

The Unicode® Standard

Version 10.0 – 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.

© 2017 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 10.0.

Includes bibliographical references and index.

ISBN 978-1-936213-16-0 (<http://www.unicode.org/versions/Unicode10.0.0/>)

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

QA268.U545 2017

ISBN 978-1-936213-16-0

Published in Mountain View, CA

June 2017

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 314
- abjads 260, 363
- abstract character sequences
 - definition 90
- abstract characters 29
 - definition 90
- abugidas 261, 262, 445, 615
- accent marks *see* diacritics
- accented characters
 - encoding 12
 - Latin 293
 - normalization 208
- accounting numbers, ideographic 178
- acrophonic numerals 207, 311
- Adlam 750–751
 - reference materials 950
- Aegean numbers 344
- Africa
 - scripts of 729–751
- Afrikaans 298
- Ahom 612–613
 - reference materials 950
- Ainu 709
- Aiton 630
- Alchemical Symbols 828
 - reference materials 950
- Algonquian 756
- Ali Gali 528
- aliases
 - character name 88, 183
 - informative 880
 - normative 881
 - property 164
 - property value 164
- allocation areas 45
- allocation of encoded characters 44–52
- Alphabetic (informative property) 190
- alphabets 260
 - European 291–339
 - mathematical 787–791
- alternate format characters (deprecated) . 194, 854–855
- Americas
 - scripts of 753–761
- Amharic 730
- Anatolian hieroglyphs 443–444
 - reference materials 950
- Ancient Symbols 831
- angle brackets (U+2329 and U+232A)
 - deprecated for technical publication 815
- Annexes, Unicode Standard (UAX) xxxiii, 901
 - as components of Unicode Standard 79
 - conformance 85
 - list of 85
- annotation characters 867–869
 - use in plain text discouraged 868
- ANSI/ISO C
 - wchar_t and Unicode 202
- apostrophe (U+0027) 276
- Arabic 371–393
 - digits 794
- Arabic-Indic digits 375–376
 - signs used with 377
- ArabicShaping.txt 379, 384, 399
- Aramaic 416, 445, 528, 555, 560
- areas of the Unicode Standard 45
- ARIB 824
- Armenian 322–323
- arrows 811–812
- ASCII
 - characters with multiple semantics 266
 - transparency of UTF-8 36
 - Unicode modeled on 1
 - zero extension 202, 913
- Assamese 469
- assigned code points 11, 30
- Athapascan 756

- atomic character boundaries 220
- Avestan 424
 - reference materials 951
- B**
- Balinese 665–670
 - reference materials 951
- Bamum 745–746
 - reference materials 951
- Bangla 469–474
- base characters 330
 - definition 106
 - multiple 59
 - ordered before combining marks 222, 330
- Basic Multilingual Plane (BMP) 1, 44
 - allocation areas 49
 - representation in UTF-16 36
- Basque 298
- Bassa Vah 747
 - reference materials 951
- Batak 676
 - reference materials 952
- benefits of Unicode 1
- Bengali 469–474
- Bhaiksuki 566–567
 - reference materials 952
- Bidi Class (normative property) 173
- Bidi Mirrored (normative property) 180
- Bidi Mirroring Glyph (informative property) 181
- BidiMirroring.txt 181
- Bidirectional Algorithm, Unicode 53, 84
- bidirectional ordering 20
 - controls 851
- bidirectional text 53, 84
 - Middle Eastern scripts 363
 - nonspacing marks in 225
 - punctuation in 265
- big-endian 40
 - definition 83
- Bihari 465
- binary comparison and sort order
 - caution for UTF-16 36
 - UTF differences 233, 235
 - UTF-8 39
- block 45, 90, 259, 875
 - headers 887
- BMP *see* Basic Multilingual Plane
- BNF (Backus-Naur Form) 895
- BOCU-1 *see* UTN #6, BOCU-1 MIME-Compatible Unicode Compression
- Bodhi 517
- Bodo 464
- BOM (U+FEFF) 40, 67, 131–134, 865–867
- Bopomofo 705–707
- boundaries, text 61, 191, 219–220, 230
 - see also* UAX #14, Unicode Line Breaking Algorithm
 - see also* UAX #29, Unicode Text Segmentation
- boustrophedon 53, 353
- box drawing symbols 819
- Brahmi 445, 555, 556–559, 560, 617
 - reference materials 952
- Braille 764–765
- Breton 298
- Buginese 663–664
- Buhid 660
- Bulgarian 316
- bullets 279
 - numeric 795
- Burmese *see* Myanmar
- Byelorussian 316
- byte order mark (BOM) (U+FEFF) 40, 67, 131–134, 865–867
- byte ordering
 - changing 81
 - conformance 83
- byte serialization 40, 67
- Byzantine Musical Symbols 771
- C**
- C language
 - wchar_t and Unicode 202
- C0 and C1 control codes 31, 189, 840
- Cambodian *see* Khmer
- Canadian Aboriginal Syllabics 756–757
 - reference materials 952
- candrabindu 467, 590
- canonical composite characters
 - see* canonical decomposable characters
- canonical composition algorithm 139
- canonical decomposable characters
 - definition 118
- canonical decomposition 63
 - definition 117
 - mappings 116
- canonical equivalence
 - definition 118
 - nonspacing marks 227
- canonical equivalent character sequences
 - conformance 81
- canonical mappings
 - see* canonical decomposition mappings
- canonical ordering algorithm 138
- canonical precomposed characters
 - see* canonical decomposable characters
- Cantonese 688

