

Electronic Edition

This file is part of the electronic edition of *The Unicode Standard, Version 5.0*, provided for online access, content searching, and accessibility. It may not be printed. Bookmarks linking to specific chapters or sections of the whole Unicode Standard are available at

<http://www.unicode.org/versions/Unicode5.0.0/bookmarks.html>

Purchasing the Book

For convenient access to the full text of the standard as a useful reference book, we recommend purchasing the printed version. The book is available from the Unicode Consortium, the publisher, and booksellers. Purchase of the standard in book format contributes to the ongoing work of the Unicode Consortium. Details about the book publication and ordering information may be found at

<http://www.unicode.org/book/aboutbook.html>

Joining Unicode

You or your organization may benefit by joining the Unicode Consortium: for more information, see [Joining the Unicode Consortium](http://www.unicode.org/consortium/join.html) at

<http://www.unicode.org/consortium/join.html>

This PDF file is an excerpt from *The Unicode Standard, Version 5.0*, issued by the Unicode Consortium and published by Addison-Wesley. The material has been modified slightly for this electronic edition, however, 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 on more recent versions of the 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. *Dai Kan-Wa Jiten*, used as the source of reference Kanji codes, was written by Tetsuji Morohashi and published by Taishukan Shoten.

Cover and CD-ROM label design: Steve Mehallo, www.mehallo.com

The publisher offers excellent discounts on this book when ordered in quantity for bulk purchases or special sales, which may include electronic versions and/or custom covers and content particular to your business, training goals, marketing focus, and branding interests. For more information, please contact U.S. Corporate and Government Sales, (800) 382-3419, corpsales@pearsoned.com. For sales outside the United States please contact International Sales, international@pearsoned.com

Visit us on the Web: www.awprofessional.com

Library of Congress Cataloging-in-Publication Data

The Unicode Standard / the Unicode Consortium ; edited by Julie D. Allen ... [et al.]. — Version 5.0.
p. cm.

Includes bibliographical references and index.

ISBN 0-321-48091-0 (hardcover : alk. paper)

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

II. Unicode Consortium.

QA268.U545 2007

005.7'22—dc22

2006023526

Copyright © 1991–2007 Unicode, Inc.

All rights reserved. Printed in the United States of America. 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, write to Pearson Education, Inc., Rights and Contracts Department, 75 Arlington Street, Suite 300, Boston, MA 02116. Fax: (617) 848-7047

ISBN 0-321-48091-0

Text printed in the United States on recycled paper at Courier in Westford, Massachusetts.

First printing, October 2006

I.2 General Index

The General Index covers the contents of this book, excluding the annexes. To find topics in the Unicode Standard Annexes, Unicode Technical Standards, and Unicode Technical Reports, refer to their 5.0.0 versions on the CD-ROM accompanying this book, or use the search feature on the Unicode Web site for the latest versions.

For definitions of terms used in this book, see the *Glossary*. To find the code points for specific characters or the code ranges for particular scripts, see *Section I.1, Unicode Names Index*.

A

- abjads 198, 263
 - abstract character sequences
 - definition 79
 - abstract characters 25
 - definition 78
 - abugidas 199, 200, 295, 373
 - accent marks *see* diacritics
 - accented characters
 - encoding 12
 - Latin 226
 - normalization 160
 - accounting numbers, ideographic 140
 - acrophonic numerals 159, 242
 - Aegean numbers 479
 - Afrikaans 231
 - Ainu 434
 - Algonquian 464
 - Ali Gali 449
 - aliases
 - character name 78, 143, 565
 - property 131
 - property value 131
 - allocation areas 39
 - allocation of encoded characters 38–46, 1100
 - Alphabetic (informative property) 144
 - alphabets 198
 - European 225–258
 - mathematical 494–498
 - Alpine 474
 - alternate format characters (deprecated) 146, 543–545
 - Amharic 445
 - angle brackets (U+2329 and U+232A)
 - deprecated for technical publication 509
 - Annexes, Unicode Standard (UAX) xxxvi, 1084
 - as components of Unicode Standard 70
 - conformance 75
 - list of 75
 - annotation characters 552–554
 - use in plain text discouraged 553
 - ANSI/ISO C
 - wchar_t and Unicode 154
 - apostrophe (U+0027) 211
 - Arabic 269–283
 - Arabic-Indic digits 271–272
 - signs used with 274
 - ArabicShaping.txt 275, 279, 288
 - Aramaic 295, 341, 364, 449
 - archaic scripts 471–480
 - areas of the Unicode Standard 39
 - Armenian 247–249
 - arrows 506–507
 - ASCII
 - characters with multiple semantics 203
 - transparency of UTF-8 32
 - Unicode modeled on 1
 - zero extension 154, 1095
 - Asian Scripts Area 43
 - Assamese 312
 - assigned code points 11, 27
 - Athapascan 464
 - atomic character boundaries 168
- ### B
- Balinese 399–404
 - Bangla 312–317
 - base characters 252
 - definition 91
 - multiple 52
 - ordered before combining marks 171, 253
 - Basic Multilingual Plane (BMP) 2, 39
 - allocation areas 43
 - representation in UTF-16 32
 - Basque 231
 - benefits of Unicode 1
 - Bengali 312–317
 - Bidi Class (normative property) 138

