

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

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

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

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

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

Copyright © 1991–2009 Unicode, Inc.

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

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

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

Includes bibliographical references and index.

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

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

II. Unicode Consortium.

QA268.U545 2009

ISBN 978-1-936213-00-9

Published in Mountain View, CA

December 2009

I General Index

The General Index covers the contents of this book. To find topics in the Unicode Standard Annexes, Unicode Technical Standards, and Unicode Technical Reports, use the search feature on the Unicode Web site.

For definitions of terms used in this book, see the glossary on the Unicode Web site. To find the code points for specific characters or the code ranges for particular scripts, use the Character Index on the Unicode Web site. (See *Section B.6, Other Unicode Online Resources.*)

A

abbreviation, Coptic 216
 abjads 178, 233
 abstract character sequences
 definition 66
 abstract characters 22
 definition 66
 abugidas 179, 180, 261, 337
 accent marks *see* diacritics
 accented characters
 encoding 9
 Latin 200
 normalization 144
 accounting numbers, ideographic 126
 acrophonic numerals 143, 214
 Aegean numbers 443
 Afrikaans 204
 Ainu 402
 Aiton 347
 Algonquian 428
 Ali Gali 412
 aliases
 character name 65, 129, 528
 property 117
 property value 117
 allocation areas 34
 allocation of encoded characters 32–40, 557
 Alphabetic (informative property) 132
 alphabets 178
 European 199–231
 mathematical 463–467
 Alpine 438
 alternate format characters (deprecated) .. 134, 509–510
 Amharic 410
 Ancient Symbols 486
 angle brackets (U+2329 and U+232A)
 deprecated for technical publication 477
 Annexes, Unicode Standard (UAX) xxxiv, 542
 as components of Unicode Standard 59
 conformance 63
 list of 63
 annotation characters 517–518
 use in plain text discouraged 518
 ANSI/ISO C
 wchar_t and Unicode 140
 apostrophe (U+0027) 189
 Arabic 238–250

Arabic-Indic digits 240–241
 signs used with 242
 ArabicShaping.txt 243, 246, 255
 Aramaic 261, 306, 332, 413
 archaic scripts 435–444
 areas of the Unicode Standard 34
 ARIB 482
 Armenian 220–221
 arrows 474–475
 ASCII
 characters with multiple semantics 182
 transparency of UTF-8 28
 Unicode modeled on 1
 zero extension 140, 553
 Assamese 279
 assigned code points 8, 23
 Athapascan 428
 atomic character boundaries 151
 Avestan 450

B

Balinese 368–373
 Bamum 427
 Bangla 279–283
 base characters 225
 definition 78
 multiple 45
 ordered before combining marks 153, 225
 Basic Multilingual Plane (BMP) 1, 33
 allocation areas 37
 representation in UTF-16 27
 Basque 204
 benefits of Unicode 1
 Bengali 279–283
 Bidi Class (normative property) 124
 Bidi Mirrored (normative property) 127
 Bidi Mirroring Glyph (informative property) ... 128
 BidiMirroring.txt 128
 Bidirectional Algorithm, Unicode 40, 62
 bidirectional ordering 15
 controls 134, 508
 bidirectional text 40, 62
 Middle Eastern scripts 233
 nonspacing marks in 156
 punctuation in 182
 big-endian 30
 definition 62

- Bihari262
 - binary comparison and sort order
 - caution for UTF-1627
 - UTF differences161, 162
 - UTF-829
 - blocks of the Unicode Standard34, 177
 - Blocks.txt34
 - BMP *see* Basic Multilingual Plane
 - BNF (Backus-Naur Form)539
 - BOCU-1 *see* UTN #6, BOCU-1
 - MIME-Compatible Unicode Compression
 - BOM (U+FEFF)30, 50, 96–98, 515–516
 - Bopomofo399–401
 - boundaries, text8, 46, 133, 151–152, 159
 - see also* UAX #14, Unicode Line Breaking Algorithm
 - see also* UAX #29, Unicode Text Segmentation
 - boustrophedon41, 439
 - Brahmi261, 306, 332, 338
 - Braille490–491
 - Breton204
 - Buginese367–368
 - Buhid366
 - Bulgarian217
 - bullets192
 - Burmese *see* Myanmar
 - Byelorussian217
 - byte order mark (BOM) (U+FEFF) ...30, 50, 96–98, 515–516
 - byte ordering
 - changing61
 - conformance62
 - byte serialization30, 50
 - Byzantine Musical Symbols496
- C**
- C language
 - wchar_t and Unicode140
 - C0 and C1 control codes23, 132, 500
 - Cambodian *see* Khmer
 - camelcase167
 - Canadian Aboriginal Syllabics428–429
 - canonical composite characters
 - see* canonical decomposable characters
 - canonical composition algorithm102
 - canonical decomposable characters
 - definition87
 - canonical decomposition48
 - definition86
 - mappings85
 - canonical equivalence
 - definition87
 - nonspacing marks157
 - canonical equivalent character sequences
 - conformance60
 - canonical mappings
 - see* canonical decomposition mappings
 - canonical ordering algorithm102
 - canonical precomposed characters
 - see* canonical decomposable characters
 - Cantonese387
 - capital letters118, 164, 199
 - Carian444
 - carriage return (U+000D) (CR)146, 501
 - carriage return and line feed (CRLF)146
 - case205
 - and text processes9
 - beyond ASCII165
 - camelcase167
 - case folding168
 - case operations (conformance)63, 110–113
 - case operations and normalization169
 - case operations, reversibility167
 - cased (definition)110
 - case-insensitive comparison ...113, 161, 162, 168
 - casing context (definition)111
 - conversion111
 - detection112
 - European alphabets199
 - exceptional Latin pairs201, 204
 - Georgian222
 - lowercase118, 164, 199
 - mapping tables138
 - mappings110, 120, 164–167
 - mappings noted in code charts529
 - see also* default case
 - titlecase118, 165
 - Turkish I166, 201
 - uppercase118, 164, 199
 - Case (normative property)118, 164
 - CaseFolding.txt120, 168
 - caseless letters205
 - Catalan203
 - 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)
 - Cham364–365
 - character encoding forms (CEF)24–30, 552
 - see also* Unicode encoding forms
 - character encoding model24, 32
 - see also* UTR #17, Unicode Character Encoding Model
 - character encoding schemes (CES)30–32
 - see also* Unicode encoding schemes
 - character encoding standards
 - coverage by Unicode2
 - Character Index545
 - character literals, Unicode
 - code point notation U+539
 - character mapping
 - interchange format *see* UTS #22, Character Mapping Markup Language (CharMapML)
 - character names65, 128–131, 555
 - aliases65, 129, 528
 - conventions537
 - for CJK ideographs531
 - for control codes130, 132
 - in code charts525–529
 - matching129
 - character properties
 - see* properties
 - see also* individual properties, e.g. combining classes
 - character semantics1, 60, 64–65, 555
 - as Unicode design principle13
 - ASCII182
 - definition65