- capital letters 166, 238, 291
- Carian 347
 - reference materials 953
- carriage return (U+000D) (CR) 211, 841
- carriage return and line feed (CRLF) 211
- case 299
 - and text processes 12
 - beyond ASCII 239
 - camelcase 241
 - case folding 242
 - case operations (conformance) 85, 153–159
 - case operations and normalization 244
 - case operations, reversibility 241
 - cased (definition) 154
 - case-insensitive comparison 158, 233, 242
 - casing context (definition) 154
 - conversion 155
 - detection 157
 - European alphabets 291
 - exceptional Latin pairs 295, 299
 - Georgian 324
 - lowercase 166, 238, 291
 - mapping tables 198
 - mappings 153, 168, 238–240
 - mappings noted in code charts 884
 - titlecase 166, 238
 - Turkish I 240, 295
 - uppercase 166, 238, 291
 - see also* default case
- Case (normative property) 166, 238
- CaseFolding.txt 168, 242
- caseless letters 299
- Catalan 297
- Caucasian Albanian 358
 - reference materials 953
- cedilla 294
- CEF *see* character encoding forms
- CES *see* character encoding schemes
- Chakma 547
 - reference materials 953
- Cham 654–655
 - reference materials 953
- character encoding forms (CEF) 33–39, 913
 - 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 902
- character literals, Unicode
 - code point notation U+ 896
- character names 88, 182–188, 917
 - aliases 88, 183
 - conventions 893
 - for CJK ideographs 889
 - for control codes 187, 189
 - in code charts 880
 - matching 182
- character properties
 - see* properties
 - see also* individual properties, e.g. Combining Class
- character semantics 1, 80, 87–88, 918
 - as Unicode design principle 18
 - ASCII 266
 - 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 183
- character sequences, combining 106
- character shaping selectors (deprecated) 854
- character tabulation (U+0009) 841
- characters
 - abstract *see* abstract characters
 - arrangement in Unicode 46
 - assigned 11, 30
 - boundaries 219
 - canonical decomposable *see* canonical decomposable characters
 - classes 896
 - code charts 875–892, 902
 - 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 247
 - conversion 198–199
 - 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, 267, 839–855
 - glyphs, relationship to 15
 - graphic 30
 - identity (definition) 87
 - ignored in processing 250–255

- interpretation 80
- layout control 68, 843–853
- modification 81
- names list 876–888
- 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, 839–874
- supplementary *see* supplementary characters
- transcoding 198–199
- unsupported 203
- characters, not glyphs
 - in spoofing 248
 - Unicode principle 15
- charsets
 - IANA registered names 41
- charts, character code *see* code charts
- Cherokee 754
 - reference materials 953
- Chinese 687–689
 - Cantonese 688
 - Hakka 706
 - Mandarin 688
 - Minnan (Hokkien/Fujian, incl. Taiwanese) . 706
 - simplified and traditional 687
- Chu hán 686
- Chu Nôm 928
- citations for
 - properties 77
 - Unicode algorithms 78
 - Unicode Standard 76
- CJK ideographs 262, 682–698
 - accounting numbers 178
 - CJK Compatibility Ideographs 697–698
 - CJK Compatibility Supplement 698
 - CJK Strokes 700, 931
 - CJK Unified Ideographs 682–696
 - CJK Unified Ideographs Extension A 684
 - CJK Unified Ideographs Extension B 696
 - CJK Unified Ideographs Extension C 697
 - CJK Unified Ideographs Extension D 697
 - CJK Unified Ideographs Extension E 697
 - CJK Unified Ideographs Extension F 697
 - code charts 889
 - compatibility ideographs in Plane 2 52
 - component structure 692
 - encoding blocks 683
 - ideographic description sequences 701–704
 - ideographic variation mark (U+303E) 703
 - KangXi radicals 695, 699–700
 - names 889
 - numbers 794
 - numeric values 178, 207
 - order of encoding 694
 - radicals 699–700
 - source standards 930
 - unknown or unavailable 288
 - Vietnamese 680
- CJK Miscellaneous Area 50
- CJK punctuation and symbols 286
 - compatibility forms 288
 - overscores and underscores 288
 - quotation marks 274
 - sesame dots 287
 - vertical forms 288
- CJK-JRG (Chinese/Japanese/Korean Joint Research Group) 926
- CJKV Ideographs Area 50
- CLDR (Unicode Common Locale Data Repository) . 903
- cluster boundaries 219
- code charts 875–892, 902
 - representative glyphs 876
- code point sequences
 - notation 894
- code points 7, 29
 - assigned 11, 30
 - assignment 46
 - categories 30
 - default ignorable 203, 254
 - definition 90
 - designated 30
 - notation 893
 - 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 894
 - 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 631, 634
 Collation Algorithm, Unicode (UCA) 12
 collation *see* sorting
 collation tables 198
 combining character sequences 56, 106
 defective 225
 definition 108
 Latin 293
 line breaking 221
 matching 221
 order of base character and marks 222, 330
 rendering 221
 selection 219
 truncation 222–223
 combining characters 55–60, 110–115, 221–229
 blocking reordering 850
 canonical ordering 62, 138, 170
 combining marks 330–331
 definition 106
 dependence 330
 display order 58
 keyboard input 222
 ligatures 59
 multiple 57
 multiple base characters 59
 normalization of 208
 ordering conventions 56
 rendering of marks 224–229
 reordrant 171
 script-specific 56
 split 171
 strikethrough 172
 subjoined 172
 typographical interaction 58, 170
 vertical stacking 58
 see also diacritics
 Combining Class (normative property) 170
 combining classes 136, 170, 227–228
 class zero characters 170
 definition 136
 combining grapheme joiner (U+034F) 849
 combining half marks 192, 338
 combining marks *see* combining characters
 comma below 294
 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
 compatibility mappings
 see compatibility decomposition mappings
 compatibility precomposed characters
 see compatibility decomposable characters
 compatibility variants 26
 mapping 245
 composite characters
 see decomposable characters
 Composition Exclusion (normative property) ... 100
 compression 210
 see also UTS #6, A Standard Compression Scheme
 for Unicode (SCSU)
 conferences 902
 conformance 73–159
 definitions 87–93
 examples 69
 ISO/IEC 10646 implementations 918
 requirements 79–84
 confusables 247
 conjunct consonants
 Indic 219, 451
 Myanmar 625
 selection of clusters 219
 contextual shaping
 apostrophe 276
 Arabic 371
 not used for Hebrew final forms 366
 quotation marks 272
 Syriac 398
 contour tones 328
 control codes 31, 68, 840
 graphics for 814
 names 189
 properties 841
 semantics 32, 841
 specified in Unicode 841
 control sequences 840
 conversion of characters 128, 198–199, 256
 convertibility
 as Unicode design principle 23
 Coptic 309, 313–315
 reference materials 954
 Coptic Epact numbers 798
 corporate use subarea 860
 corrigenda 76
 CR (U+000D carriage return) 211, 841
 CRLF (carriage return and line feed) 211
 Croatian 298
 digraphs 298
 culturally expected sorting 12, 232