- Bidi Mirrored (normative property)141
 - Bidi Mirroring Glyph (informative property)142
 - BidiMirroring.txt142
 - Bidirectional Algorithm, Unicode47, 74
 - bidirectional ordering19
 - controls146, 542
 - bidirectional text46, 74
 - Middle Eastern scripts263
 - nonspacing marks in174
 - punctuation in203
 - big-endian35
 - definition74
 - Bihari296
 - binary comparison and sort order
 - caution for UTF-1632
 - UTF differences180, 182
 - UTF-835
 - blocks of the Unicode Standard39, 197
 - Blocks.txt40, 1112
 - BMP *see* Basic Multilingual Plane
 - BNF (Backus-Naur Form)1079
 - BOCU-1 *see* UTN #6, BOCU-1
 - MIME-Compatible Unicode Compression
 - BOM (U+FEFF)35, 57, 105–108, 550–552
 - Bopomofo431–433
 - boundaries, text11, 54, 145, 168–169, 178
 - see also* UAX #14, Line Breaking Properties
 - see also* UAX #29, Text Boundaries
 - boustrophedon47, 476
 - Brahmi295, 341, 364, 373
 - Braille519–520
 - Breton231
 - Buginese397–398
 - Buhid395
 - Bulgarian245
 - bullets214
 - Burmese *see* Myanmar
 - Byelorussian245
 - byte order mark (BOM) (U+FEFF)35, 57, 105–108, 550–552
 - byte ordering
 - changing72
 - conformance74
 - byte serialization35, 57
 - Byzantine Musical Symbols525
- C**
- C language
 - wchar_t and Unicode154
 - C0 and C1 control codes27, 43, 144, 532
 - Cambodian *see* Khmer
 - camelcase187
 - Canadian Aboriginal Syllabics464
 - canonical composite characters
 - see* canonical decomposable characters
 - canonical decomposable characters
 - definition97
 - canonical decomposition55
 - definition96
 - canonical decomposition mappings95
 - canonical equivalence
 - definition97
 - nonspacing marks175
 - canonical equivalent character sequences
 - conformance71, 72
 - canonical mappings
 - see* canonical decomposition mappings
 - canonical ordering of combining marks115–117
 - canonical precomposed characters
 - see* canonical decomposable characters
 - Cantonese416
 - capital letters132, 184, 225
 - carriage return (U+000D) (CR)162, 533
 - carriage return and line feed (CRLF)162
 - case232
 - beyond ASCII185
 - camelcase187
 - case folding187
 - case operations (conformance)75, 123–126
 - case operations and normalization189
 - case operations, reversibility187
 - cased (definition)123
 - case-insensitive comparison . . .126, 180, 181, 187
 - casing context (definition)124
 - conversion125
 - detection125
 - European alphabets225
 - exceptional Latin pairs228, 231
 - Georgian249
 - lowercase132, 184, 225
 - mapping tables152
 - mappings123, 133, 184–186
 - mappings noted in code charts567
 - and text processes12
 - titlecase132, 184
 - Turkish I186, 228
 - uppercase132, 184, 225
 - Case (normative property)132, 184
 - CaseFolding.txt133, 188
 - caseless letters232
 - Catalan230
 - CD-ROMxxxvii
 - 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)

- character encoding forms (CEF) 28–35, 1095
 - see also* Unicode encoding forms
- character encoding model 28, 37
 - see also* UTR #17, Character Encoding Model
- character encoding schemes (CES) 35–38
 - see also* Unicode encoding schemes
- character encoding standards
 - coverage by Unicode 3
- character literals, Unicode
 - code point notation U+ 1080
- character mapping
 - interchange format *see* UTS #22, Character Mapping Markup Language (CharMapML)
- character names 77, 142–143, 1098
 - aliases 78, 143, 565
 - for CJK ideographs 142, 569
 - in code charts 563–565
 - for control codes 142, 144
 - conventions 1077
 - matching 143
- character properties
 - see* properties
 - see also individual properties, e.g.* combining classes
- character semantics 1, 71, 76–77, 1098
 - ASCII 203
 - definition 77
 - as Unicode design principle 16
- character sequences
 - abstract *see* abstract character sequences
 - canonical equivalent *see* canonical equivalent character sequences
 - compatibility equivalent *see* compatibility equivalent character sequences
 - conformance 71
 - named 142
- character shaping selectors (deprecated) 544
- character tabulation (U+0009) 533
- characters
 - abstract *see* abstract characters
 - arrangement in Unicode 41
 - assigned 11, 27
 - blocks 39, 197
 - boundaries 168
 - canonical decomposable *see* canonical decomposable characters
 - classes 1080
 - code charts 563–570
 - coded *see* encoded characters
 - combining *see* combining characters
 - compatibility decomposable *see* compatibility decomposable characters
 - composite *see* decomposable characters
 - concept of 14, 53
 - conformance definitions 78–80
 - confusable 191
 - conversion 151–153
 - 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 53
 - format control 27, 58, 204, 531–559
 - glyphs, relationship to 14
 - graphic 27
 - identity (definition) 77
 - interpretation 71
 - layout control 58, 534–543
 - modification 72
 - names list 563–565
 - names *see* character names
 - not encoded in Unicode 3
 - number encoded in Version 5.0 2, 1100
 - obsolete 80
 - online charts 1088
 - precomposed *see* decomposable characters
 - properties *see* properties
 - semantics *see* character semantics
 - special 57, 531–559
 - supplementary *see* supplementary characters
 - transcoding 151–153
 - unsupported 155–156
- characters, not glyphs
 - in spoofing 191
 - Unicode principle 14
- CharMapML
 - see* UTS #22, Character Mapping Markup Language (CharMapML)
- charsets
 - IANA registered names 36
- charts, character code *see* code charts
- Cherokee 463
- Chinese 415–416
 - Cantonese 416
 - Hakka 432
 - Mandarin 416
 - Minnan (Hokkien/Fujian, incl. Taiwanese) 432
 - simplified and traditional 415
- Chu hán 414
- Chu Nôm 1117
- citations for
 - properties 69
 - Unicode algorithms 69
 - Unicode Standard 68
- CJK ideographs 200, 408–425
 - accounting numbers 140
 - CJK Compatibility Ideographs 424–425