- character sequences
 - abstract *see* abstract character sequences
 - canonical equivalent *see* canonical equivalent character sequences
 - compatibility equivalent *see* compatibility equivalent character sequences
 - conformance 60
 - named 129
- character sequences, combining 77
- character shaping selectors (deprecated) 510
- character tabulation (U+0009) 501
- characters
 - abstract *see* abstract characters
 - arrangement in Unicode 35
 - assigned 8, 23
 - blocks 34, 177
 - boundaries 151
 - canonical decomposable *see* canonical decomposable characters
 - classes 539
 - code charts 525–533, 545
 - coded *see* encoded characters
 - combining *see* combining characters
 - compatibility decomposable *see* compatibility decomposable characters
 - composite *see* decomposable characters
 - concept of 11, 46
 - conformance definitions 66–68
 - confusable 172
 - conversion 137–139
 - 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 46
 - format control 23, 51, 183, 499–523
 - glyphs, relationship to 11
 - graphic 23
 - identity (definition) 64
 - interpretation 60
 - layout control 51, 501–509
 - modification 60
 - names list 525–529
 - names *see* character names
 - not encoded in Unicode 2
 - number encoded in Version 5.0 557
 - number encoded in Version 5.2 2
 - precomposed *see* decomposable characters
 - properties *see* properties
 - semantics *see* character semantics
 - special 50, 499–523
 - supplementary *see* supplementary characters
 - transcoding 137–139
 - unsupported 140–141
- characters, not glyphs
 - in spoofing 172
 - Unicode principle 11
- CharMapML
 - see* UTS #22, Character Mapping Markup Language (CharMapML)
- charsets
 - IANA registered names 31
- charts, character code *see* code charts
- Cherokee 427
- Chinese 386–387
 - Cantonese 387
 - Hakka 401
 - Mandarin 387
 - Minnan (Hokkien/Fujian, incl. Taiwanese) .. 401
 - simplified and traditional 386
- Chu hán 385
- Chu Nôm 566
- citations for
 - properties 58
 - Unicode algorithms 58
 - Unicode Standard 57
- CJK ideographs 180, 380–394
 - accounting numbers 126
 - CJK Compatibility Ideographs 393–394
 - CJK Compatibility Supplement 394
 - CJK Unified Ideographs 380–393
 - CJK Unified Ideographs Extension A 383
 - CJK Unified Ideographs Extension B 393
 - CJK Unified Ideographs Extension C 393
 - code charts 531
 - compatibility ideographs in Plane 2 40
 - component structure 390
 - encoding blocks 383
 - ideographic description sequences 396–399
 - ideographic variation mark (U+303E) 399
 - KangXi radicals 394–395, 535
 - names 531
 - numeric values 126, 143
 - order of encoding 392
 - radicals 394–395
 - radical-stroke index 535
 - source standards 381–383
 - unknown or unavailable 197
 - Vietnamese 379
- CJK Miscellaneous Area 38
- CJK punctuation and symbols 196
 - compatibility forms 198
 - overscores and underscores 198
 - quotation marks 188
 - sesame dots 197
 - vertical forms 198
- CJK Radical (property) 397
- CJK-JRG (Chinese/Japanese/Korean Joint Research Group) 565
- CJKV Ideographs Area 38
- CLDR (Unicode Common Locale Data Repository) 546
- cluster boundaries 151
- code charts 525–533, 545
 - representative glyphs 526
- code point sequences
 - notation 538
- code points 5, 22
 - assigned 8, 23
 - assignment 35, 557
 - categories 23
 - default ignorable 141, 173
 - definition 66
 - designated 23
 - notation 537
 - number in Unicode Standard 1
 - private-use *see* private-use code points
 - reserved *see* reserved code points
 - semantics 24

- surrogate *see* surrogates
- unassigned *see* unassigned code points
- undesignated 23
- code positions *see* code points
- code set independence 13
- code unit sequences
 - definition 88
 - ill-formed (definition) 89
 - notation 538
 - well-formed (definition) 90
- code units
 - definition 88
 - isolated 87
- code values *see* code units
- coded character representations
 - see* coded character sequences
- coded character sequences
 - definition 67
- coded characters *see* encoded characters
- codespace *see* Unicode codespace
- coeng 349, 351
- Collation Algorithm, Unicode (UCA) 10
- collation *see* sorting
- collation tables 138
- combining character sequences 42, 77
 - defective 156
 - definition 79
 - Latin 200
 - line breaking 153
 - matching 153
 - order of base character and marks 153, 225
 - rendering 153
 - selection 151
 - truncation 153–154
- combining characters 41–46, 81–85, 152–159
 - blocking reordering 507
 - canonical ordering 47, 102, 120
 - class zero 121
 - combining marks 225–226
 - definition 78
 - dependence 225
 - display order 43
 - keyboard input 153
 - ligatures 45
 - multiple 43
 - multiple base characters 45
 - normalization of 145
 - ordering conventions 42
 - rendering of marks 155–159
 - reordrant 121
 - script-specific 42
 - split 122
 - strikethrough 124
 - subjoined 123
 - typographical interaction 43, 120
 - vertical stacking 43
 - see also* diacritics
- combining class 100
- Combining Class (normative property) 120
- combining classes 120, 157–158
 - class zero characters 120
 - definition 100
- combining grapheme joiner (U+034F) 507
- combining half marks 133, 230
- combining marks *see* combining characters
- Compatibility and Specials Area 20, 38
- compatibility characters 18
- compatibility composite characters 21
 - see* compatibility decomposable characters
- compatibility decomposable characters 20
 - definition 86
- compatibility decomposition 48
 - definition 85
- compatibility decomposition mappings 85
- Compatibility Encoding Scheme for UTF-16
 - see* UTR #26, Compatibility Encoding Scheme for UTF-16: 8-Bit (CESU-8)
- compatibility equivalence
 - definition 86
- compatibility equivalent character sequences
 - conformance 60
- compatibility mappings
 - see* compatibility decomposition mappings
- compatibility precomposed characters
 - see* compatibility decomposable characters
- compatibility variants 20
 - mapping 170
- composite characters
 - see* decomposable characters
 - compatibility *see* compatibility decomposable characters
- Composition Exclusion (normative property) 73
- compression 145
 - see also* UTS #6, A Standard Compression Scheme for Unicode (SCSU)
- conferences 545
- conformance 55–113
 - clause and definition updates 559
 - definitions 64–68
 - examples 51
 - ISO/IEC 10646 implementations 555
 - requirements 59–62
- confusables 172
- conjunct consonants
 - Indic 151, 265
 - Myanmar 343
 - selection of clusters 151
- contextual shaping
 - apostrophe 190
 - Arabic 238
 - not used for Hebrew final forms 235
 - quotation marks 187
 - Syriac 254
- contour tones 224
- control codes 23, 51, 500
 - graphics for 476
 - names 132
 - properties 501
 - semantics 24, 501
 - specified in Unicode 501
- control sequences 500
- conversion of characters 137–139
- convertibility
 - as Unicode design principle 19
- Coptic 213, 215–217
- corporate use subarea 512
- corrigenda 57
- CR (U+000D carriage return) 146, 501
- CRLF (carriage return and line feed) 146