- Cuneiform
 Old Persian 435
 Sumero-Akkadian 430–433
 Ugaritic 434
Cuneiform and Hieroglyphic Area 51
Cuneiform and Hieroglyphs 429–444
currency symbols block 781–784
 currency symbols encoded in other blocks .. 782
 currency symbols, other 783
 dollar sign, form and usage 782
 euro sign 783
 lari sign 783
 lira sign, compatibility usage 782
 lira sign, Turkish 783
 peso signs, usage 782
 ruble sign 783
 rupee signs, Indian, usage 783
 yen and yuan signs, usage 782
cursive joining 845–849
 Arabic 379–386
 control characters for 193, 373–374, 531, 844
 Mandaic 406
 Mongolian 530–532
 N’Ko 741
 Phags-pa 573
 Syriac 398–401
 transparency 848
cursive scripts 363
Cypriot 346
 reference materials 960
 see also Linear B
Cyrillic 316–319
Czech 298
- D**
danda, in Devanagari block 463
Danish 297
dashes 269
Database, Unicode Character
 see Unicode Character Database (UCD)
dead consonants, Indic 450
dead keys 222
decomposable characters 63
 definition 116
 normalization of 208
decomposition 63, 116–118
 canonical *see* canonical decomposition
 compatibility *see* compatibility decomposition
 definition 116
 in normalization 208
 mapping, definition 116
 mappings noted in code charts 884
 default case
 algorithms 85, 153–159
 conversion 155
 detection 157
 folding 156
 default caseless matching 158
 default grapheme clusters 219
 see also UAX #29, Unicode Text Segmentation
 Default Ignorable Code Point (property) 254
 default ignorable code points 203, 254
 default property values
 definition 97
 defective combining character sequences 225
 definition 108
 dependent vowel signs
 Indic 449
 Khmer 636
 Philippine scripts 660
 deprecated characters 74, 879
 alternate format 194, 854–855
 definition 92
 Derived Age (property) 203
 derived properties
 definition 104
 DerivedCoreProperties.txt 154, 166, 254
 DerivedNormalizationProps.txt 244
 Deseret 759–761
 reference materials 954
 design goals of Unicode 4
 design principles of Unicode 14–24
 designated code points 30
 Devanagari 447–468
 Dhivehi 511
 diacritics 55, 330
 alternative glyphs 293, 330
 Czech 294
 display in isolation 60, 269, 331
 double 114, 192, 332
 German dialectology 336
 Greek 306–307, 310
 Latin 293–296
 Latvian 294
 mathematical 790
 on i and j 295
 rendering 224–229
 Slovak 294
 spacing clones of 328, 332
 symbol 55, 337
 see also combining characters
 dictionary symbols 824
 digit form names 375
 digits 207
 Arabic 794
 Arabic-Indic 375–376

- compatibility 794
 - decimal 177
 - glyph variants 796
 - hexadecimal 794
 - Myanmar 794
 - national shapes 855
 - Shan 794
 - superscript and subscript 795
 - Tai Laing 794
 - Tai Tham 794
 - digraphs 298, 301, 303
 - dingbats 826–828
 - directionality 20, 53
 - East Asian scripts 680
 - Middle Eastern scripts 363
 - Mongolian 529
 - musical symbols 767
 - normative property 173
 - Ogham 360
 - Old Italic 350
 - Philippine scripts 661
 - Runic 353
 - discussion list for Unicode 902
 - Dogri 464
 - Domino Tiles 829
 - dotless i 240, 295
 - dotted circle
 - in code charts 107, 331
 - in fallback rendering 224
 - to indicate diacritic 55
 - to indicate vowel sign placement 56
 - double diacritics 114, 192, 332
 - Duployan 775–776
 - reference materials 954
 - Dutch 297, 298
 - dynamic composition
 - as Unicode design principle 22
 - Dzongkha 517
- E**
- East Asian scripts 679–727
 - writing direction 53
 - see also* CJK ideographs
 - Eastern Arabic-Indic digits 375
 - EBCDIC
 - newline function 212
 - editing, text boundaries for 219–220
 - efficiency
 - as Unicode design principle 15
 - Egyptian hieroglyphs 436–440
 - reference materials 954
 - Elbasan 357
 - reference materials 955
 - ellipsis 277–278
 - e-mail discussion list for Unicode 902
 - emoji 822, 902
 - animal symbols 825
 - charts 902
 - cultural symbols 825
 - zodiacal symbols 825
 - emoji modifiers 826
 - emoticons 826
 - Enclosed Alphanumerics 835
 - enclosing marks 338
 - definition 107
 - encoded characters 7, 29
 - allocation 44–52
 - definition 92
 - encoding form conversion
 - definition 127
 - encoding forms 33–39
 - ISO/IEC 10646 definitions 913
 - 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 861
 - English 297
 - equivalent sequences 208
 - as Unicode design principle 23
 - case-insensitivity 233, 242
 - combining characters in matching 221
 - conformance 82
 - Hangul syllables 715
 - in sorting and searching 232
 - language-specific 118
 - security implications 247
 - see also* canonical equivalence
 - see also* compatibility equivalence
 - see also* encoding forms, encoding schemes
 - errata xxxvi, 76, 903
 - escape sequences 841
 - not used in Unicode 1, 4
 - Esperanto 298
 - Estonian 298
 - Ethiopic 730–733
 - reference materials 955
 - Etruscan 349
 - European scripts 291–339
 - ancient 341–361
 - eyelash-RA 456

- F**
- fallback rendering 254
 - of nonspacing marks 224
 - FAQ (Frequently Asked Questions) 902
 - Faroese 297
 - Farsi 371, 374
 - featural syllabaries 261
 - FF (U+000C form feed) 211, 841
 - file separator (U+001C) 841
 - Finnish 297
 - Finno-Ugric Transcription (FUT)
 - see Uralic Phonetic Alphabet (UPA)
 - fixed-width Unicode encoding form (UTF-32) ... 35, 124
 - flat tables 198
 - Flemish 297
 - fleurons 828
 - fonts
 - and Unicode characters 16
 - for mathematical alphabets 789–791
 - style variation for symbols 779
 - form feed (U+000C) (FF) 211, 841
 - format control characters 30, 68, 267, 839–855
 - deprecated 854–855
 - prefixed 194, 334
 - stateful 852
 - fraction characters 805
 - fraction slash (U+2044) 277, 801
 - French 298
 - Frisian 298
 - FTP site, Unicode Consortium 902
 - fullwidth forms in East Asian encodings 712
 - futhark 352
- G**
- Garshuni 394
 - Ge'ez 730
 - General Category (normative property) 174
 - list of values 174
 - general punctuation 265–289
 - General Scripts Area 50
 - geometrical symbols 819–821
 - Georgian 324–325
 - German 297
 - geta mark (U+3013) 288
 - Glagolitic 321
 - reference materials 955
 - Glossary 902
 - glyph selection tables 198
 - glyphs 6, 15
 - characters, relationship to 15
 - diacritics alternative 293, 330
 - Greek alternative 307–309
 - Latin alternative 293
 - mathematical alternative 807
 - missing 254
 - representative in code charts 876
 - standardized variants 856
 - symbols alternative 779
 - golden numbers 354
 - Gothic 356
 - reference materials 955
 - Grantha 609–611
 - reference materials 956
 - grapheme base 330
 - definition 108
 - grapheme clusters 11, 60–61
 - see also UAX #29, Unicode Text Segmentation
 - default 219
 - definition 109
 - grapheme extender
 - definition 109
 - grapheme joiner, combining (U+034F) 849
 - graphic characters 30
 - Greek 306–311
 - acrophonic numerals 207, 311
 - alternative glyphs 307–309
 - ancient musical notation 772–774
 - editorial marks 282
 - letters as symbols 307–309, 808
 - see also Cypriot, Linear B
 - Greek editorial marks
 - reference materials 956
 - Greenlandic 298
 - group separator (U+001D) 841
 - guillemets 272
 - Gujarati 480
 - Gurmukhi 475–479
- H**
- Hakka 706
 - halant 445
 - see also virama
 - half marks, combining 192, 338
 - half-consonants, Indic 452
 - halfwidth forms in East Asian encodings 712
 - Han ideographs see CJK ideographs
 - Han unification 689–696
 - and language tags 217
 - history 925–930
 - language usage 687
 - source separation rule 684, 690
 - source standards 930
 - hand symbols 825
 - Hangul Area 50