- CJK Compatibility Supplement425
- CJK Unified Ideographs 408–423
- CJK Unified Ideographs Extension A412
- CJK Unified Ideographs Extension B423–424
- code charts569
- compatibility ideographs in Plane 246
- component structure419
- encoding blocks411
- ideographic description sequences 427–430
- ideographic variation mark (U+303E)430
- KangXi radicals 425–426, 1023
- names142, 569
- numeric values140, 159
- order of encoding420
- radicals 425–426
- radical-stroke index1023
- source standards 409–411, 423
- unknown or unavailable220
- Vietnamese408
- CJK Miscellaneous Area43
- CJK punctuation and symbols219
 - compatibility forms220
 - overscores and underscores221
 - quotation marks210
 - sesame dots220
 - vertical forms220
- CJK Radical (property)428
- CJK-JRG (Chinese/Japanese/Korean Joint Research Group)1116
- CJKV Ideographs Area43
- CLDR (Common Locale Data Registry)1088
- cluster boundaries168
- code charts 563–570
 - online1088
 - representative glyphs564
- code point sequences
 - notation1078
- code points6, 25
 - assigned11, 27
 - assignment41, 1100
 - categories26
 - default ignorable156, 192
 - definition79
 - designated27
 - notation1077
 - number in Unicode Standard2
 - private-use *see* private-use code points
 - reserved *see* reserved code points
 - semantics28
 - surrogate *see* surrogates
 - unassigned *see* unassigned code points
 - undesignated27
- code positions *see* code points
- code set independence17
- code unit sequences
 - definition99
 - ill-formed (definition)100
 - notation1078
 - well-formed (definition)100
- code units
 - definition98
 - isolated98
- code values *see* code units
- coded character representations
 - see* coded character sequences
- coded character sequences
 - definition79
- coded characters *see* encoded characters
- codespace *see* Unicode codespace
- coeng382, 385
- Collation Algorithm, Unicode (UCA)12
- collation *see* sorting
- collation tables152
- combining character sequences49
 - canonical ordering115–117
 - defective173
 - definition93
 - Latin226
 - line breaking170
 - matching170
 - order of base character and marks171, 253
 - rendering170
 - selection168
 - truncation171–172
- combining characters 48–53, 109–115, 169–178
 - blocking reordering541
 - class zero135
 - combining marks253
 - definition91
 - dependence252
 - display order50
 - keyboard input170
 - ligatures52
 - multiple50
 - multiple base characters52
 - normalization of160
 - ordering conventions49
 - rendering of marks172–178
 - reordrant135
 - script-specific48
 - and sorting117
 - split136
 - strikethrough137
 - subjoined137
 - typographical interaction50, 133
 - vertical stacking50
 - see also* diacritics
- Combining Class (normative property)133

