913
 - Dutch 291, 292
 - dynamic composition
 - as Unicode design principle 22
 - Dzongkha 499
- E**
- East Asian scripts 651–692
 - writing direction 53
 - see also* CJK ideographs
 - Eastern Arabic-Indic digits 365
 - EBCDIC
 - newline function 208
 - see* UTR #16, UTF-EBCDIC
 - editing, text boundaries for 215–216
 - efficiency
 - as Unicode design principle 15
 - Egyptian hieroglyphs 424–427
 - reference materials 914
 - Elbasan 348
 - reference materials 914
 - e-mail discussion list for Unicode 859
 - emoji 779
 - animal symbols 782
 - cultural symbols 782
 - zodiacal symbols 782
 - Emoticons 783
 - Enclosed Alphanumerics 792
 - enclosing marks 329
 - definition 106
 - encoded characters 7, 29
 - allocation 44–52
 - definition 90
 - encoding form conversion
 - definition 126
 - encoding forms 33–39
 - ISO/IEC 10646 definitions 868
 - 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 817
 - English 291
 - equivalent sequences 204
 - as Unicode design principle 23
 - case-insensitivity 229, 238
 - combining characters in matching 217
 - conformance 82
 - Hangul syllables 683
 - in sorting and searching 228
 - language-specific 117
 - security implications 243
 - see also* canonical equivalence
 - see also* compatibility equivalence
 - see also* encoding forms, encoding schemes
 - errata xxxvi, 76, 860
 - escape sequences 797
 - not used in Unicode 1, 4
 - Esperanto 292
 - Estonian 292
 - Ethiopic 694–697
 - reference materials 914
 - Etruscan 341
 - euro sign (U+20AC) 741
 - European scripts 285–331
 - ancient 333–352
 - eyelash-RA 442

F

- fallback rendering 250
 - of nonspacing marks 220
- FAQ (Frequently Asked Questions) 859
- Faroese 291
- Farsi 361, 364
- featural syllabaries 257
- FF (U+000C form feed) 207, 797
- file separator (U+001C) 797
- Finnish 291
- Finno-Ugric Transcription (FUT)
 - see Uralic Phonetic Alphabet (UPA)
- fixed-width Unicode encoding form (UTF-32) 35, 123
- flat tables 194
- Flemish 291
- fleurons 784
- fonts
 - and Unicode characters 16
 - for mathematical alphabets 746–748
 - style variation for symbols 737
- form feed (U+000C) (FF) 207, 797
- format control characters 30, 68, 263, 795–830
 - deprecated 810–811
 - prefixed 191
 - stateful 808
- fraction characters 762
- fraction slash (U+2044) 273, 758
- French 292
- Frisian 292
- FTP site, Unicode Consortium 859
- fullwidth forms in East Asian encodings 680
- futhark 344
- Latin alternative 287
- mathematical alternative 763
- missing 250
- representative in code charts 832
- standardized variants 812
- symbols alternative 737
- golden numbers 346
- Gothic 347
 - reference materials 915
- Grantha 582–584
 - reference materials 915
- grapheme base 322
 - definition 107
- grapheme clusters 11, 60–61
 - see also UAX #29, Unicode Text Segmentation
 - default 215
 - definition 108
- grapheme extender
 - definition 108
- grapheme joiner, combining (U+034F) 805
- graphic characters 30
- Greek 300–305
 - acrophonic numerals 203, 305
 - alternative glyphs 301–303
 - ancient musical notation 732–734
 - editorial marks 278, 915
 - letters as symbols 301–303, 765
 - see also Cypriot, Linear B
- Greenlandic 292
- group separator (U+001D) 797
- guillemets 268
- Gujarati 466
- Gurmukhi 461–465

G

- Garshuni 383
- Ge'ez 694
- General Category (normative property) 174
 - list of values 174
- general punctuation 261–284
- General Scripts Area 50
- geometrical symbols 776–778
- Georgian 316–317
- German 291
- geta mark (U+3013) 283
- Glagolitic 313
 - reference materials 914
- Glossary 859
- glyph selection tables 194
- glyphs 6, 15
 - characters, relationship to 15
 - diacritics alternative 287, 322
 - Greek alternative 301–303
 - Hakka 676
 - halant 431
 - see also virama
 - half marks, combining 190, 330
 - half-consonants, Indic 438
 - halfwidth forms in East Asian encodings 680
 - Han ideographs see CJK ideographs
 - Han unification 660–666
 - and language tags 213
 - history 881–890
 - language usage 657
 - source separation rule 655, 661
 - source standards 886–890
 - hand symbols 782
 - Hangul Area 50
 - Hangul syllables 651, 681–684
 - and combining marks 113
 - as grapheme clusters 61