- Hangul syllables 679, 713–716
 and combining marks 114
 as grapheme clusters 61
 canonical decomposition 145
 collation 715
 composition 147
 conjoining jamo 143–152
 equivalent sequences 715
 Hangul Compatibility Jamo 714
 Hangul Jamo 713–716
 Hangul Syllables block 715–716
 Johab set 715
 name generation 148
 normalization 714
 standard 144
- Hangzhou numerals 801
- Hanja *see* CJK ideographs
- Hanunóo 660
- Hanzi *see* CJK ideographs
- harakat 372
- hasant 469
- hash tables 199
- Hatran 428
 reference materials 956
- Hebrew 365–370
- hentaigana 709–710
- hieroglyphs
 Anatolian 443–444
 Egyptian 436–440
 Meroitic 441–442
- high surrogate
 definition 119
 high-surrogate code points 79, 862
 high-surrogate code units 119
- higher-level protocols
 definition 93
- Hindi 447
- Hiragana 708
- horizontal tab (U+0009) 841
- HTML newline function 212
- Hungarian 298
- hyphenation 844
 as a text process 10
- hyphens 269, 844
- I**
- I Ching symbols 830
- IANA charset names 41
- Icelandic 297
- identifiers 231
see also UAX #31, Unicode Identifier and Pattern Syntax
- Ideographic (informative property) 190
- ideographic description sequences 702
- Ideographic Rapporteur Group (IRG) 928
- ideographs *see also* CJK ideographs
- IICore 685, 928
- ill-formed
 definition 122
- Imperial Aramaic 416–417
 reference materials 956
- implementation guidelines 197–256
 in a Unicode encoding form
 definition 123
- in-band mechanisms 874
- India
 Official scripts 445–508
- Indian rupee signs, usage 783
- Indic scripts 445–508
 principles, in terms of Devanagari 448–455
 relation to ISCII standard 447
- Indonesia and Oceania
 scripts of 659–678
- Indonesian 297
- industry character sets
 covered in Unicode 3
- information separators (U+001C..U+001F) 841
- informative properties
 definition 100
- Inscriptional Pahlavi 422
- Inscriptional Parthian 422
- inside-out rule 224
- interchange restrictions 31
- International Phonetic Alphabet (IPA) 260, 300–301
 reference materials 957
 Spacing Modifier Letters 327
see also phonetic alphabets
- internationalization 18
- Internationalization & Unicode Conference 902
- Internet protocols
 UTF-8 as preferred encoding 37
- Inuktitut 756
- invisible operators 813
- iota subscript 307
- IPA *see* International Phonetic Alphabet
- IRG (Ideographic Rapporteur Group) 928
- Irish 297, 360
- ISCII standard and Unicode 447
- ISO/IEC 10646 905–918
 conformance of Unicode implementations 918
 encoding forms 913
 synchrony with Unicode Standard 915
 timeline compared to Unicode versions 907
- Italian 297
- ITC Zapf Dingbats 826
- IUC *see* Internationalization & Unicode Conference

- J**
- jamos *see* Hangul syllables
- Japanese 679
- Javanese 671–674
 reference materials 958
- Jawi 390
- jihvamuliya 468, 590
- Johab 715
- joiners 373
 combining grapheme joiner (U+034F) 849
 word joiner (U+2060) 843
 zero width joiner (U+200D) 373–374, 846
- justification 226
- K**
- Kaithi 587–589
 reference materials 958
- Kana (Hiragana and Katakana) 708–709
- Kanbun 698
- KangXi radicals 695, 699–700
- Kanji *see* CJK ideographs
- Kannada 497–500
- Kashmiri 465
- Katakana 708–709
- Kawi 665, 667
- Kayah Li 653
 reference materials 958
- KC (normalization form)
 see Normalization Form KC
- KD (normalization form)
 see Normalization Form KD
- keytop labels 814
- Khamti Shan 628
- Kharoshthi 560–561
 reference materials 958
- Khmer 631–641
 characters not recommended 638
 syllable components, order of 639
- Khojki 598–599
 reference materials 959
- Khudawadi 600–601
 reference materials 959
- killer 262
 Batak 676
 Brahmi 556
 Meetei Mayek 541
 Myanmar (asat) 626
 see also virama
- Konkani 464
- Korean Hangul *see* Hangul
- Kurdish 390
- L**
- Ladino 365
- language tags 217, 870–874
 and Han unification 217
 use strongly discouraged 870, 873
- Lanna 644
- Lao 621–623
- last-resort glyphs 254
- Latin 293–305
 alternative glyphs 293
 Basic Latin 297
 encoding blocks 45
 IPA Extensions 300–301
 Latin Extended Additional 303–305
 Latin Extended-A 297
 Latin Extended-B 298–300
 Latin Extended-C 303
 Latin Extended-D 304
 Latin Extended-E 305
 Latin Ligatures 303
 Latin-1 Supplement 297
 Phonetic Extensions 302–305
- Latvian 298, 305
 cedilla 294
- layout control characters 68, 843–853
- leading surrogates
 see high-surrogate code units
- legibility criterion for plain text 19
- Lepcha 548–550
 reference materials 959
- letter spacing 844
- letterlike symbols 785–791
- LF (U+000A line feed) 211, 841
- ligatures 845–849
 Arabic 382–383
 combining characters on 59
 control characters for 193
 for nonspacing marks 228
 Latin 303
 selection 220
 Syriac 401
- Limbu 537–540
 reference materials 959
- line breaking 211–215, 843–845
 control characters 192
 in South Asian scripts 619, 627, 641
 recommendations 213
 see also UAX #14, Unicode Line Breaking Algorithm
- line feed (U+000A) (LF) 211, 841
- line separator (U+2028) (LS) 211, 845
- line tabulation (U+000B) (VT) 841