- combining classes 133, 175–176
 - class zero characters 115, 133
 - definition 114
 - combining grapheme joiner (U+034F) 540
 - combining half marks 146, 258
 - combining marks *see* combining characters
 - Common Locale Data Registry (CLDR) 1088
 - Compatibility and Specials Area 23, 43
 - compatibility characters 22
 - mapping 25
 - compatibility composite characters 24
 - compatibility decomposable characters 24
 - definition 95
 - compatibility decomposition 55
 - definition 95
 - compatibility decomposition mappings 95
 - Compatibility Encoding Scheme for UTF-16
 - see* UTR #26, Compatibility Encoding Scheme for UTF-16: 8-Bit (CESU-8)
 - compatibility equivalence
 - definition 96
 - compatibility equivalent character sequences
 - conformance 72
 - compatibility mappings
 - see* compatibility decomposition mappings
 - compatibility precomposed characters
 - see* compatibility decomposable characters
 - compatibility variants 23
 - composite characters
 - see* decomposable characters
 - compatibility *see* compatibility decomposable characters
 - Composition Exclusion (normative property) ... 86
 - compression 161
 - see also* UTS #6, A Standard Compression Scheme for Unicode (SCSU)
 - conferences 1088
 - conformance 65–126
 - definitions 76–80
 - examples 59
 - ISO/IEC 10646 implementations 1098
 - requirements 70–74
 - confusables 191
 - conjunct consonants
 - Indic 168, 300
 - Myanmar 380
 - selection of clusters 168
 - contextual shaping
 - apostrophe 212
 - Arabic 270
 - not used for Hebrew final forms 265
 - quotation marks 209
 - Syriac 288
 - contour tones 251
 - control codes 27, 58, 532
 - graphics for 508
 - names 144
 - properties 534
 - semantics 28, 533
 - specified in Unicode 533
 - control sequences 532
 - conversion of characters 151–153
 - convertibility
 - as Unicode design principle 23
 - Coptic 240, 243–245
 - corporate use subarea 547
 - corrigenda 67
 - CR (U+000D carriage return) 162, 533
 - CRLF (carriage return and line feed) 162
 - Croatian 231
 - digraphs 231
 - culturally expected sorting 12, 179
 - Cuneiform
 - Old Persian 483
 - Sumero-Akkadian 483–486
 - Ugaritic 482
 - currency symbols 490–492
 - encoded in script blocks 491
 - cursive joining 536–540
 - Arabic 275–281
 - control characters for ... 146, 270–271, 452, 535
 - Mongolian 451–452
 - N’Ko 461
 - Syriac 288–290
 - transparency 540
 - cursive scripts 263
 - Cypriot 479–480
 - see also* Linear B
 - Cyrillic 245–246
 - for Mongolian 449
 - Czech 231
- ## D
- danda, in Devanagari block 311
 - Danish 230
 - dashes 206
 - Database, Unicode Character
 - see* Unicode Character Database (UCD)
 - dead consonants, Indic 299
 - dead keys 170
 - decomposable characters 55
 - definition 95
 - normalization of 160
 - decomposition 55, 95–97
 - canonical *see* canonical decomposition
 - compatibility *see* compatibility decomposition
 - definition 95

- mapping, definition 95
 - mappings noted in code charts 567
 - in normalization 160
 - default grapheme clusters 168
 - see also* UAX #29, Text Boundaries
 - default ignorable code points 156, 192
 - Default Ignorable Code Points (property) 192
 - default property values 156
 - definition 84
 - defective combining character sequences 173
 - definition 93
 - dependent vowel signs
 - Indic 298
 - Khmer 387
 - Philippine scripts 396
 - deprecated characters 66
 - alternate format 146, 543–545
 - definition 80
 - Derived Age (property) 156
 - derived properties
 - definition 89
 - DerivedAge.txt 1109, 1111
 - DerivedCoreProperties.txt 124, 133, 192
 - DerivedNormalizationProps.txt 190
 - Deseret 465–467
 - design goals of Unicode 4
 - design principles of Unicode 13–23
 - designated code points 27
 - Devanagari 296–312
 - Dhivehi 291
 - diacritics 48, 253
 - alternative glyphs 226, 253
 - Czech 227
 - display in isolation 53, 205, 254
 - double 113, 146, 255
 - Greek 237–238, 241
 - Latin 226–229
 - Latvian 227
 - on i and j 228
 - mathematical 498
 - rendering 172–178
 - Romanian 228
 - Slovak 227
 - spacing clones of 251, 254
 - symbol 48, 257
 - Turkish 228
 - see also* combining characters
 - digit form names 272
 - digits 140, 158
 - Arabic-Indic 271–272
 - national shapes 544
 - digraphs 230, 231, 234, 235
 - dingbats 515–516
 - directionality 19, 46
 - East Asian scripts 408
 - Middle Eastern scripts 263
 - Mongolian 450
 - musical symbols 522
 - normative property 138
 - Ogham 472
 - Old Italic 474
 - Philippine scripts 396
 - Runic 476
 - discussion list for Unicode 1088
 - dotless i 186, 228
 - dotted circle
 - in code charts 92, 254
 - in fallback rendering 173
 - to indicate diacritic 48
 - to indicate vowel sign placement 50
 - double diacritics 113, 146, 255
 - Dutch 230
 - dynamic composition
 - as Unicode design principle 22
 - Dzongkha 343
- ## E
- East Asian scripts 407–441
 - writing direction 47
 - see also* CJK ideographs
 - Eastern Arabic-Indic digits 272
 - EBCDIC
 - newline function 163
 - see* UTR #16, UTF-EBCDIC
 - editing, text boundaries for 167–169
 - efficiency
 - as Unicode design principle 14
 - e-mail discussion list for Unicode 1088
 - Enclosed Alphanumerics 517
 - enclosing marks 258
 - definition 92
 - encoded characters 6, 25
 - allocation 38–46, 1100
 - definition 79
 - encoding form conversion
 - definition 104
 - encoding forms 28–35
 - ISO/IEC 10646 definitions 1095
 - encoding forms, Unicode
 - see* Unicode encoding forms
 - encoding model for Unicode characters 28, 37
 - see also* UTR #17, Character Encoding Model
 - encoding schemes 35–38
 - encoding schemes, Unicode
 - see* Unicode encoding schemes