- Croatian 204
 - digraphs 204
 - culturally expected sorting 10, 160
 - Cuneiform
 - Old Persian 452
 - Sumero-Akkadian 452–455
 - Ugaritic 451
 - Cuneiform and Hieroglyphic Area 39
 - currency symbols 460–462
 - encoded in script blocks 461
 - cursive joining 503–507
 - Arabic 243–248
 - control characters for 134, 239–240, 415, 503
 - Mongolian 414–416
 - N’Ko 424
 - Syriac 254–257
 - transparency 506
 - cursive scripts 233
 - Cypriot 443–444
 - see also* Linear B
 - Cyrillic 217–219
 - for Mongolian 413
 - Czech 204
- D**
- danda, in Devanagari block 275
 - Danish 203
 - dashes 185
 - Database, Unicode Character
 - see* Unicode Character Database (UCD)
 - dead consonants, Indic 265
 - dead keys 153
 - decomposable characters 48
 - definition 85
 - normalization of 144
 - decomposition 48, 85–87
 - canonical *see* canonical decomposition
 - compatibility *see* compatibility decomposition
 - definition 85
 - in normalization 144
 - mapping, definition 85
 - mappings noted in code charts 530
 - default case
 - algorithms 63
 - default case algorithms 110–113
 - default case conversion 111
 - default case detection 112
 - default caseless matching 113
 - default casing algorithms 63
 - default grapheme clusters 151
 - see also* UAX #29, Unicode Text Segmentation
 - default ignorable code points 141, 173
 - Default Ignorable Code Points (property) 173
 - default property values 141
 - definition 71
 - defective combining character sequences 156
 - definition 80
 - dependent vowel signs
 - Indic 264
 - Khmer 352
 - Philippine scripts 366
 - deprecated characters 55, 528
 - alternate format 134, 509–510
 - definition 67
 - Derived Age (property) 141
 - derived properties
 - definition 76
 - DerivedCoreProperties.txt 110, 118, 173
 - DerivedNormalizationProps.txt 170
 - Deseret 429–431
 - design goals of Unicode 3
 - design principles of Unicode 10–19
 - designated code points 23
 - Devanagari 262–279
 - Dhivehi 259
 - diacritics 41, 225
 - alternative glyphs 200, 226
 - Czech 201
 - display in isolation 45, 184, 226
 - double 84, 133, 227
 - Greek 210–211, 213
 - Latin 200–202
 - Latvian 201
 - mathematical 466
 - on i and j 201
 - rendering 155–159
 - Romanian 201
 - Slovak 201
 - spacing clones of 224, 227
 - symbol 42, 230
 - Turkish 201
 - see also* combining characters
 - dictionary symbols 482
 - digit form names 241
 - digits 126, 143
 - Arabic-Indic 240–241
 - national shapes 510
 - digraphs 203, 204, 206, 207
 - dingbats 483–484
 - directionality 15, 40
 - East Asian scripts 380
 - Middle Eastern scripts 233
 - Mongolian 414
 - musical symbols 493
 - normative property 124
 - Ogham 436
 - Old Italic 438
 - Philippine scripts 367
 - Runic 439
 - discussion list for Unicode 545
 - Domino Tiles 485
 - dotless i 166, 201
 - dotted circle
 - in code charts 79, 226
 - in fallback rendering 155
 - to indicate diacritic 41
 - to indicate vowel sign placement 42
 - double diacritics 84, 133, 227
 - Dutch 203
 - dynamic composition
 - as Unicode design principle 18
 - Dzongkha 307
- E**
- East Asian scripts 379–408
 - writing direction 40
 - see also* CJK ideographs
 - Eastern Arabic-Indic digits 240