- Linear A 343
 - reference materials 960
 - Linear B 344–345
 - reference materials 960
 - see also* Cypriot
 - linear boundaries 220
 - Lisu 721–723
 - reference materials 961
 - Lithuanian 298
 - little-endian 40
 - definition 83
 - logical order
 - as Unicode design principle 19
 - exceptions to 171
 - logograph 262
 - logosyllabaries 262
 - low surrogate
 - definition 119
 - low-surrogate code points 79, 862
 - low-surrogate code units 119
 - lowercase 166, 238, 291
 - LS (U+2028 line separator) 211, 845
 - Lycian 347
 - reference materials 961
 - Lydian 347
 - reference materials 962
- M**
- MacOS newline function 212
 - Mahajani 596–597
 - reference materials 962
 - Mahjong Tiles 829
 - mail discussion list for Unicode 902
 - Maithili 464
 - major version 75
 - Malay 297
 - Malay, Patani 620
 - Malayalam 501–508
 - Suriyani 402, 502
 - Maltese 298
 - Manchu 529
 - Mandaic 405–407
 - reference materials 962
 - Mandarin 688
 - Manden 738
 - Manichaean 418–421
 - reference materials 962
 - map symbols 824
 - mapping tables *see* tables of character data
 - Marathi 447, 456, 463
 - Marchen 575
 - reference materials 962
 - markup languages
 - and Unicode conformance 874
 - line breaking 211
 - Masaram Gondi 552–553
 - reference materials 963
 - Mathematical (informative property) 805
 - mathematical expression format characters 194
 - see also* UTR #25, Unicode Support for Mathematics
 - mathematical symbols 805–812
 - alphabets 787–791
 - alphanumeric 786–791
 - fonts 789–791
 - format characters 813
 - fragments for typesetting 815
 - invisible operators 813
 - operators 806–809
 - reference materials 963
 - standardized variants 812
 - MathML 809
 - matras 170, 449
 - Meetei Mayek 541–542
 - reference materials 963
 - Mende Kikakui 748–749
 - reference materials 963
 - Meroitic
 - cursive 441–442
 - hieroglyphs 441–442
 - reference materials 964
 - Miao 724–725
 - reference materials 964
 - Middle Eastern scripts 363–512
 - ancient 409–428
 - Min 688
 - Minnan (Hokkien/Fujian, incl. Taiwanese) 706
 - minor version 75
 - minus sign
 - commercial (U+2052) 280
 - mirrored property
 - see* Bidi Mirrored (normative property)
 - mirroring of paired punctuation 271
 - Miscellaneous Symbols 823
 - missing glyphs 254
 - Modi 606–608
 - reference materials 964
 - modifier letters 326–329
 - Modifier Letters, Spacing 302
 - Mongolian 528–536, 568
 - writing direction 529
 - Mro 543
 - reference materials 964, 965
 - Multani 602
 - reference materials 964

- multibyte encodings
 - compared to UTF-8 37
 - multistage tables 198
 - musical symbols 766–774
 - ancient Greek 772–774
 - Balinese 669
 - Byzantine 771
 - directionality 767
 - Gregorian 769
 - Kievan 770
 - reference materials 965
 - Western 766–770
 - Myanmar 624–630
 - digits 794
 - Myanmar Extended-A 628
 - Myanmar Extended-B 628
 - reference materials 965
- N**
- N’Ko 738–742
 - reference materials 966
 - Nabataean 426
 - reference materials 966
 - named character sequences 183
 - names, character *see* character names
 - namespace 89
 - NEL (U+0085 next line) 211, 841
 - Nepali 447
 - neutral directional characters 173
 - New Tai Lue 644–646
 - Newa 515–516
 - reference materials 966
 - newline function (NLF) 212, 842
 - newline guidelines 211–215
 - next line (U+0085) (NEL) 211, 841
 - NFC (Normalization Form C) 62
 - NFD (Normalization Form D) 62
 - NFKC (Normalization Form KC) 62
 - NFKD (Normalization Form KD) 62
 - NLF (newline function) 212, 842
 - no-break space (U+00A0) 843
 - base for diacritic in isolation 60, 269, 331
 - no-break space, narrow (U+202F) 534
 - noncharacter code points *see* noncharacters
 - noncharacters 31, 863
 - conformance 79
 - definition 93
 - handling 82
 - in code charts 879
 - interchange restrictions 31
 - semantics 32
 - U+10FFFF (not a character code) 863
 - U+FDD0..U+FDEF 31, 863
 - U+FFFE (not a character code) 67, 864
 - U+FFFF (not a character code) 31, 863
 - nondecomposable characters 64
 - non-joiner, zero width (U+200C) 373–374, 847
 - nonlinear boundaries 220
 - non-overlap principle in Unicode encoding forms 33
 - nonspacing marks 330
 - definition 107
 - display in isolation 60, 269, 331
 - positioning 228
 - rendering 224–229
 - see also* combining characters
 - see also* diacritics
 - normalization 62, 208–209
 - and case operations 244
 - canonical ordering algorithm 62, 138, 170
 - conformance 84
 - of private-use characters 860
 - see also* UAX #15, Unicode Normalization Forms
 - stability 135
 - Normalization Form C (NFC) 62
 - Normalization Form D (NFD) 62
 - Normalization Form KC (NFKC) 62
 - Normalization Form KD (NFKD) 62
 - normalization forms 135–142
 - definition 141
 - specification 137
 - normative behaviors
 - definition 87
 - normative properties
 - definition 99
 - list 100
 - may change 99
 - Norwegian 297
 - notational conventions 893–897
 - notational systems 263, 763–778
 - nukta 372, 392, 457
 - null (U+0000)
 - as Unicode string terminator 842
 - number forms
 - CJK ideographs 207
 - numbers
 - Coptic Epact 798
 - handling 207
 - ideographic accounting 178
 - numerals 792–802
 - acrophonic 311
 - Chinese counting rods 803
 - Coptic 315
 - Cuneiform 433
 - Ethiopic 732
 - Greek acrophonic 207
 - Hangzhou 801
 - Meroitic cursive 442

old-style	277
Roman	207, 805
Rumi	799
Suzhou-style	801
numeric separators	280
numeric shape selectors (deprecated)	855
Numeric Type (normative property)	177
Numeric Value (normative property)	177
numero sign (U+2116)	785
Nüshu	720
reference materials	967