H

- canonical decomposition 144
 - collation 683
 - composition 146
 - conjoining jamo 142–151
 - equivalent sequences 683
 - Hangul Compatibility Jamo 682
 - Hangul Jamo 681–684
 - Hangul Syllables block 683–684
 - Johab set 683
 - name generation 147
 - normalization 682
 - standard 143
 - Hangzhou numerals 757
 - Hanja *see* CJK ideographs
 - Hanunóo 632
 - Hanzi *see* CJK ideographs
 - harakat 362
 - hasant 455
 - hash tables 195
 - Hebrew 355–360
 - hentaigana 679
 - hieroglyphs
 - Egyptian 424–427
 - Meroitic 428–429
 - high surrogate
 - definition 118
 - high-surrogate code points 79, 818
 - high-surrogate code units 118
 - higher-level protocols
 - definition 92
 - Hindi 433
 - Hiragana 678
 - horizontal tab (U+0009) 797
 - HTML newline function 208
 - Hungarian 292
 - hyphenation 800
 - as a text process 10
 - hyphens 265, 800
- I**
- I Ching symbols 787
 - IANA charset names 41
 - Icelandic 291
 - identifiers 227
 - see also* UAX #31, Unicode Identifier and Pattern Syntax
 - Ideographic (informative property) 188
 - ideographic description sequences 672
 - Ideographic Rapporteur Group (IRG) 884
 - Ideographic Variation Database *see* UTS #37, Unicode Ideographic Variation Database
 - ideographs *see also* CJK ideographs
 - IDNA *see* UTS #46, Unicode IDNA Compatibility Processing
 - IICore 655, 884
 - ill-formed
 - definition 121
 - Imperial Aramaic 404–405
 - reference materials 915
 - implementation guidelines 193–252
 - in a Unicode encoding form
 - definition 122
 - in-band mechanisms 830
 - India
 - Official scripts 431–492
 - Indian rupee sign (U+20B9) 741
 - Indic scripts 431–492
 - principles, in terms of Devanagari 434–441
 - relation to ISCII standard 433
 - Indonesia and Oceania
 - scripts of 631–650
 - Indonesian 291
 - industry character sets
 - covered in Unicode 3
 - information separators (U+001C..U+001F) 797
 - informative properties
 - definition 99
 - Inscriptional Pahlavi 410
 - Inscriptional Parthian 410
 - inside-out rule 220
 - interchange restrictions 31
 - International Phonetic Alphabet (IPA) 256, 294–295
 - reference materials 916
 - Spacing Modifier Letters 319
 - see also* phonetic alphabets
 - internationalization 18
 - Internationalization & Unicode Conference 859
 - Internet protocols
 - UTF-8 as preferred encoding 37
 - Inuktitut 717
 - invisible operators 770
 - iota subscript 301
 - IPA *see* International Phonetic Alphabet
 - IRG (Ideographic Rapporteur Group) 884
 - Irish 291, 351
 - ISCII standard and Unicode 433
 - ISO/IEC 10646 861–873
 - conformance of Unicode implementations 873
 - encoding forms 868
 - synchrony with Unicode Standard 870
 - timeline compared to Unicode versions 863
 - Italian 291
 - ITC Zapf Dingbats 783
 - IUC *see* Internationalization & Unicode Conference

- J**
- jamos *see* Hangul syllables
- Japanese 651
- Javanese 643–646
 reference materials 917
- Jawi 379
- jihvamuliya 454, 564
- Johab 683
- joiners 363
 combining grapheme joiner (U+034F) 805
 word joiner (U+2060) 799
 zero width joiner (U+200D) 363–364, 802
- justification 222
- K**
- Kaithi 561–563
 reference materials 917
- Kana (Hiragana and Katakana) 678–679
- Kanbun 668
- KangXi radicals 665, 668–669
- Kanji *see* CJK ideographs
- Kannada 483–486
- Kashmiri 451
- Katakana 678–679
- Kawi 637, 639
- Kayah Li 624
 reference materials 917
- KC (normalization form)
 see Normalization Form KC
- KD (normalization form)
 see Normalization Form KD
- keytop labels 771
- Khamti Shan 600
- Kharoshthi 542–543
 reference materials 917
- Khmer 603–613
 characters not recommended 610
 syllable components, order of 611
- Khojki 572–573
 reference materials 918
- Khudawadi 574–575
 reference materials 918
- killer 258
 Batak 648
 Brahmi 538
 Meetei Mayek 524
 Myanmar (asat) 598
 see also virama
- Konkani 450
- Korean Hangul *see* Hangul
- Kurdish 379
- L**
- Ladino 355
- language tags 213, 826–830
 and Han unification 213
 use strongly discouraged 829
- Lanna 616
- Lao 593–595
- last-resort glyphs 250
- Latin 287–299
 alternative glyphs 287
 Basic Latin 291
 encoding blocks 45
 IPA Extensions 294–295
 Latin Extended Additional 297–299
 Latin Extended-A 291
 Latin Extended-B 292–294
 Latin Extended-C 297
 Latin Extended-D 298
 Latin Extended-E 299
 Latin Ligatures 297
 Latin-1 Supplement 291
 Phonetic Extensions 296–299
- Latvian 292, 298
 cedilla 288
- layout control characters 68, 799–809
- leading surrogates
 see high-surrogate code units
- legibility criterion for plain text 19
- Lepcha 532–534
 reference materials 918
- letter spacing 800
- letterlike symbols 742–748
- LF (U+000A line feed) 207, 797
- ligatures 801–805
 Arabic 371–373
 combining characters on 59
 control characters for 190
 for nonspacing marks 224
 Latin 297
 selection 216
 Syriac 390
- Limbu 520–523
 reference materials 918
- line breaking 207–211, 799–801
 control characters 192
 in South Asian scripts 591, 599, 613
 recommendations 209
 see also UAX #14, Unicode Line Breaking Algorithm
- line feed (U+000A) (LF) 207, 797
- line separator (U+2028) (LS) 207, 801
- line tabulation (U+000B) (VT) 797