- EBCDIC
 newline function147
see UTR #16, UTF-EBCDIC
- editing, text boundaries for 151–152
- efficiency
 as Unicode design principle11
- Egyptian Hieroglyphs 455–458
- e-mail discussion list for Unicode545
- Enclosed Alphanumerics489
- enclosing marks230
 definition79
- encoded characters5, 22
 allocation 32–40, 557
 definition66
- encoding form conversion
 definition94
- encoding forms24–30
 ISO/IEC 10646 definitions552
- encoding forms, Unicode
see Unicode encoding forms
- encoding model for Unicode characters24, 32
see also UTR #17, Unicode Character Encoding Model
- encoding schemes30–32
- encoding schemes, Unicode
see Unicode encoding schemes
- endian ordering
see byte order mark (BOM) (U+FEFF)
- end-user subarea513
- English203
- equivalent sequences144
 as Unicode design principle18
 case-insensitivity162, 168
 combining characters in matching153
 conformance61
 Hangul syllables405
 in sorting and searching160
 language-specific87
 security implications171
see also canonical equivalence
see also compatibility equivalence
see also encoding forms, encoding schemes
- errata xxxvi, 57, 546
- escape sequences500
 not used in Unicode1, 3
- Esperanto204
- Estonian204
- Ethiopic410–412
- Etruscan437
- euro sign (U+20AC)461
- European alphabetic scripts199–231
- eyelash-RA270
- F**
- fallback rendering of nonspacing marks155
- FAQ (Frequently Asked Questions)546
- Faroese203
- Farsi238, 239
- featural syllabaries179
- FF (U+000C form feed)146, 501
- file separator (U+001C)501
- Finnish203
- Finno-Ugric Transcription (FUT)
see Uralic Phonetic Alphabet (UPA)
- fixed-width Unicode encoding form (UTF-32) .. 26, 92
- flat tables138
- Flemish203
- fonts
 and Unicode characters12
 for mathematical alphabets466–467
 style variation for symbols459
- form feed (U+000C) (FF)146, 501
- format control characters23, 51, 183, 499–523
 deprecated509–510
 prefixed134
 reserved ranges141
 stateful174
- fraction characters467
- fraction slash (U+2044)190, 467
- French204
- Frisian204
- FTP site, Unicode Consortium545
- fullwidth forms in East Asian encodings402–403
- futhark440
- G**
- Garshuni251
- Ge'ez410
- General Category (normative property)124
 list of values124
- general punctuation181–198
- General Scripts Area38
- geometrical symbols480–481
- Georgian221–222
- German203
- geta mark (U+3013)197
- Glagolitic219–220
- Glossary546
- glyph selection tables138
- glyphs5, 12
 characters, relationship to11
 diacritics alternative200, 226
 Greek alternative211–212
 Latin alternative200
 mathematical alternative472
 missing173
 representative in code charts526
 standardized variants511
 symbols alternative459
- golden numbers441
- Gothic441–442
- grapheme base225
 definition80
- grapheme clusters9, 46
see also UAX #29, Unicode Text Segmentation
- default151
 definition80
- grapheme extender
 definition80
- grapheme joiner, combining (U+034F)507
- graphic characters23
- Greek209–214
 acrophonic numerals143, 214
 alternative glyphs211–212
 ancient musical notation496–498
 letters as symbols211–212, 472
see also Cypriot, Linear B
- Greenlandic204

- group separator (U+001D) 501
 guillemets 187
 Gujarati 286–287
 Gurmukhi 283–286
- H**
- Hakka 401
 halant 261
 see also virama
 half marks, combining 133, 230
 half-consonants, Indic 266
 halfwidth forms in East Asian encodings ... 402–403
 Han ideographs *see* CJK ideographs
 Han unification 388–393
 and language tags 150
 history 565–567
 language usage 386
 source separation rule 384, 389
 source standards 381–383
 Hangul Area 38
 Hangul syllables 379, 403–406
 as grapheme clusters 46
 boundary determination 106
 canonical decomposition 109
 collation 406
 composition 108
 conjoining jamo 104–109
 equivalent sequences 405
 Hangul Compatibility Jamo 404–405
 Hangul Jamo 403–406
 Hangul Syllables block 404–406
 Johab set 405
 name generation 109
 normalization 404
 precomposed 106
 standard 106
 Hangzhou numerals 469
 Hanja *see* CJK ideographs
 Hanunóo 366
 Hanzi *see* CJK ideographs
 harakat, Arabic pronunciation marks 238
 hasant 279
 hash tables 138
 Hebrew 234–238
 high surrogate
 definition 87
 high-surrogate code points 59, 513
 high-surrogate code units 87
 higher-level protocols
 definition 68
 Hindi 262
 Hiragana 401–402
 historic scripts 435–444
 horizontal tab (U+0009) 501
 HTML newline function 147
 Hungarian 204
 hyphenation 503
 as a text process 8
 hyphens 185, 503
- I**
- I Ching symbols 485
 IANA charset names 31
 Icelandic 203
- identifiers 160
 see also UAX #31, Unicode Identifier and Pattern
 Syntax
 Ideographic (informative property) 133
 Ideographic Rapporteur Group (IRG) 381, 566
 Ideographic Variation Database *see* UTS #37, Unicode
 Ideographic Variation Database
 ideographs *see* CJK ideographs
 IICore 384, 566
 ill-formed
 definition 89
 Imperial Aramaic 448
 implementation guidelines 137–175
 in a Unicode encoding form
 definition 90
 in-band mechanisms 522
 Indic scripts 261–304, 305–307
 principles, in terms of Devanagari 263–269
 relation to ISCII standard 262
 Indonesian 203
 industry character sets
 covered in Unicode 2
 information separators (U+001C..U+001F) 501
 informative properties
 definition 73
 Inscriptional Pahlavi 449
 Inscriptional Parthian 449
 inside-out rule 155
 interchange restrictions 24
 International Phonetic Alphabet (IPA) ..178, 205–206
 Spacing Modifier Letters 224
 see also phonetic alphabets
 internationalization 13
 Internationalization & Unicode Conference 545
 Internet protocols
 UTF-8 as preferred encoding 28
 Inuktitut 428
 invisible operators 476
 iota subscript 211
 IPA *see* International Phonetic Alphabet
 IRG (Ideographic Rapporteur Group) 381, 566
 Irish 203, 436
 ISCII standard and Unicode 262
 ISO/IEC 10646 549–555
 codespace 553
 conformance of Unicode implementations .. 555
 encoding forms 552
 synchrony with Unicode Standard 554
 timeline compared to Unicode versions 550
 UCS transformation formats (UTF) 553
 Italian 203
 ITC Zapf Dingbats 483
 IUC *see* Internationalization & Unicode Conference
- J**
- Jamo.txt 109
 jamos *see* Hangul syllables
 Japanese 379
 Javanese 373–375
 Jawi 249
 Johab 405
 joiners 239
 combining grapheme joiner (U+034F) 507
 word joiner (U+2060) 502

zero width joiner (U+200D) 239–240, 504
 justification 157

K

Kaithi 327–329
 Kana (Hiragana and Katakana) 401–402
 Kanbun 394
 KangXi radicals 394–395, 535
 Kanji *see* CJK ideographs
 Kannada 297–299
 Katakana 402
 Kawi 368, 370
 Kayah Li 363–364
 KC (normalization form)
 see Normalization Form KC
 KD (normalization form)
 see Normalization Form KD
 keytop labels 477
 Khamti Shan 346
 Kharoshthi 332–333
 Khmer 348–357
 characters not recommended 354
 syllable components, order of 355
 killer
 see also virama
 Korean Hangul *see* Hangul
 Kurdish 238