O

object replacement character (U+FFFC)	869
octet	895
Ogham	360
reference materials	967
Ol Chiki	545–546
reference materials	967
Old Church Slavonic	316
Old Hungarian	355
reference materials	967
Old Italic	349–351
reference materials	967
Old North Arabian	411
reference materials	968
Old Permic	359
reference materials	968
Old Persian	435
reference materials	968
Old South Arabian	412–413
reference materials	968
Old Turkic	581
reference materials	969
old-style numerals	277
Oriya	482–484
ornamental dingbats	828
Oromo	730
Osage	758
reference materials	967
Osmanya	734
reference materials	969
out-of-band mechanisms	874
overlapping encodings	33
overscores	277

P

Pahawh Hmong	656–657
reference materials	969
Pahlavi, Inscriptional	422
reference materials	957
Pahlavi, Psalter	423

Palmyrene	427
reference materials	970
Panjabi	475
paragraph or section marks	280
paragraph separator (U+2029) (PS)	211, 845
Parthian, Inscriptional	422
reference materials	957
Pashto	371
Patani Malay	620
Pau Cin Hau	658
reference materials	970
Persian	371, 374
Phags-pa	568–574
reference materials	970
Phaistos Disc symbols	832
Phake	630
Philippine scripts	660–662
reference materials	971
Phoenician	414
reference materials	971
phonemes	263
phonetic alphabets	260
IPA Extensions	300–301
Phonetic Extensions	302–305
Spacing Modifier Letters	327–329
Uralic Phonetic Alphabet (UPA)	280, 302
<i>see also</i> International Phonetic Alphabet (IPA)	
Pinyin	297
pipeline table	
proposed new characters	903
pivot code, Unicode as	198
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, 861
Playing Cards	829
points, Hebrew pronunciation marks	365
policies of the Unicode Consortium	903
Polish	298
Portuguese	297
precomposed characters	
<i>see</i> decomposable characters	
compatibility <i>see</i> compatibility decomposable characters	
prefixed format control characters	194
prepended concatenation marks	255, 334
Private Use Area (PUA)	50, 860
Private Use planes	45, 52, 861

private-use characters
 properties 859
 semantics 32

private-use code points 31, 203
 conformance 80
 definition 105
 high surrogates 862

processing code, Unicode as 38

properties 18, 95–105, 161–195
 aliases 164
 aliases (definition) 104
 and Unicode algorithms 99
 data tables 198
 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 841
 provisional *see* provisional properties
 simple *see* simple properties
 see also individual properties, e.g. combining classes

property values
 aliases 164
 aliases (definition) 105
 default 97
 default (definition) 97
 normative references to 84

PropertyAliases.txt 104, 896

PropertyValueAliases.txt 105, 896

PropList.txt 168

Provençal 298

provisional properties
 definition 101

PS (U+2029 paragraph separator) 211, 845

Psalter Pahlavi 423
 reference materials 972

PUA (Private Use Area) 50, 860

pulli 485

punctuation 265–289
 blocks containing 259
 CJK 286
 doubled 277
 in bidirectional text 265
 paired 271
 small form variants 289
 typographic forms 265
 vertical forms 288

Punctuation and Symbols Area 50

Punjabi 475

Q

quotation marks 272–275
 East Asian 274
 European 272

R

radicals, KangXi and other CJK 699–700

radical-stroke index 695

record separator (U+001E) 841

recycling symbols 824–825

referencing 84
 properties 77
 Unicode algorithms 78
 Unicode Standard 76

regional indicator symbols 836

regular expressions 216
 and line breaking 211
 see also UTS #18, Unicode Regular Expressions

Rejang 675
 reference materials 972

rendering of text 6, 10, 17
 fallback 254
 unsupported characters 203

repertoire of abstract characters 29

replacement character (U+FFFD) ... 43, 68, 83, 128, 256, 869

reserved code points 30, 203
 definition 93
 in code charts 879
 preservation in interchange 31
 see also unassigned code points

Rhaeto-Romanic 298

rich text 18

right single quotation mark (U+2019)
 preferred for apostrophe 276

right-to-left text 53
 East Asian scripts 680
 Middle Eastern scripts 363

roadmap for script additions 46, 903

Roman numerals 207, 805

Romanian 298
 comma below 295

Romany 298

Rong 548–550

Rumi numeral symbols 799

Runic 352–354
 reference materials 972

Russian 316

S

Samaritan 403–404
 reference materials 972

Sami 298

- Sanskrit 447
- Saurashtra 551
 - reference materials 973
- scalar values, Unicode
 - see* Unicode scalar values
- scripts
 - in Unicode Standard 3
 - roadmap for future additions 46, 903
 - types of 264
 - see also* UAX #24, Unicode Script Property
- SCSU
 - see* UTS #6, A Standard Compression Scheme for Unicode
- searching 232–234
 - as a text process 10
 - case-insensitive 233, 242
- section or paragraph marks 280
- security issues 247
- self-synchronization of encoding forms 34
- semantics
 - see* character semantics
- sequences
 - notation 894
- Serbian
 - corresponding digraphs in Croatian 298
- Shan 642
 - digits 794
- Sharada 590–591
 - reference materials 973
- Shavian 361, 721
 - reference materials 973
- Show Hidden 81, 224, 254, 857
- SHY (U+00AD soft hyphen) 844
- Sibe 529
- Siddham 594–595
 - reference materials 973
- signature for Unicode data 67, 865–867
- simple properties
 - definition 104
- simplified Chinese 687
- Sindhi 371, 464
- Sinhala 513–514
 - reference materials 974
- Sinological dot 304
- SIP (Supplementary Ideographic Plane) 44, 52
- slash, fraction (U+2044) 277
- Slovak 298
- Slovenian 298
- small letters 166, 238, 291
- SMP (Supplementary Multilingual Plane) 44, 51
- soft hyphen (U+00AD) (SHY) 844
- Somali 734
- Sora Sompeng 614
 - reference materials 974
- Sorbian 298
- sorting 12, 232
 - and combining grapheme joiner 850
 - as a text process 10
 - case-insensitive 233
 - culturally expected 12, 232
 - language-insensitive 232
 - see also* Unicode Collation Algorithm (UCA)
- source separation rule 684, 690
- South and Central Asian scripts
 - Ancient 555–581
 - Other historic 583–614
 - Other modern 509–551
- South Asian scripts 445–540
- Southeast Asian scripts 615–658
- Soyombo 579–580
 - reference materials 974
- space (U+0020)
 - base for diacritic in isolation 60, 269, 331
- space characters 268, 843–845
 - graphics for 814
- space, zero width (U+200B) 268
- spacing clones of diacritics 328, 332
- spacing marks 330
 - definition 108
- Spacing Modifier Letters 327–329
- Spanish 297
- special characters 67, 839–874
- SpecialCasing.txt 153, 168
- Specials 865–869
- spell-checking
 - as a text process 11
- spellings, alternative
 - see* equivalent sequences
- spoofing 247
- SSP (Supplementary Special-purpose Plane) 45
- stability 102, 163
 - as Unicode design principle 23
- stacked boundaries 219
- stacking sequences 57
 - nondefault 58
- standardized variants 532, 856
 - in the code charts 886
 - mathematical symbols 812
- StandardizedVariants.txt 532, 812
- standards coverage 3
- starters 137
- stateful encoding
 - not used in Unicode 4
 - paired format controls 852
- string comparison 12
- string literals, Unicode
 - code point notation `\u1234` 896