Linear A 335
 reference materials 919
 Linear B 336–337
 reference materials 919
 see also Cypriot
 linear boundaries 216
 lira sign, Turkish (U+20BA) 741
 Lisu 688–690
 reference materials 920
 Lithuanian 292
 little-endian 40
 definition 83
 Locale Data Markup Language
 see UTS #35, Unicode Locale Data Markup Language (LDML)
 logical order
 as Unicode design principle 19
 exceptions to 170
 logograph 258
 logosyllabaries 258
 low surrogate
 definition 118
 low-surrogate code points 79, 818
 low-surrogate code units 118
 lowercase 164, 234, 285
 LS (U+2028 line separator) 207, 801
 Lycian 339
 reference materials 920
 Lydian 339
 reference materials 921

M

MacOS newline function 208
 Mahajani 570–571
 reference materials 921
 Mahjong Tiles 785
 mail discussion list for Unicode 859
 Maithili 450
 major version 75
 Malay 291
 Malay, Patani 592
 Malayalam 487–492
 Maltese 292
 Manchu 512
 Mandaic 394–395
 reference materials 921
 Mandarin 659
 Manden 702
 Manichaean 406–409
 reference materials 921
 map symbols 781
 mapping tables *see* tables of character data
 Marathi 433, 442, 448

markup languages
 and Unicode conformance 830
 line breaking 207
 see also UTR #20, Unicode in XML and Other Markup Languages
 Mathematical (informative property) 762
 mathematical expression format characters 190
 see also UTR #25, Unicode Support for Mathematics
 mathematical symbols 762–769
 alphabets 744–748
 alphanumeric 743–748
 fonts 746–748
 format characters 770
 fragments for typesetting 772
 invisible operators 770
 operators 763–766
 reference materials 921
 standardized variants 769
 MathML 766
 matras 168, 435
 Meetei Mayek 524–525
 reference materials 922
 Mende Kikakui 712–713
 reference materials 922
 Meroitic
 cursive 428–429
 hieroglyphs 428–429
 reference materials 922
 Miao 691–692
 reference materials 922
 Middle Eastern scripts 353–496
 ancient 397–415
 Min 659
 Minnan (Hokkien/Fujian, incl. Taiwanese) 676
 minor version 75
 minus sign 765
 commercial (U+2052) 276
 mirrored property
 see Bidi Mirrored (normative property)
 mirroring of paired punctuation 267
 Miscellaneous Symbols 780
 missing glyphs 250
 Modi 579–581
 reference materials 923
 modifier letters 318–321
 Modifier Letters, Spacing 296
 Mongolian 511–519, 548
 writing direction 512
 Mro 526
 reference materials 923
 multibyte encodings
 compared to UTF-8 37
 multistage tables 194