L

Ladino 234
 language tags 149, 519–522
 and Han unification 150
 use strongly discouraged 522
 Lanna 358
 Lao 340–342
 last-resort glyphs 173
 Latin 200–209
 alternative glyphs 200
 Basic Latin 203
 encoding blocks 34
 IPA Extensions 205–206
 Latin Extended Additional 208–209
 Latin Extended-A 203
 Latin Extended-B 204–205
 Latin Extended-C 208
 Latin Extended-D 208
 Latin Ligatures 208
 Latin-1 Supplement 203
 Phonetic Extensions 207–208
 Latvian 204
 layout control characters 51, 501–509
 leading surrogates
 see high-surrogate code units 15
 Lepcha 316–318
 letter spacing 502
 letterlike symbols 462–467
 LF (U+000A line feed) 146, 501
 ligatures 503–507
 Arabic 245–246
 combining characters on 45
 control characters for 134
 for nonspacing marks 158
 Latin 208

 selection 152
 Syriac 257
 Limbu 323–326
 line breaking 146–149, 502–503
 control characters 135
 in South Asian scripts 340, 345, 357
 recommendations 148
 see also UAX #14, Unicode Line Breaking Algorithm
 line feed (U+000A) (LF) 146, 501
 line separator (U+2028) (LS) 146, 503
 line tabulation (U+000B) (VT) 501
 Linear B 442–443
 see also Cypriot
 linear boundaries 152
 Lisu 432–434
 Lithuanian 204
 little-endian 30
 definition 62
 Locale Data Markup Language
 see UTS #35, Unicode Locale Data Markup Language (LDML)
 logical order
 as Unicode design principle 15
 exceptions to 122
 logosyllabaries 180
 low surrogate
 definition 87
 low-surrogate code points 59, 513
 low-surrogate code units 87
 lowercase 118, 164, 199
 LS (U+2028 line separator) 146, 503
 Lycian 444
 Lydian 444

M

MacOS newline function 147
 Mahjong Tiles 484
 mail discussion list for Unicode 545
 major version 56
 Malay 203
 Malayalam 299–304
 Maltese 204
 Manchu 413
 Mandarin 387
 Manden 421
 map symbols 482
 mapping tables *see* tables of character data
 Marathi 262, 270, 274
 markup languages
 and Unicode conformance 522
 line breaking 146
 see also UTR #20, Unicode in XML and Other Markup Languages
 Mathematical (informative property) 471
 mathematical expression format characters 134
 see also UTR #25, Unicode Support for Mathematics
 mathematical symbols 471–475
 alphabets 463–467
 alphanumeric 463–467
 fonts 466–467
 format characters 476
 fragments for typesetting 478

invisible operators	476
operators	471–473
standardized variants	475
MathML	473
matras	121, 264
Meetei Mayek	330–331
Middle Eastern scripts	233–260
Min	387
Minnan (Hokkien/Fujian, incl. Taiwanese)	401
minor version	56
minus sign	473
commercial (U+2052)	193
mirrored property	
<i>see</i> Bidi Mirrored (normative property)	
mirroring of paired punctuation	187
Miscellaneous Symbols	482
missing glyphs	173
modifier letters	222–225
Modifier Letters, Spacing	207
Mongolian	318, 412–419
writing direction	414
multibyte encodings	
compared to UTF-8	28
multistage tables	138
musical symbols	492–498
ancient Greek	496–498
Balinese	372
Byzantine	496
directionality	493
Gregorian	493
Western	492–495
Myanmar	342–348
Myanmar Extended-A	345

N

N’Ko	421–425
named character sequences	129
names, character <i>see</i> character names	
namespace	65
NEL (U+0085 next line)	146, 501
Nepali	262
neutral directional characters	124
New Tai Lue	358–359
newline function (NLF)	147, 501
newline guidelines	146–149
next line (U+0085) (NEL)	146, 501
NFC (Normalization Form C)	47
NFD (Normalization Form D)	47
NFKC (Normalization Form KC)	47
NFKD (Normalization Form KD)	47
NLF (newline function)	147, 501
no-break space (U+00A0)	502
base for diacritic in isolation	45, 184, 226
no-break space, narrow (U+202F)	418
noncharacter code points <i>see</i> noncharacters	
noncharacters	23, 50, 514
conformance	59
definition	67
handling	61
in code charts	528
interchange restrictions	24
semantics	24
U+10FFFF (not a character code)	514
U+FDD0..U+FDEF	23, 514

U+FFFE (not a character code)	50, 514
U+FFFF (not a character code)	23, 514
nondecomposable characters	48
non-joiner, zero width (U+200C)	239–240, 505
nonlinear boundaries	152
non-overlap principle in Unicode encoding forms	25
nonspacing marks	225
definition	79
display in isolation	45, 184, 226
positioning	158
rendering	155–159
<i>see also</i> combining characters	
<i>see also</i> diacritics	
normalization	47, 144–145
and case operations	169
canonical ordering algorithm	47, 102, 120
conformance	62
of private-use characters	512
<i>see also</i> UAX #15, Unicode Normalization Forms	
Normalization Form C (NFC)	47
Normalization Form D (NFD)	47
Normalization Form KC (NFKC)	47
Normalization Form KD (NFKD)	47
normalization forms	99–104
definition	104
specification	101
normalization stability	99
normative behaviors	
definition	64
normative properties	
definition	72
list	73
may change	72
Norwegian	203
notational conventions	537–540
notational systems	181
nukta	271
null (U+0000)	
as Unicode string terminator	501
number forms	467–470
CJK ideographs	143
numbers	
handling	143
ideographic accounting	126
numerals	
acrophonic	214
Chinese counting rods	469
Coptic	217
Cuneiform	455
Greek acrophonic	143
Hangzhou	469
old-style	191
Roman	143, 467
Rumi	468
Suzhou-style	469
numeric separators	193
numeric shape selectors (deprecated)	510
Numeric Value (normative property)	126
numero sign (U+2116)	462