strings, Unicode 43, 121
 null termination 842
 strong directional characters 173
 styled text 18
 sublinear searching 233
 subsets, supported 71
 conformance 80
 ISO/IEC 10646 specification for 916
 substitution character
 see replacement character
 Sumero-Akkadian 430–433
 Sumero-Akkadian Cuneiform
 reference materials 974
 Sundanese 677–678
 reference materials 975
 superscripts 328
 and subscripts 803
 supplementary characters
 in UTF-16 strings 43
 tables for 199
 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, 861
 Supplementary Special-purpose Plane (SSP) 45
 supported subsets 71
 conformance 80
 supralineation 314
 surrogate code points
 see surrogates
 surrogate pairs 36, 125
 definition 119
 processing 38, 205–206
 surrogates 31, 119, 862
 interchange restrictions 31
 isolated surrogates, handling 43
 isolated surrogates, ill-formed 125
 isolated surrogates, uninterpreted 119
 support levels 205
 Surrogates Area 50, 862
 Sutton SignWriting 777–778
 reference materials 975
 Suzhou-style numerals 801
 svasti signs 524
 Swahili 297
 Swedish 297
 syllabaries 261
 alphabetic property 190
 featural 261
 Syloti Nagri 585–586

symbols 779–837
 animal 825
 appearance variation 779
 arrows 811–812
 box drawing 819
 cultural 825
 currency symbols block 781–784
 dictionary 824
 dingbats 826–828
 emoji 822, 836
 Enclosed Alphanumerics 835
 fragments for mathematical typesetting 815
 game 825
 gender 825
 genealogical 825
 geometrical 819–821
 hand 825
 Khmer lunar calendar 641
 letterlike 785–791
 map 824
 mathematical 805–812
 mathematical alphanumeric 786–791
 miscellaneous 823
 musical 766–774
 numerals 792–802
 recycling 824–825
 regional indicator 836
 technical 814–818
 weather 824
 zodiacal 825
 symmetric swapping format characters 854
 Syriac 394–401
 reference materials 975

T

tab (U+0009 character tabulation) 841
 tab, vertical (U+000B) 211, 841
 tables of character data 198–199
 optimization 199
 supplementary characters 199
 tag characters 870–874
 Tagalog 660
 Tagbanwa 660
 tags, language 217, 870–874
 use strongly discouraged 873
 Tai Laing
 digits 794
 Tai Le 642–643
 reference materials 976
 Tai Tham 647–649
 digits 794
 reference materials 976
 Tai Viet 650–652

- Tai Xuan Jing symbols 831
- Takri 592–593
 reference materials 976
- Tamil 485–493
- Tangut 726–727
 components 727
 radicals 727
 reference materials 976
- tashkil 372
- tashkil, harakat, points 374
- TCHAR in Win32 API 202
- Technical Reports (UTR) 901
- Technical Standards (UTS) xxxvi, 901
 abstracts 902
- technical symbols 814–818
- Telugu 494–496
- terminal emulation 780
- text boundaries 61, 191, 219–220, 230
 see also UAX #14, Unicode Line Breaking Algorithm
 see also UAX #29, Unicode Text Boundaries
- text elements 6, 10, 219
 boundaries 230
 for sorting 232
 variable-width nature 38
- text processes 6, 10–13
- text rendering 6, 10, 17
- text selection, boundaries for 219–220
- Thaana 511–512
 reference materials 976
- Thai 617–620
- Tibetan 517–527
- Tifinagh 735
- Tigre 730
- tilde (U+007E) 280
- Tirhuta 603–605
 reference materials 977
- titlecase 166, 238
- Todo 529
- tone letters 328–329
- tone marks
 Bopomofo spacing 705, 706
 Chinantec 329
 Chinese 329
 Tai Le 642
 Thai 617
 Vietnamese 296
- traditional Chinese 687
- traffic signs 824
- trailing surrogates
 see low-surrogate code units
- transcoding 198–199
 tables 198
- Transport and Map Symbols 826
- triangulation in transcoding 198
- tries 198
- truncation
 combining character sequences 222–223
 surrogates and 206
- Turkish 298
 case mapping of İ 240, 295
 cedilla 295
 lira sign 783
- two-stage tables 199
- ## U
- U+ notation 896
- U+10FFFF (not a character code) 863
- U+FEFF (BOM) 865–867
- U+FFFE (not a character code) 864
- U+FFFF (not a character code) 863
- UAX (Unicode Standard Annex) xxxiii, 901
 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-2 913
- UCS-4 913
- Ugaritic 434
 reference materials 977
- Ukrainian 316
- unassigned code points 30, 79, 203
 defined as reserved code points 93
 handling 74
 properties of 97
 semantics 79
 see also reserved code points
- underscores 277
- undesignated code points 30
- Unicode 1.0 Name (informative property) 189
- Unicode algorithms
 and properties 99
 conformance 84
 definition 93
 normative references to 78, 84
- Unicode Bidirectional Algorithm 20, 53
 see also UAX #9, Unicode Bidirectional Algorithm
- Unicode Character Database (UCD) xxxv, 163, 903
 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+ 896
- Unicode codespace
 allocation numbers 920
 definition 90
 planes 44
 size 1, 29
- Unicode Collation Algorithm (UCA) 12
- Unicode Common Locale Data Repository (CLDR) . 903
- Unicode conferences 902
- Unicode Consortium 900
 addresses 904
 Consortium membership in standards bodies 900
 e-mail discussion list 902
 FTP site 902
 membership 900
 policies 903
 website 902
- Unicode data signature 67, 865–867
- Unicode data types 201–202
 for C 201–202
- Unicode encoding forms 120–127
 advantages of each 38
 conformance 34, 82
 definition 121
 fixed-width (UTF-32) 35, 124
 signatures 866, 867
 variable-width 36, 125
 see also encoding forms
- Unicode encoding schemes
 conformance 131–134
 definition 131
 endian ordering 40
 see also encoding schemes
- Unicode escape sequence notation `\u1234` 896
- Unicode scalar values
 definition 120
- Unicode security 247
 see also UTS #39, Unicode Security Mechanisms
- Unicode Standard
 allocation of encoded characters 44–52
 architecture 10–13
 areas 45
 benefits 1
 blocks 259
 code charts 875–892, 902
 components 79
 conformance 73–159
 conformance of ISO/IEC 10646 implementations
 918
 corrections 76
 definitions for conformance 87–93
 design goals 4
 design principles 14–24
 errata 76, 903
 normative references to 76, 84
 number of characters 3
 number of code points 1, 29
 script coverage 3
 security issues 247
 synchrony with ISO/IEC 10646 915
 updates 903
 versions *see* versions of the Unicode Standard
 see also Version 10.0
- Unicode Standard Annexes (UAX) xxxiii, 901
 as components of Unicode Standard 79
 conformance 85
 list of 85
- Unicode string literals
 code point notation `\u1234` 896
- Unicode strings 43
 definition 121
- Unicode Technical Committee (UTC) 900
- Unicode Technical Reports (UTR) 901
- Unicode Technical Standards (UTS) xxxvi, 901
 abstracts 902
- UnicodeData.txt 153, 168
- unification
 as Unicode design principle 21
 see also Han unification
- Unified Repertoire and Ordering (URO) .. 691, 927
 see also Han unification
- Unihan Database 163, 695, 696, 889, 903, 928
- Unihan.zip 102, 163
- unit separator (U+001F) 841
- Universal Character Set (UCS)
 see ISO/IEC 10646
- universality
 as Unicode design principle 14
- Unix
 and UTFs 38
 newline function 212
 UTF-32 in 35
 UTF-8 in 18
- unsupported characters 203
- upadhmaniya 468, 590
- update version 75
- uppercase 166, 238, 291
- Uralic Phonetic Alphabet (UPA) 280, 302
- Urdu 371
- URO (Unified Repertoire and Ordering) .. 691, 927
 see also Han unification
- UTF, Unicode Transformation Formats 33, 121
 advantages of each 38
 as encoding form or scheme 134