- musical symbols 726–734
 - ancient Greek 732–734
 - Balinese 641
 - Byzantine 731
 - directionality 727
 - Gregorian 727
 - reference materials 923
 - Western 726–730
- Myanmar 596–602
 - digits 751
 - Myanmar Extended-A 600
 - Myanmar Extended-B 600
 - reference materials 924
- N**
- N’Ko 702–706
 - reference materials 924
- Nabataean 414
 - reference materials 924
- named character sequences 182
- names, character *see* character names
- namespace 89
- NEL (U+0085 next line) 207, 797
- Nepali 433
- neutral directional characters 173
- New Tai Lue 616–617
- newline function (NLF) 208, 798
- newline guidelines 207–211
- next line (U+0085) (NEL) 207, 797
- NFC (Normalization Form C) 62
- NFD (Normalization Form D) 62
- NFKC (Normalization Form KC) 62
- NFKD (Normalization Form KD) 62
- NLF (newline function) 208, 798
- no-break space (U+00A0) 799
 - base for diacritic in isolation 60, 265, 323
- no-break space, narrow (U+202F) 517
- noncharacter code points *see* noncharacters
- noncharacters 31, 819
 - conformance 79
 - definition 92
 - handling 82
 - in code charts 835
 - interchange restrictions 31
 - semantics 32
 - U+10FFFF (not a character code) 819
 - U+FDD0..U+FDEF 31, 819
 - U+FFFE (not a character code) 67, 820
 - U+FFFF (not a character code) 31, 819
- nondecomposable characters 64
- non-joiner, zero width (U+200C) 363–364, 803
- nonlinear boundaries 216
- non-overlap principle in Unicode encoding forms 33
- nonspacing marks 322
 - definition 106
 - display in isolation 60, 265, 323
 - positioning 224
 - rendering 220–225
 - see also* combining characters
 - see also* diacritics
- normalization 62, 204–205
 - and case operations 239
 - canonical ordering algorithm 62, 137, 168
 - conformance 84
 - of private-use characters 816
 - 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 98
 - list 99
 - may change 98
- Norwegian 291
- notational conventions 845–849
- notational systems 259, 723–736
- nukta 362, 381, 443
- null (U+0000)
 - as Unicode string terminator 798
- number forms
 - CJK ideographs 203
- numbers
 - Coptic Epact 755
 - handling 203
 - ideographic accounting 178
- numerals 749–759
 - acrophonic 305
 - Chinese counting rods 760
 - Coptic 309
 - Cuneiform 421
 - Ethiopic 696
 - Greek acrophonic 203
 - Hangzhou 757
 - old-style 273
 - Roman 203, 762
 - Rumi 756
 - Suzhou-style 757
- numeric separators 275
- numeric shape selectors (deprecated) 811
- Numeric Type (normative property) 177

Numeric Value (normative property) 177
 numero sign (U+2116) 742

O

object replacement character (U+FFFC) 825
 octet 847
 Ogham 351
 reference materials 925
 Ol Chiki 528–529
 reference materials 925
 Old Italic 341–343
 reference materials 925
 Old North Arabian 399
 reference materials 925
 Old Permic 350
 reference materials 926
 Old Persian 423
 reference materials 926
 Old South Arabian 400–401
 reference materials 926
 Old Turkic 555
 reference materials 926
 old-style numerals 273
 Oriya 468–470
 ornamental dingbats 784
 Oromo 694
 Osmanya 698
 reference materials 927
 out-of-band mechanisms 830
 overlapping encodings 33
 overscores 273

P

Pahawh Hmong 627–628
 reference materials 927
 Pahlavi, Inscriptional 410
 reference materials 916
 Pahlavi, Psalter 411
 Palmyrene 415
 reference materials 927
 Panjabi 461
 paragraph or section marks 275
 paragraph separator (U+2029) (PS) 207, 801
 Parthian, Inscriptional 410
 reference materials 916
 Pashto 361
 Patani Malay 592
 Pau Cin Hau 629
 reference materials 928
 Persian 361, 364
 peso 740
 Phags-pa 548–554
 reference materials 928

Phaistos Disc symbols 788
 Phake 602
 Philippine scripts 632–634
 reference materials 928
 Phoenician 402
 reference materials 929
 phonemes 259
 phonetic alphabets 256
 IPA Extensions 294–295
 Phonetic Extensions 296–299
 Spacing Modifier Letters 319–321
 Uralic Phonetic Alphabet (UPA) 276, 296
 see also International Phonetic Alphabet (IPA)
 Pinyin 291
 pivot code, Unicode as 194
 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, 817
 Playing Cards 786
 points, Hebrew pronunciation marks 355
 policies of the Unicode Consortium 860
 Polish 292
 Portuguese 291
 precomposed characters
 see decomposable characters
 compatibility *see* compatibility decomposable
 characters
 prefixed format control characters 191
 Private Use Area (PUA) 50, 816
 Private Use planes 45, 52, 817
 private-use characters
 properties 815
 semantics 32
 private-use code points 31, 199
 conformance 80
 definition 104
 high surrogates 818
 properties 18, 94–104, 159–192
 aliases 162
 aliases (definition) 103
 and Unicode algorithms 98
 data tables 194
 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 797