O

object replacement character (U+FFFC)	518
octet	539
Ogham	436–437

- Ol Chiki 331–332
 Old Italic 437–439
 Old Persian 452
 Old South Arabian 445–447
 Old Turkic 442
 old-style numerals 191
 Oriya 287–289
 Oromo 410
 Osmanya 419
 out-of-band mechanisms 522
 overlapping encodings 25
 overscores 190
- P**
- Panjabi 283
 paragraph or section marks 192
 paragraph separator (U+2029) (PS) 146, 503
 Pashto 238
 Persian 238, 239
 Phags-pa 318–323
 Phaistos Disc symbols 486
 Phake 347
 Philippine scripts 366–367
 Phoenician 447
 phonemes 180
 phonetic alphabets 178
 IPA Extensions 205–206
 Phonetic Extensions 207–208
 Spacing Modifier Letters 224–225
 Uralic Phonetic Alphabet (UPA) 193, 207
 see also International Phonetic Alphabet (IPA)
 phonetic extensions 208
 Pinyin 202
 pivot code, Unicode as 137
 plain text
 as Unicode design principle 14
 legibility criterion 15
 planes of Unicode codespace 33
 Plane 0 (BMP) 33
 Plane 1 (SMP) 33, 39
 Plane 14 (SSP) 33
 Plane 2 (SIP) 33, 39
 Planes 15–16 (Private Use) 40, 513
 points, Hebrew pronunciation marks 234
 policies of the Unicode Consortium 546
 Polish 204
 Portuguese 203
 precomposed characters
 see decomposable characters
 compatibility *see* compatibility decomposable
 characters
 prefixed format control characters 134
 Private Use Area (PUA) 38, 512
 Private Use planes 33, 40, 513
 private-use characters
 semantics 24
 private-use code points 23, 140
 conformance 60
 definition 77
 high surrogates 513
 processing code, choice of Unicode encoding form 28
 properties 13, 69–77, 115–135
 aliases 117
 aliases (definition) 76
 and Unicode algorithms 73
 data tables 138
 derived *see* derived properties
 in Unicode Character Database (UCD) 34
 informative *see* informative properties
 normative references to 58, 63
 normative *see* normative properties
 of control codes 501
 provisional *see* provisional properties
 simple *see* simple properties
 see also individual properties, e.g. combining
 classes
 property values
 aliases 117
 aliases (definition) 77
 default 141, 512
 default (definition) 71
 normative references to 63
 PropertyAliases.txt 77, 539
 PropertyValueAliases.txt 77, 539
 PropList.txt 120
 Provençal 204
 provisional properties
 definition 74
 PS (U+2029 paragraph separator) 146, 503
 PUA (Private Use Area) 38, 512
 pulli 290
 punctuation 181–198
 blocks containing 177
 CJK 196
 doubled 190
 in bidirectional text 182
 paired 187
 small form variants 198
 typographic forms 182
 vertical forms 198
 Punctuation and Symbols Area 38
 Punjabi 283
- Q**
- quotation marks 187–189
 East Asian 188
 European 187
- R**
- radicals, KangXi and other CJK 394–395
 radical-stroke index 535
 record separator (U+001E) 501
 recycling symbols 482
 referencing 63
 properties 58
 Unicode algorithms 58
 Unicode Standard 57
 regular expressions 149
 and line breaking 146
 see also UTS #18, Unicode Regular Expressions
 Rejang 375–376
 rendering of text 5, 8, 13
 unsupported characters 141
 repertoire of abstract characters 22
 replacement character (U+FFFD) 32, 51, 62, 518
 reserved code points 23, 140
 definition 67
 in code charts 528

- preservation in interchange 24
 - see also* unassigned code points
- Rhaeto-Romanic 204
- rich text 14
- right single quotation mark (U+2019)
 - preferred for apostrophe 189
- right-to-left text 40
 - East Asian scripts 380
 - Middle Eastern scripts 233
- roadmap for script additions 34
- Roman numerals 143, 467
- Romanian 204
- Romany 204
- Rumi numeral forms 468
- Runic 439–441
- Russian 217

- S**
- Samaritan 257–259
- Sami 204
- Sanskrit 262
- Saurashtra 329–330
- scalar values, Unicode
 - see* Unicode scalar values
- scripts
 - in Unicode Standard 2
 - roadmap for future additions 34
 - types of 181
 - see also* UAX #24, Unicode Script Property
- SCSU
 - see* UTS #6, A Standard Compression Scheme for Unicode
- searching 160–162
 - as a text process 8
 - case-insensitive 162, 168
- section or paragraph marks 192
- security issues 171
- self-synchronization of encoding forms 26
- semantics
 - see* character semantics
- sequences
 - notation 538
- Serbian
 - corresponding digraphs in Croatian 204
- Shan 357
- Shavian 431
- Show Hidden 60, 155, 174, 511
- SHY (U+00AD soft hyphen) 503
- Sibe 413
- signature for Unicode data 51, 515–516
- simple properties
 - definition 76
- simplified Chinese 386
- Sindhi 238, 276
- Sinhala 306–307
- SIP (Supplementary Ideographic Plane) 33, 39
- slash, fraction (U+2044) 190
- Slovak 204
- Slovenian 204
- small letters 118, 164, 199
- SMP (Supplementary Multilingual Plane) 33, 39
- soft hyphen (U+00AD) (SHY) 503
- Somali 419
- Sorbian 204

- sorting 10, 160
 - and combining grapheme joiner 508
 - as a text process 8
 - case-insensitive 161
 - culturally expected 10, 160
 - language-insensitive 161
 - see also* Unicode Collation Algorithm (UCA)
- source separation rule 384, 389
- South Asian scripts 261–304, 305–326
- Southeast Asian scripts 337–367
- space (U+0020)
 - base for diacritic in isolation 46, 184, 226
- space characters 184, 502–503
 - graphics for 477
- space, zero width (U+200B) 185
- spacing clones of diacritics 224, 227
- spacing marks 225
 - definition 79
- Spacing Modifier Letters 224–225
- Spanish 203
- special characters 50, 499–523
- SpecialCasing.txt 110, 120
- Specials 515–519
- spell-checking
 - as a text process 8
- spellings, alternative
 - see* equivalent sequences
- spoofing 172
- SSP (Supplementary Special-purpose Plane) 33
- stability 75, 116
 - as Unicode design principle 18
- stacked boundaries 151
- stacking sequences 43
 - nondefault 44
- Standard Compression Scheme for Unicode (SCSU)
 - see* UTS #6, A Standard Compression Scheme for Unicode
- standard Korean syllables 106
- standardized variants 416, 511
 - mathematical symbols 475
- StandardizedVariants.txt 416, 475
- standards coverage 2
- starters 101
- stateful encoding
 - not used in Unicode 3
 - paired format controls 174
- string comparison 10
- string literals, Unicode
 - code point notation `\u1234` 539
- strings, Unicode 32, 89
 - null termination 501
- strong directional characters 124
- styled text 14
- sublinear searching 162
- subsets, supported 53
 - conformance 60
 - ISO/IEC 10646 specification for 554
- substitution character 51
- Sumero-Akkadian 452–455
- Sundanese 376–377
- superscripts and subscripts 470
- supplementary characters
 - in UTF-16 strings 32
 - tables for 138
- Supplementary General Scripts Area 38