- endian ordering
 see byte order mark (BOM) (U+FEFF)
- end-user subarea 547
- English 230
- equivalent sequences 160
 case-insensitivity 181, 187
 combining characters in matching 170
 conformance 72
 Hangul syllables 437
 language-specific 97
 security implications 190
 in sorting and searching 179
 as Unicode design principle 22
 see also canonical equivalence
 see also compatibility equivalence
 see also encoding forms, encoding schemes
- errata xxxviii, 67, 1089, 1111
- escape sequences 533
 not used in Unicode 1, 4
- Esperanto 231
- Estonian 231
- Ethiopic 445–448
- Etruscan 473
- euro sign (U+20AC) 492
- European alphabetic scripts 225–258
- eyelash-RA 305
- F**
- fallback rendering of nonspacing marks 173
- FAQ (Frequently Asked Questions) 1088
- Faroese 230
- Farsi 269, 270
- featural syllabaries 199
- FF (U+000C form feed) 162, 533
- file separator (U+001C) 533
- Finnish 230
- Finno-Ugric Transcription (FUT)
 see Uralic Phonetic Alphabet (UPA)
- fixed-width Unicode encoding form (UTF-32) ... 31, 102
- flat tables 152
- Flemish 230
- fonts
 for mathematical alphabets 497–498
 style variation for symbols 489
 and Unicode characters 15
- form feed (U+000C) (FF) 162, 533
- format control characters 27, 58, 204, 531–559
 deprecated 543–545
 prefixed 146
 reserved ranges 156
 stateful 194
- fraction characters 499
- fraction slash (U+2044) 212, 499
- French 231
- Frisian 231
- FTP site, Unicode Consortium 1088
- fullwidth forms in East Asian encodings ... 434–435
- futhark 476
- G**
- Garshuni 283
- Ge'ez 445
- General Category (normative property) 138
 list of values 138
- general punctuation 202–221
- General Scripts Area 43
- geometrical symbols 512–513
- Georgian 249–250
- German 230
- geta mark (U+3013) 220
- Glagolitic 246–247
- glyph selection tables 152
- glyphs 6, 15
 characters, relationship to 14
 diacritics alternative 226, 253
 Greek alternative 238–240
 Latin alternative 226
 mathematical alternative 503
 missing 193
 representative in code charts 564
 standardized variants 545
 symbols alternative 489
- golden numbers 477
- Gothic 477–478
- grapheme base 252
 definition 93
- grapheme clusters 11, 53–54
 see also UAX #29, Text Boundaries
 default 168
 definition 94
- grapheme extender
 definition 93
- grapheme joiner, combining (U+034F) 540
- graphic characters 27
- Greek 237–242
 acrophonic numerals 159, 242
 alternative glyphs 238–240
 ancient musical notation 526–528
 letters as symbols 238–240, 504
 see also Cypriot, Linear B
- Greenlandic 231
- group separator (U+001D) 533
- guillemets 209
- Gujarati 321–322
- Gurmukhi 317–320

H

Hakka 432
 halant 295
 see also virama
 half-consonants, Indic 301
 half marks, combining 146, 258
 halfwidth forms in East Asian encodings ... 434–435
 Han ideographs *see* CJK ideographs
 Han unification 417–423
 history 1115–1117
 and language tags 166
 language usage 414
 source separation rule 412, 418
 source standards 409–411, 423
 Hangul syllables 407, 435–438
 boundary determination 119
 canonical decomposition 122
 collation 436
 composition 121
 conjoining jamo 117–123
 equivalent sequences 437
 as grapheme clusters 54
 Hangul Compatibility Jamo 436–437
 Hangul Jamo 435–436
 Hangul Syllables block 437–438
 Johab set 437
 name generation 123
 normalization 436
 precomposed 119
 standard 120
 Hangzhou numerals 500
 Hanja *see* CJK ideographs
 Hanunóo 395
 Hani *see* CJK ideographs
 harakat, Arabic pronunciation marks 269
 hasant 312
 hash tables 152
 Hebrew 264–269
 high-surrogate
 definition 97
 high-surrogate code points 70, 548
 high-surrogate code units 97
 higher-level protocols
 definition 80
 Hindi 296
 Hiragana 433
 historic scripts 471–480
 horizontal tab (U+0009) 533
 HTML newline function 163
 Hungarian 231
 hyphenation 536
 as a text process 10
 hyphens 206, 536