- provisional *see* provisional properties
 simple *see* simple properties
see also individual properties, e.g. combining classes
- property values
 aliases 162
 aliases (definition) 104
 default 96
 default (definition) 96
 normative references to 84
 PropertyAliases.txt 103, 848
 PropertyValueAliases.txt 104, 848
 PropList.txt 166
 Provençal 292
 provisional properties
 definition 100
 PS (U+2029 paragraph separator) 207, 801
 Psalter Pahlavi 411
 reference materials 929
 PUA (Private Use Area) 50, 816
pulli 471
 punctuation 261–284
 blocks containing 255
 CJK 281
 doubled 273
 in bidirectional text 261
 paired 267
 small form variants 284
 typographic forms 261
 vertical forms 283
 Punctuation and Symbols Area 50
 Punjabi 461
- Q**
- quotation marks 268–271
 East Asian 270
 European 268
- R**
- radicals, KangXi and other CJK 668–669
 radical-stroke index 665
 record separator (U+001E) 797
 recycling symbols 781
 referencing 84
 properties 77
 Unicode algorithms 77
 Unicode Standard 76
 regional indicator symbols 793
 regular expressions 212
 and line breaking 207
 see also UTS #18, Unicode Regular Expressions
 Rejang 647
 reference materials 929
 rendering of text 6, 10, 17
 fallback 250
 unsupported characters 199
 repertoire of abstract characters 29
 replacement character (U+FFFD) 43, 68, 83, 127, 252, 825
 reserved code points 30, 199
 definition 92
 in code charts 835
 preservation in interchange 31
 see also unassigned code points
 Rhaeto-Romanic 292
 rich text 18
 right single quotation mark (U+2019)
 preferred for apostrophe 272
 right-to-left text 53
 East Asian scripts 652
 Middle Eastern scripts 353
 roadmap for script additions 46
 Roman numerals 203, 762
 Romanian 292
 comma below 289
 Romany 292
 Rong 532–534
 ruble sign 741
 Rumi numeral forms 756
 Runic 344–346
 reference materials 929
 rupee sign, Indian (U+20B9) 741
 Russian 310
- S**
- Samaritan 392–393
 reference materials 930
 Sami 292
 Sanskrit 433
 Saurashtra 535
 reference materials 930
 scalar values, Unicode
 see Unicode scalar values
 scripts
 in Unicode Standard 3
 roadmap for future additions 46
 types of 260
 see also UAX #24, Unicode Script Property
 SCSU
 see UTS #6, A Standard Compression Scheme for Unicode
 searching 228–230
 as a text process 10
 case-insensitive 229, 238
 section or paragraph marks 275
 security issues 243