- Supplementary Ideographic Plane (SIP)33, 39
 - Supplementary Multilingual Plane (SMP)33, 39
 - supplementary planes
 - representation in UTF-1627
 - representation in UTF-828
 - Supplementary Private Use Areas40, 513
 - Supplementary Special-purpose Plane (SSP)33
 - supported subsets53
 - conformance60
 - supralineation216
 - surrogate code points
 - see surrogates
 - surrogate pairs27, 92
 - definition87
 - processing29, 142–143
 - surrogates24, 87–88, 513
 - interchange restrictions24
 - isolated surrogates, handling32
 - isolated surrogates, ill-formed92
 - isolated surrogates, uninterpreted87
 - support levels142
 - Surrogates Area38, 513
 - Suzhou-style numerals469
 - svasti signs313
 - Swahili203
 - Swedish203
 - syllabaries178
 - alphabetic property132
 - featural179
 - Syoti Nagri326–327
 - symbols459–498
 - appearance variation459
 - arrows474–475
 - currency460–462
 - dictionary482
 - dingbats483–484
 - Enclosed Alphanumerics489
 - fragments for mathematical typesetting478
 - geometrical480–481
 - Khmer lunar calendar357
 - letterlike462–467
 - map482
 - mathematical471–475
 - mathematical alphanumeric463–467
 - miscellaneous482
 - musical492–498
 - number forms467–470
 - recycling482
 - technical476–480
 - weather482
 - symmetric swapping format characters (deprecated)509
 - Syriac250–257
- T**
- tab (U+0009 character tabulation)501
 - tab, vertical (U+000B)146, 501
 - tables of character data137–139
 - optimization138
 - supplementary characters138
 - tag characters519–523
 - Tagalog366
 - Tagbanwa366
 - tags, language149, 519–522
 - use strongly discouraged522
 - Tai Le357–358
 - Tai Tham359–361
 - Tai Viet361–363
 - Tai Xuan Jing symbols485
 - Tamil289–296
 - TCHAR in Win32 API140
 - Technical Notes (UTN)544
 - Technical Reports (UTR)542
 - abstracts543
 - Technical Standards (UTS)xxxvi, 542
 - abstracts542
 - technical symbols476–480
 - Telugu296–297
 - terminal emulation460
 - text boundaries8, 46, 133, 151–152, 159
 - see also UAX #14, Unicode Line Breaking Algorithm
 - see also UAX #29, Unicode Text Boundaries
 - text elements5, 8, 151
 - boundaries159
 - for sorting161
 - variable-width nature29
 - text processes4, 8–10
 - text rendering5, 8, 13
 - text selection, boundaries for151–152
 - Thaana259–260
 - Thai338–340
 - Tibetan307–316
 - Tifinagh420
 - Tigre410
 - tilde (U+007E)193
 - titlecase118, 165
 - Todo413
 - tone letters224–225
 - tone marks
 - Bopomofo spacing400, 401
 - Chinantec225
 - Chinese225
 - Tai Le357
 - Thai338
 - Vietnamese202
 - traditional Chinese386
 - traffic signs482
 - trailing surrogates
 - see low-surrogate code units
 - transcoding137–139
 - tables138
 - triangulation in transcoding137
 - truncation
 - combining character sequences153–154
 - surrogates and143
 - Turkish204
 - case mapping of I166, 201
 - two-stage tables138
- U**
- U+ notation539
 - U+10FFFF (not a character code)514
 - U+FEFF (BOM)515–516
 - U+FFFE (not a character code)514
 - U+FFFF (not a character code)514
 - UAX (Unicode Standard Annex)xxxiv, 542

- as component of Unicode Standard 59
 - conformance 63
 - list of 63
- UCA *see* Unicode Collation Algorithm
- UCD *see* Unicode Character Database
- UCS (Universal Character Set)
 - see* ISO/IEC 10646
- UCS-2 553
- UCS-4 552
- Ugaritic 451–452
- Uighur 318, 413
- Ukrainian 217
- unassigned code points 23, 59, 141
 - defined as reserved code points 67
 - handling 55
 - properties of 141
 - semantics 59
 - see also* reserved code points
- underscores 190
- undesignated code points 23
- Unicode 1.0 Name (informative property) 132
- Unicode algorithms
 - and properties 73
 - conformance 63
 - definition 68
 - normative references to 58, 63
- Unicode Bidirectional Algorithm 16, 40
 - see also* UAX #9, Unicode Bidirectional Algorithm
- Unicode Character Database (UCD) xxxv, 116, 546
 - as component of Unicode Standard 59
 - changes 56
 - properties in 34
- Unicode character encoding model 24, 32
 - see also* UTR #17, Unicode Character Encoding Model
- Unicode character literals
 - code point notation U+ 539
- Unicode codespace
 - allocation numbers 557
 - definition 66
 - planes 33
 - same as ISO/IEC 10646 553
 - size 1, 22
- Unicode Collation Algorithm (UCA) 10
 - see also* UTS #10, Unicode Collation Algorithm
- Unicode Common Locale Data Repository (CLDR) 546
- Unicode conferences 545
- Unicode Consortium 541
 - addresses 546
 - Consortium membership in standards bodies 541
 - e-mail discussion list 545
 - FTP site 545
 - membership 541
 - policies 546
 - Web site xxxv, 545
- Unicode data signature 51, 515–516
- Unicode data types 139–140
 - for C 139–140
- Unicode encoding forms 88–94
 - advantages of each 28
 - conformance 26, 61
 - definition 89
 - fixed-width (UTF-32) 26, 92
 - signatures 516, 517
 - variable-width 27, 28, 92
 - see also* encoding forms
- Unicode encoding schemes
 - conformance 95–99
 - definition 95
 - endian ordering 30
 - see also* encoding schemes
- Unicode escape sequence notation \u1234 539
- Unicode Regular Expressions *see* UTS #18, Unicode Regular Expressions
- Unicode scalar values
 - definition 88
- Unicode security mechanisms
 - see also* UTS #39, Unicode Security Mechanisms
 - Unicode security 171
- Unicode Standard
 - allocation of encoded characters 32–40
 - application areas 3
 - architecture 7–10
 - areas 34
 - benefits 1
 - blocks 34, 177
 - code charts 525–533, 545
 - components 59
 - conformance 55–113
 - conformance of ISO/IEC 10646 implementations 555
 - corrections 57
 - definitions for conformance 64–68
 - design goals 3
 - design principles 10–19
 - errata 57, 546
 - normative references to 57, 63
 - number of characters 2, 557
 - number of code points 1, 22
 - script coverage 2
 - security issues 171
 - synchrony with ISO/IEC 10646 554
 - updates 546
 - user community 3
 - versions *see* versions of the Unicode Standard
 - see also* Version 5.2
- Unicode Standard Annexes (UAX) xxxiv, 542
 - as components of Unicode Standard 59
 - conformance 63
 - list of 63
- Unicode string literals
 - code point notation \u1234 539
- Unicode strings 32
 - definition 89
- Unicode Technical Committee (UTC) 541
- Unicode Technical Notes (UTN) 544
- Unicode Technical Reports (UTR) 542
 - abstracts 543
- Unicode Technical Standards (UTS) xxxvi, 542
 - abstracts 542
- UnicodeData.txt 110, 120
- unification
 - as Unicode design principle 17
 - see also* Han unification
- Unified CJK Ideograph (property) 397
- Unified Repertoire and Ordering (URO) 389, 566
 - see also* Han unification
- Unihan Database 116, 392, 393, 531, 546, 567
- Unihan.txt 74