I

I Ching symbols 516
 IANA charset names 36
 Icelandic 230
 identifiers 179
 see also UAX #31, Identifier and Pattern Syntax
 Ideographic (informative property) 145
 Ideographic Rapporteur Group (IRG) ... 409, 1116
 Ideographic Variation Database *see* UTS #37, Unicode
 Ideographic Variation Database
 ideographs *see* CJK ideographs
 IICore 412, 1117
 ill-formed
 definition 100
 implementation guidelines 151–194
 in a Unicode encoding form
 definition 101
 in-band mechanisms 559
 Indic scripts 295–337, 341–343
 principles, in terms of Devanagari 297–304
 relation to ISCII standard 297
 Indonesian 230
 industry character sets
 covered in Unicode 3
 information separators (U+001C..U+001F) 533
 informative properties
 definition 86
 inside-out rule 172
 interchange restrictions 28
 International Phonetic Alphabet (IPA) .. 198, 233–234
 Spacing Modifier Letters 250
 see also phonetic alphabets
 internationalization 17
 Internationalization & Unicode Conference (IUC) .. 1088
 Internet protocols
 UTF-8 as preferred encoding 33
 Inuktitut 464
 invisible operators 507
 iota subscript 238
 IPA *see* International Phonetic Alphabet
 IRG (Ideographic Rapporteur Group) ... 409, 1116
 Irish 230, 472
 ISCII standard and Unicode 297
 ISO/IEC 10646 1091–1098
 codespace 1095
 conformance of Unicode implementations . 1098
 encoding forms 1095
 synchrony with Unicode Standard 1097
 timeline compared to Unicode versions ... 1092
 UCS transformation formats (UTF) 1096
 Italian 230
 ITC Zapf Dingbats 515
 IUC *see* Internationalization & Unicode Conference

J

Jamo Short Name (normative property)	86
Jamo.txt	123
jamos <i>see</i> Hangul syllables	
Japanese	407
Jawi	281
Johab	437
joiners	271
combining grapheme joiner (U+034F)	540
word joiner (U+2060)	534
zero width joiner (U+200D)	270–271, 537
justification	175

K

Kana (Hiragana and Katakana)	433–434
Kanbun	425
KangXi radicals	425–426, 1023
Kanji <i>see</i> CJK ideographs	
Kannada	331–334
Katakana	433–434
Kawi	399, 401
KC (Normalization Form) <i>see</i> Normalization Form KC	
KD (Normalization Form) <i>see</i> Normalization Form KD	
keytop labels	509
Kharoshthi	364–366
Khmer	382–392
characters not recommended	389
syllable components, order of	390
killer Myanmar	380
<i>see also</i> virama	
Korean Hangul <i>see</i> Hangul	
Kurdish	269

L

Ladino	264
language tags	166, 555–558
and Han unification	166
use strongly discouraged	558
Lanna	394
Lao	376–378
last-resort glyphs	193
Latin	226–236
alternative glyphs	226
Basic Latin	230
encoding blocks	40
IPA Extensions	233–234
Latin Extended Additional	235–236
Latin Extended-A	230
Latin Extended-B	231–233
Latin Extended-C	236

Latin Extended-D	236
Latin Ligatures	236
Latin-1 Supplement	230
Phonetic Extensions	234–235
Latvian	231
layout control characters	58, 534–543
leading surrogates <i>see</i> high-surrogate code units	
legibility criterion for plain text	19
letter spacing	535
letterlike symbols	492–498
LF (U+000A line feed)	162, 533
liaison members, Unicode Consortium	1083
ligatures	536–540
Arabic	278–279
combining characters on	52
control characters for	146
Latin	236
for nonspacing marks	176
selection	169
Syriac	290
Limbu	360–363
line breaking	161–165, 534–536
control characters	148
recommendations	164
in South Asian scripts	376, 381, 392
<i>see also</i> UAX #14, Line Breaking Properties	
line feed (U+000A) (LF)	162, 533
line separator (U+2028) (LS)	162, 536
line tabulation (U+000B) (VT)	533
Linear B	478–479
<i>see also</i> Cypriot	
linear boundaries	168
Lithuanian	231
little-endian	35
definition	74
Locale Data Markup Language <i>see</i> UTS #35, Locale Data Markup Language (LDML)	
logical order exceptions to	135
as Unicode design principle	19
logosyllabaries	200
low-surrogate definition	97
low-surrogate code points	70, 548
low-surrogate code units	97
lowercase	132, 184, 225
LS (U+2028 line separator)	162, 536