- self-synchronization of encoding forms 34
- semantics
 - see* character semantics
- sequences
 - notation 846
- Serbian
 - corresponding digraphs in Croatian 292
- Shan 614
 - digits 751
- Sharada 564–565
 - reference materials 930
- Shavian 352, 688
 - reference materials 930
- Show Hidden 81, 220, 250, 813
- SHY (U+00AD soft hyphen) 800
- Sibe 512
- Siddham 568–569
 - reference materials 931
- signature for Unicode data 67, 821–823
- simple properties
 - definition 103
- simplified Chinese 658
- Sindhi 361, 450
- Sinhala 497–498
 - reference materials 931
- SIP (Supplementary Ideographic Plane) 44, 52
- slash, fraction (U+2044) 273
- Slovak 292
- Slovenian 292
- small letters 164, 234, 285
- SMP (Supplementary Multilingual Plane) 44, 51
- soft hyphen (U+00AD) (SHY) 800
- Somali 698
- Sora Sompeng 585
 - reference materials 931
- Sorbian 292
- sorting 12, 228
 - and combining grapheme joiner 806
 - as a text process 10
 - case-insensitive 229
 - culturally expected 12, 228
 - language-insensitive 228
 - see also* Unicode Collation Algorithm (UCA)
- source separation rule 655, 661
- South and Central Asian scripts
 - Ancient 537–555
 - Other historic 557–585
 - Other modern 493–535
- South Asian scripts 431–523
- Southeast Asian scripts 587–629
- space (U+0020)
 - base for diacritic in isolation 60, 265, 323
- space characters 264, 799–801
 - graphics for 771
- space, zero width (U+200B) 264
- spacing clones of diacritics 320, 324
- spacing marks 322
 - definition 106
- Spacing Modifier Letters 319–321
- Spanish 291
- special characters 67, 795–830
- SpecialCasing.txt 152, 166
- Specials 821–825
- spell-checking
 - as a text process 11
- spellings, alternative
 - see* equivalent sequences
- spoofing 243
- SSP (Supplementary Special-purpose Plane) 45
- stability 101, 161
 - as Unicode design principle 23
- stacked boundaries 215
- stacking sequences 57
 - nondefault 58
- Standard Compression Scheme for Unicode (SCSU)
 - see* UTS #6, A Standard Compression Scheme for Unicode
- standardized variants 515, 812
 - in the code charts 839
 - mathematical symbols 769
- StandardizedVariants.txt 515, 769
- standards coverage 3
- starters 136
- stateful encoding
 - not used in Unicode 4
 - paired format controls 808
- string comparison 12
- string literals, Unicode
 - code point notation `\u1234` 848
- strings, Unicode 43, 120
 - null termination 798
- strong directional characters 173
- styled text 18
- sublinear searching 229
- subsets, supported 71
 - conformance 80
 - ISO/IEC 10646 specification for 871
- substitution character
 - see* replacement character
- Sumero-Akkadian 418–421
- Sundanese 649–650
 - reference materials 932
- superscripts 320
 - and subscripts 760
- supplementary characters
 - in UTF-16 strings 43
 - tables for 195
- 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, 817
 - Supplementary Special-purpose Plane (SSP) 45
 - supported subsets 71
 - conformance 80
 - supralineation 308
 - surrogate code points
 - see* surrogates
 - surrogate pairs 36, 124
 - definition 118
 - processing 38, 201–202
 - surrogates 31, 118, 818
 - interchange restrictions 31
 - isolated surrogates, handling 43
 - isolated surrogates, ill-formed 124
 - isolated surrogates, uninterpreted 118
 - support levels 201
 - Surrogates Area 50, 818
 - Suzhou-style numerals 757
 - svasti signs 506
 - Swahili 291
 - Swedish 291
 - syllabaries 257
 - alphabetic property 188
 - featural 257
 - Syloiti Nagri 559–560
 - symbols 737–794
 - animal 782
 - appearance variation 737
 - arrows 767–768
 - box drawing 776
 - cultural 782
 - currency 739–741
 - dictionary 781
 - dingbats 783–784
 - emoji 779, 793
 - Enclosed Alphanumerics 792
 - fragments for mathematical typesetting 772
 - game 782
 - gender 781
 - genealogical 782
 - geometrical 776–778
 - hand 782
 - Khmer lunar calendar 613
 - letterlike 742–748
 - map 781
 - mathematical 762–769
 - mathematical alphanumeric 743–748
 - miscellaneous 780
 - musical 726–734
 - numerals 749–759
 - recycling 781
 - regional indicator 793
 - technical 771–775
 - weather 781
 - zodiacal 782
 - symmetric swapping format characters 810
 - Syriac 383–390
 - reference materials 932
- ## T
- tab (U+0009 character tabulation) 797
 - tab, vertical (U+000B) 207, 797
 - tables of character data 194–195
 - optimization 195
 - supplementary characters 195
 - tag characters 826–830
 - Tagalog 632
 - Tagbanwa 632
 - tags, language 213, 826–830
 - use strongly discouraged 829
 - Tai Laing
 - digits 751
 - Tai Le 614–615
 - reference materials 932
 - Tai Tham 618–620
 - digits 751
 - reference materials 932
 - Tai Viet 621–623
 - Tai Xuan Jing symbols 787
 - Takri 566–567
 - reference materials 933
 - Tamil 471–479
 - tashkil 362
 - tashkil, harakat, points 364
 - TCHAR in Win32 API 198
 - Technical Notes (UTN) 858
 - Technical Reports (UTR) 853
 - abstracts 856
 - Technical Standards (UTS) xxxvi, 853
 - abstracts 854
 - technical symbols 771–775
 - Telugu 480–482
 - terminal emulation 738
 - text boundaries 61, 189, 215–216, 226
 - see also* UAX #14, Unicode Line Breaking Algorithm
 - see also* UAX #29, Unicode Text Boundaries
 - text elements 6, 10, 215
 - boundaries 226
 - for sorting 228
 - variable-width nature 38
 - text processes 6, 10–13