binary comparison and sort order differences . . .
 233, 235
 in APIs 202
 UTF-16 36, 125, 914
 binary comparison and sort order caution . . . 36
 bit distribution (table) 125
 BOM in 132, 865
 encoding form (definition) 125
 encoding scheme (definition) 132
 encoding schemes 40
 in ISO/IEC 10646 914
 in UTF-8 order 236
 surrogates and string handling 43, 205
 UTF-16BE (Big-endian) 866
 encoding scheme 41
 encoding scheme (definition) 131
 UTF-16LE (Little-endian) 866
 encoding scheme 41
 encoding scheme (definition) 131
 UTF-32 35, 124
 as processing code 38
 BOM in 133
 encoding form (definition) 124
 encoding scheme (definition) 133
 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) 133
 UTF-8 36, 125, 914
 ASCII transparency 36
 binary comparison and sort order 39
 bit distribution (table) 126
 BOM in 131, 134, 866
 byte ranges 126
 compared to multibyte encodings 37
 encoding form (definition) 125
 encoding scheme 40
 encoding scheme (definition) 131
 in Unix 18
 in UTF-16 order 235
 non-shortest form is invalid 125, 247
 preferred encoding for Internet protocols . . . 37
 security and 247
 signature 131, 134, 866
 UTR (Unicode Technical Report) 901
 UTS (Unicode Technical Standard) xxxvi, 901
 abstracts 902
 Uyghur 371, 528, 568

V

Vai 743–744
 reference materials 977
 valid (synonym for well-formed) 123
 variable-width Unicode encoding form 36, 125
 variants
 compatibility 26
 fullwidth and halfwidth 289
 mathematical symbols 812
 small form 289
 standardized 856
 variation selectors 195, 856
 ideographic variation mark (U+303E) 703
 Mongolian free variation selectors 532
 variation sequences 856
 for Phags-pa 572–574
 Version 10.0 79
 number of characters 3
 versions of the Unicode Standard xxxv, 74, 903, 919–
 920
 backward compatibility 74
 compared to ISO/IEC 10646 editions 919
 content 75
 interaction in implementations 203
 numbering 75
 property changes 74
 stability 74
 updates 903
 vertical tab (U+000B) 211, 841
 vertical text 53, 266, 288
 East Asian scripts 680
 Mongolian 529
 Vietnamese 296, 303
 ideographs 680
 virama 262, 445
 definition 450
 Kharoshthi 564
 Khmer 634
 Myanmar 625
 Philippine scripts 660
 virama-like characters 193
 visual order used for Thai and Lao 21
 vowel harmony
 Mongolian 533
 vowel marks, Middle Eastern scripts 363
 vowel separator
 Mongolian 534
 vowel signs
 Indic 56, 449
 Khmer 636
 Philippine scripts 660

W

- Warang Citi 544
- reference materials 977
- wchar_t
 - and Unicode encoding forms 38
 - in C language 202
- weak directional characters 173
- weather symbols 824
- website, Unicode Consortium 902
- Weierstrass elliptic function symbol 786
- well-formed
 - definition 122
- Welsh 298
- Where Is My Character? 904
- wide characters
 - data type in C 202
- wiggly fence (U+29DB) 810
- Windows newline function 212
- word breaks 221, 843–845
 - in South Asian scripts 619, 627, 641
- word joiner (U+2060) 843
- writing direction *see* directionality
- writing systems 260–264
- Wu (Shanghainese) 688

X

- Xibe 529
- Xishuangbanna Dai 644

Y

- Yi 717–719
 - reference materials 978
- Yiddish 365
- Yijing Hexagram Symbols 830
- ypogegrammeni 307

Z

- Zanabazar Square 576–578
 - reference materials 978
- Zapf Dingbats 826
- zero extension relation among encodings 913
- zero width joiner (U+200D) 373–374, 846
- zero width no-break space (U+FEFF) ... 67, 83, 843
 - initial 134, 866
- zero width non-joiner (U+200C) ... 373–374, 847
- zero width space (U+200B) 844
 - for word breaks in South Asian scripts .619, 627, 641
- zero-width space characters 844
- 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)