M

MacOS newline function	163
mail discussion list for Unicode	1088

- major version 67
 - Malay 230
 - Malayalam 334–337
 - Maltese 231
 - Manchu 449
 - Mandarin 416
 - Manden 458
 - mapping tables *see* tables of character data
 - Marathi 296, 305, 311
 - markup languages
 - line breaking 162
 - and Unicode conformance 559
 - see also* UTR #20, Unicode in XML and Other Markup Languages
 - Mathematical (informative property) 502
 - mathematical expression format characters 146
 - see also* UTR #25, Unicode Support for Mathematics
 - mathematical symbols 502–507
 - alphabets 494–498
 - alphanumeric 494–498
 - fonts 497–498
 - format characters 507
 - fragments for typesetting 510
 - invisible operators 507
 - operators 503–504
 - standardized variants 507
 - MathML 505
 - matras 134, 298
 - Middle Eastern scripts 263–292
 - Min 416
 - Minnan (Hokkien/Fujian, incl. Taiwanese) 432
 - minor version 67
 - minus sign 504
 - commercial (U+2052) 215
 - mirrored property
 - see* Bidi Mirrored (normative property)
 - mirroring of paired punctuation 208
 - Miscellaneous Symbols 514
 - missing glyphs 193
 - modifier letters 236, 250–252
 - Modifier Letters, Spacing 235
 - Mongolian 354, 448–457
 - writing direction 450
 - multibyte encodings
 - compared to UTF-8 33
 - multistage tables 152
 - musical symbols 520–528
 - ancient Greek 526–528
 - Balinese 403
 - Byzantine 525
 - directionality 522
 - Gregorian 521
 - Western 520–525
 - Myanmar 379–381
- ## N
- N’Ko 458–463
 - named character sequences 142
 - names, character *see* character names
 - namespace 78
 - NEL (U+0085 next line) 162, 533
 - Nepali 296
 - neutral directional characters 138
 - New Tai Lue 394–395
 - newline function (NLF) 163, 534
 - newline guidelines 161–165
 - next line (U+0085) (NEL) 162, 533
 - NFC (Normalization Form C) 55
 - NFD (Normalization Form D) 55
 - NFKC (Normalization Form KC) 55
 - NFKD (Normalization Form KD) 55
 - NLF (newline function) 163, 534
 - no-break space (U+00A0) 534
 - base for diacritic in isolation 53, 205, 254
 - no-break space, narrow (U+202F) 454
 - noncharacter code points *see* noncharacters
 - noncharacters 27, 57, 549
 - in code charts 568
 - conformance 70
 - definition 80
 - deletion 72
 - handling 72
 - interchange restrictions 28
 - semantics 28
 - U+10FFFF (not a character code) 549
 - U+FDD0..U+FDEF 27, 549
 - U+FFFE (not a character code) 58, 549
 - U+FFFF (not a character code) 27, 549
 - nondecomposable characters 56
 - non-joiner, zero width (U+200C) 270–271, 538
 - nonlinear boundaries 169
 - non-overlap principle in Unicode encoding forms 29
 - nonspacing marks 252
 - definition 92
 - display in isolation 53, 205, 254
 - positioning 176
 - rendering 172–178
 - see also* combining characters
 - see also* diacritics
 - normalization 55, 160–161
 - and case operations 189
 - conformance 74
 - of private-use characters 546
 - see also* UAX #15, Unicode Normalization Forms
 - Normalization Form C (NFC) 55
 - Normalization Form D (NFD) 55

- Normalization Form KC (NFKC) 55
 - Normalization Form KD (NFKD) 55
 - normative behaviors
 - definition 76
 - normative properties
 - definition 85
 - list 85
 - may change 85
 - Norwegian 230
 - notational conventions 1077–1081
 - notational systems 201
 - nukta 306
 - null (U+0000)
 - as Unicode string terminator 534
 - number forms 498–500
 - CJK ideographs 159
 - numbers
 - handling 158
 - ideographic accounting 140
 - numerals
 - acrophonic 242
 - Chinese counting rods 499
 - Coptic 244
 - Cuneiform 486
 - Greek acrophonic 159
 - Hangzhou 500
 - old-style 213
 - Roman 159, 499
 - Suzhou-style 500
 - numeric separators 215
 - numeric shape selectors (deprecated) 544
 - Numeric Value (normative property) 139
 - numero sign (U+2116) 492
- O**
- object replacement character (U+FFFC) 554
 - obsolete characters 80
 - octet 1079
 - Ogham 472–473
 - Old Italic 473–475
 - Old Persian 483
 - old-style numerals 213
 - Oriya 322–324
 - Oromo 445
 - Osmanya 457
 - out-of-band mechanisms 559
 - overlapping encodings 29
 - overscores 212
- P**
- Panjabi 317
 - paragraph or section marks 214
 - paragraph separator (U+2029) (PS) 162, 536
 - Pashto 269
 - Persian 269, 270
 - Phags-pa 353–359
 - Philippine scripts 395–397
 - Phoenician 480
 - phonemes 201
 - phonetic alphabets 198
 - IPA Extensions 233–234
 - Phonetic Extensions 234–235
 - Spacing Modifier Letters 250–252
 - Uralic Phonetic Alphabet (UPA) 215, 234
 - see also* International Phonetic Alphabet (IPA)
 - phonetic extensions 236
 - Pinyin 229
 - pivot code, Unicode as 152
 - plain text
 - legibility criterion 19
 - as Unicode design principle 18
 - planes of Unicode codespace 39
 - Plane 0 (BMP) 39
 - Plane 1 (SMP) 39, 45
 - Plane 14 (SSP) 39
 - Plane 2 (SIP) 39, 46
 - Planes 15–16 (Private Use) 46, 548
 - points, Hebrew pronunciation marks 264
 - policies of the Unicode Consortium 1089
 - Polish 231
 - Portuguese 230
 - precomposed characters
 - compatibility *see* compatibility decomposable characters
 - see* decomposable characters
 - prefixed format control characters 146
 - Private Use Area (PUA) 43, 547
 - Private Use planes 39, 46, 548
 - private-use characters
 - semantics 28
 - private-use code points 27, 155
 - conformance 71
 - definition 91
 - high-surrogates 548
 - processing code, choice of Unicode encoding form . . 33
 - properties 16, 81–91, 129–148
 - aliases 131
 - aliases (definition) 90
 - of control codes 534
 - data tables 152
 - derived *see* derived properties
 - informative *see* informative properties
 - normative references to 69, 75
 - normative *see* normative properties
 - provisional *see* provisional properties
 - simple *see* simple properties
 - and Unicode algorithms 85