- text rendering 6, 10, 17
 - text selection, boundaries for 215–216
 - Thaana 495–496
 - reference materials 933
 - Thai 589–592
 - Tibetan 499–510
 - Tifinagh 699
 - Tigre 694
 - tilde (U+007E) 276
 - Tirhuta 576–578
 - reference materials 933
 - titlecase 164, 234
 - Todo 512
 - tone letters 320–321
 - tone marks
 - Bopomofo spacing 675, 676
 - Chinantec 321
 - Chinese 321
 - Tai Le 614
 - Thai 589
 - Vietnamese 290
 - traditional Chinese 658
 - traffic signs 781
 - trailing surrogates
 - see* low-surrogate code units
 - transcoding 194–195
 - tables 194
 - Transport and Map Symbols 783
 - triangulation in transcoding 194
 - tries 194
 - truncation
 - combining character sequences 218–219
 - surrogates and 202
 - Turkish 292
 - case mapping of I 236, 289
 - cedilla 289
 - lira sign (U+20BA) 741
 - two-stage tables 195
- U**
- U+ notation 848
 - U+10FFFF (not a character code) 819
 - U+FEFF (BOM) 821–823
 - U+FFFE (not a character code) 820
 - U+FFFF (not a character code) 819
 - UAX (Unicode Standard Annex) xxxiv, 853
 - as component of Unicode Standard 79
 - conformance 85
 - list of 85
 - UCA *see* Unicode Collation Algorithm
 - UCD *see* Unicode Character Database
 - UCS (Universal Character Set)
 - see* ISO/IEC 10646
 - UCS-2 868
 - UCS-4 868
 - Ugaritic 422
 - reference materials 933
 - Uighur 511, 548
 - Ukrainian 310
 - unassigned code points 30, 79, 199
 - defined as reserved code points 92
 - handling 74
 - properties of 96
 - semantics 79
 - see also* reserved code points
 - underscores 273
 - undesignated code points 30
 - Unicode 1.0 Name (informative property) 187
 - Unicode algorithms
 - and properties 98
 - conformance 84
 - definition 92
 - normative references to 77, 84
 - Unicode Bidirectional Algorithm 20, 53
 - see also* UAX #9, Unicode Bidirectional Algorithm
 - Unicode Character Database (UCD) . . .xxxv, 161, 859
 - 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+ 848
 - Unicode codespace
 - allocation numbers 876
 - definition 90
 - planes 44
 - size 1, 29
 - Unicode Collation Algorithm (UCA) 12
 - see also* UTS #10, Unicode Collation Algorithm
 - Unicode Common Locale Data Repository (CLDR) . . . 860
 - Unicode conferences 859
 - Unicode Consortium 852
 - addresses 860
 - Consortium membership in standards bodies 852
 - e-mail discussion list 859
 - FTP site 859
 - membership 852
 - policies 860
 - website 859
 - Unicode data signature 67, 821–823
 - Unicode data types 197–198
 - for C 197–198
 - Unicode encoding forms 119–126
 - advantages of each 38

- conformance 34, 82
- definition 120
- fixed-width (UTF-32) 35, 123
- signatures 822, 823
- variable-width 36, 124
- see also* encoding forms
- Unicode encoding schemes
 - conformance 130–133
 - definition 130
 - endian ordering 40
 - see also* encoding schemes
- Unicode escape sequence notation `\u1234` 848
- Unicode Regular Expressions *see* UTS #18, Unicode Regular Expressions
- Unicode scalar values
 - definition 119
- Unicode security mechanisms
 - see also* UTS #39, Unicode Security Mechanisms
 - Unicode security 243
- Unicode Standard
 - allocation of encoded characters 44–52
 - architecture 10–13
 - areas 45
 - benefits 1
 - blocks 45, 255
 - code charts 831–844, 859
 - components 79
 - conformance 73–158
 - conformance of ISO/IEC 10646 implementations
 - 873
 - corrections 76
 - definitions for conformance 87–92
 - design goals 4
 - design principles 14–24
 - errata 76, 860
 - normative references to 76, 84
 - number of characters 3
 - number of code points 1, 29
 - script coverage 3
 - security issues 243
 - synchrony with ISO/IEC 10646 870
 - updates 860
 - versions *see* versions of the Unicode Standard
 - see also* Version 7.0
- Unicode Standard Annexes (UAX) xxxiv, 853
 - as components of Unicode Standard 79
 - conformance 85
 - list of 85
- Unicode string literals
 - code point notation `\u1234` 848
- Unicode strings 43
 - definition 120
- Unicode Technical Committee (UTC) 852
- Unicode Technical Notes (UTN) 858
- Unicode Technical Reports (UTR) 853
 - abstracts 856
- Unicode Technical Standards (UTS) xxxvi, 853
 - abstracts 854
- UnicodeData.txt 152, 166
- unification
 - as Unicode design principle 21
 - see also* Han unification
- Unified Repertoire and Ordering (URO) 661, 883
 - see also* Han unification
- Unihan Database 161, 665, 666, 841, 860, 884
- Unihan.zip 101, 161
- unit separator (U+001F) 797
- Universal Character Set (UCS)
 - see* ISO/IEC 10646
- universality
 - as Unicode design principle 14
- Unix
 - and UTFs 38
 - newline function 208
 - UTF-32 in 35
 - UTF-8 in 18
- unsupported characters 199
- upadhmaniya 454, 564
- update version 75
- uppercase 164, 234, 285
- Uralic Phonetic Alphabet (UPA) 276, 296
- Urdu 361
- URO (Unified Repertoire and Ordering) 661, 883
 - see also* Han unification
- UTF, Unicode Transformation Formats 33, 120
 - advantages of each 38
 - as encoding form or scheme 133
 - binary comparison and sort order differences ...
 - 229, 231
 - in APIs 198
- UTF-16 36, 124, 869
 - binary comparison and sort order caution 36
 - bit distribution (table) 124
 - BOM in 131, 821
 - encoding form (definition) 124
 - encoding scheme (definition) 131
 - encoding schemes 40
 - in ISO/IEC 10646 869
 - in UTF-8 order 232
 - surrogates and string handling 43, 201
- UTF-16BE (Big-endian) 822
 - encoding scheme 41
 - encoding scheme (definition) 130
- UTF-16LE (Little-endian) 822
 - encoding scheme 41
 - encoding scheme (definition) 130
- UTF-32 35, 123
 - as processing code 38