- Unihan.zip116
 - unit separator (U+001F)501
 - Universal Character Set (UCS)
 - see ISO/IEC 10646
 - universality
 - as Unicode design principle10
 - Unix
 - and UTFs29
 - newline function147
 - UTF-32 in27
 - UTF-8 in14
 - unsupported characters140–141
 - update version57
 - uppercase118, 164, 199
 - Uralic Phonetic Alphabet (UPA)193, 207
 - Urdu238
 - URO (Unified Repertoire and Ordering) ...389, 566
 - see also Han unification
 - UTF, Unicode Transformation Formats24, 89
 - advantages of each28
 - as encoding form or scheme98
 - binary comparison and sort order differences ...
 - 161, 162
 - in APIs140
 - in ISO/IEC 10646553
 - UTF-1627, 92
 - binary comparison and sort order caution ...27
 - bit distribution (table)92
 - BOM in96, 515
 - encoding form (definition)92
 - encoding scheme (definition)96
 - encoding schemes30
 - in ISO/IEC 10646554
 - in UTF-8 order163
 - surrogates and string handling32, 142
 - UTF-16BE (Big-endian)516
 - encoding scheme31
 - encoding scheme (definition)96
 - UTF-16LE (Little-endian)516
 - encoding scheme31
 - encoding scheme (definition)96
 - UTF-3226, 92
 - BOM in97
 - encoding form (definition)92
 - encoding scheme (definition)97
 - encoding schemes30
 - in Unix27
 - UTF-32BE (Big-endian)
 - encoding scheme31
 - encoding scheme (definition)97
 - UTF-32LE (Little-endian)
 - encoding scheme31
 - encoding scheme (definition)97
 - UTF-828, 92
 - ASCII transparency28
 - binary comparison and sort order29
 - bit distribution (table)93
 - BOM in96, 99, 515
 - byte ranges93
 - compared to multibyte encodings28
 - encoding form (definition)92
 - encoding scheme30
 - encoding scheme (definition)96
 - in ISO/IEC 10646553
 - in Unix14
 - in UTF-16 order163
 - non-shortest form is invalid93, 172
 - preferred encoding for Internet protocols ...28
 - security and172
 - signature96, 99, 515
 - UTF-EBCDIC
 - see UTR #16, UTF-EBCDIC
 - UTN (Unicode Technical Note)544
 - UTR (Unicode Technical Report)542
 - abstracts543
 - UTS (Unicode Technical Standard)xxxvi, 542
 - abstracts542
- ## V
- Vai425–426
 - valid (synonym for well-formed)91
 - variable-width Unicode encoding form ...27, 28, 92
 - variants
 - compatibility20
 - fullwidth and halfwidth198
 - mathematical symbols475
 - small form198
 - standardized511
 - variation selectors135, 510
 - ideographic variation mark (U+303E)399
 - Mongolian free variation selectors416
 - variation sequences510
 - for Phags-pa322–323
 - Version 5.059
 - correlation with ISO/IEC 10646552
 - number of characters557
 - Version 5.2
 - number of characters2
 - versions of the Unicode Standard ...xxv, 55, 546, 557–558
 - backward compatibility55
 - compared to ISO/IEC 10646 editions557
 - content56
 - interaction in implementations141
 - numbering56
 - property changes56
 - stability56
 - updates546
 - vertical tab (U+000B)146, 501
 - vertical text40, 182, 198
 - East Asian scripts380
 - Mongolian414
 - Vietnamese202, 208
 - ideographs379
 - virama179, 261
 - definition265
 - Kharoshthi336
 - Khmer351
 - Philippine scripts366
 - virama-like characters134
 - visual order used for Thai and Lao17
 - vowel harmony
 - Mongolian417
 - vowel marks, Middle Eastern scripts233
 - vowel separator
 - Mongolian418
 - vowel signs
 - Indic42, 264
 - Khmer352

Philippine scripts 366

W

wchar_t
 and Unicode encoding forms 29
 in C language 140
 weak directional characters 124
 weather symbols 482
 Web site, Unicode Consortium xxxv, 545
 Weierstrass elliptic function symbol 463
 well-formed
 definition 90
 Welsh 204
 Where Is My Character? 546
 wide characters
 data type in C 140
 wiggly fence (U+29DB) 474
 Windows newline function 147
 word breaks 153, 502–503
 in South Asian scripts 340, 345, 357
 word joiner (U+2060) 502
 writing direction *see* directionality
 writing systems 178–181
 Wu (Shanghainese) 387

X

Xibe 413
 Xishuang Banna Dai 358
 XML
see UTR #20, Unicode in XML and Other Markup
 Languages

Y

yen currency sign 461
 Yi 406–408
 Yiddish 234
 Yijing Hexagram Symbols 485
 ypogegrammeni 211
 yuan currency sign 461

Z

Zapf Dingbats 483
 zero extension relation among encodings 553
 zero width joiner (U+200D) 239–240, 504
 zero width no-break space (U+FEFF) ... 50, 62, 502
 initial 98, 516
 zero width non-joiner (U+200C) 239–240, 505
 zero width space (U+200B) 502
 for word breaks in South Asian scripts .. 340, 345,
 357
 zero-width space characters 503
 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)