BOM in 132
 encoding form (definition) 123
 encoding scheme (definition) 132
 encoding schemes 40
 in Unix 35

UTF-32BE (Big-endian)
 encoding scheme 41
 encoding scheme (definition) 131

UTF-32LE (Little-endian)
 encoding scheme 41
 encoding scheme (definition) 132

UTF-8 36, 124, 869
 ASCII transparency 36
 binary comparison and sort order 39
 bit distribution (table) 125
 BOM in 130, 133, 822
 byte ranges 125
 compared to multibyte encodings 37
 encoding form (definition) 124
 encoding scheme 40
 encoding scheme (definition) 130
 in Unix 18
 in UTF-16 order 231
 non-shortest form is invalid 124, 243
 preferred encoding for Internet protocols 37
 security and 243
 signature 130, 133, 822

UTF-EBCDIC
see UTR #16, UTF-EBCDIC

UTN (Unicode Technical Note) 858

UTR (Unicode Technical Report) 853
 abstracts 856

UTS (Unicode Technical Standard) xxxvi, 853
 abstracts 854

Uyghur 361

V

Vai 707–708
 reference materials 933

valid (synonym for well-formed) 122

variable-width Unicode encoding form 36, 124

variants
 compatibility 26
 fullwidth and halfwidth 284
 mathematical symbols 769
 small form 284
 standardized 812

variation selectors 191, 812
 ideographic variation mark (U+303E) 673
 Mongolian free variation selectors 515

variation sequences 812
 for Phags-pa 552–554

Version 7.0 79
 number of characters 3

versions of the Unicode Standard xxxvi, 74, 860, 876–877
 backward compatibility 74
 compared to ISO/IEC 10646 editions 876
 content 75
 interaction in implementations 199
 numbering 75
 property changes 74
 stability 74
 updates 860

vertical tab (U+000B) 207, 797

vertical text 53, 262, 283
 East Asian scripts 652
 Mongolian 512

Vietnamese 290, 297
 ideographs 652

virama 258, 431
 definition 436
 Kharoshthi 546
 Khmer 606
 Myanmar 597
 Philippine scripts 632
 virama-like characters 191

visual order used for Thai and Lao 21

vowel harmony
 Mongolian 516

vowel marks, Middle Eastern scripts 353

vowel separator
 Mongolian 517

vowel signs
 Indic 56, 435
 Khmer 608
 Philippine scripts 632

W

Warang Citi 527
 reference materials 934

wchar_t
 and Unicode encoding forms 38
 in C language 198

weak directional characters 173

weather symbols 781

website, Unicode Consortium 859

Weierstrass elliptic function symbol 743

well-formed
 definition 121

Welsh 292

Where Is My Character? 860

wide characters
 data type in C 198

wiggly fence (U+29DB) 767

Windows newline function 208
 word breaks 217, 799–801
 in South Asian scripts 591, 599, 613
 word joiner (U+2060) 799
 writing direction *see* directionality
 writing systems 256–260
 Wu (Shanghainese) 659

X

Xibe 512
 Xishuang Banna Dai 616
 XML
 see UTR #20, Unicode in XML and Other Markup
 Languages

Y

yen currency sign 740
 Yi 685–687
 reference materials 934
 Yiddish 355
 Yijing Hexagram Symbols 787
 ypogegrammeni 301
 yuan currency sign 740

Z

Zapf Dingbats 783
 zero extension relation among encodings 868
 zero width joiner (U+200D) 363–364, 802
 zero width no-break space (U+FEFF) ... 67, 83, 799
 initial 133, 822
 zero width non-joiner (U+200C) 363–364, 803
 zero width space (U+200B) 800
 for word breaks in South Asian scripts .. 591, 599,
 613
 zero-width space characters 800
 ZWJ *see* zero width joiner (U+200D)
 ZWNBS *see* zero width no-break space (U+FEFF)
 ZWNJ *see* zero width non-joiner (U+200C)
 ZWSP *see* zero width space (U+200B)