- in Unicode Character Database (UCD) 40
 - see also individual properties, e.g. combining classes*
 - property values
 - aliases 131
 - aliases (definition) 90
 - default 156, 546
 - default (definition) 84
 - normative references to 75
 - PropertyAliases.txt 90, 1080, 1112
 - PropertyValueAliases.txt 90, 1080, 1112
 - PropList.txt 133
 - Provençal 231
 - provisional properties
 - definition 87
 - PS (U+2029 paragraph separator) 162, 536
 - PUA (Private Use Area) 43, 547
 - pulli* 325
 - punctuation 202–221
 - in bidirectional text 203
 - blocks containing 197
 - CJK 219
 - doubled 212
 - paired 208
 - small form variants 221
 - typographic forms 202
 - vertical forms 220
 - Punjabi 317
- Q**
- quotation marks 209–211
 - East Asian 210
 - European 209
- R**
- radicals, KangXi and other CJK 425–426
 - radical-stroke index 1023
 - record separator (U+001E) 533
 - recycling symbols 514
 - referencing 75
 - properties 69
 - Unicode algorithms 69
 - Unicode Standard 68
 - regular expressions 166
 - and line breaking 162
 - see also* UTS #18, Unicode Regular Expression Guidelines
 - rendering of text 6, 10, 16
 - unsupported characters 155
 - repertoire of abstract characters 25
 - replacement character (U+FFFD) 38, 58, 73, 554
 - reserved code points 26, 155
 - definition 80
 - in code charts 568
 - preservation in interchange 28
 - see also* unassigned code points
 - Rhaeto-Romanic 231
 - rich text 18
 - right single quotation mark (U+2019)
 - preferred for apostrophe 211
 - right-to-left text 46
 - East Asian scripts 408
 - Middle Eastern scripts 263
 - roadmap for script additions 40
 - Roman numerals 159, 499
 - Romanian 231
 - Romany 231
 - Runic 475–477
 - Russian 245
- S**
- Sami 231
 - sample code, on CD-ROM xxxvii
 - Sanskrit 296
 - scalar values, Unicode
 - see* Unicode scalar values
 - scripts
 - added in Version 5.0 3
 - roadmap for future additions 40
 - types of 201
 - in Unicode Standard 2
 - see also* UAX #24, Script Names
 - SCSU
 - see* UTS #6, A Standard Compression Scheme for Unicode
 - searching 179–181
 - case-insensitive 181, 187
 - as a text process 10
 - section or paragraph marks 214
 - security issues 190
 - self-synchronization of encoding forms 30
 - semantics
 - see* character semantics
 - sequences
 - notation 1078
 - Serbian
 - corresponding digraphs in Croatian 231
 - Shan 393
 - Shavian 467
 - Show Hidden 72, 173, 193, 545
 - SHY (U+00AD soft hyphen) 536
 - Sibe 450
 - signature for Unicode data 58, 550–552
 - simple properties
 - definition 89
 - simplified Chinese 415

- Sindhi 269, 312
- Sinhala 341–343
- SIP (Supplementary Ideographic Plane) 39, 46
- slash, fraction (U+2044) 212
- Slovak 231
- Slovenian 231
- small letters 132, 184, 225
- SMP (Supplementary Multilingual Plane) 39, 45
- soft hyphen (U+00AD) (SHY) 536
- Somali 457
- Sorbian 231
- sorting 12, 179
 - case-insensitive 180
 - and combining characters 117
 - and combining grapheme joiner 541
 - culturally expected 12, 179
 - language-insensitive 180
 - as a text process 10
 - see also* Unicode Collation Algorithm (UCA)
- source separation rule 412, 418
- South Asian scripts 295–337, 341–363
- Southeast Asian scripts 373–397
- space (U+0020)
 - base for diacritic in isolation 53, 205, 254
- space characters 205, 534–536
 - graphics for 508
- space, zero width (U+200B) 206
- spacing clones of diacritics 251, 254
- spacing marks 252
 - definition 93
- Spacing Modifier Letters 250–252
- Spanish 230
- special characters 57, 531–559
- SpecialCasing.txt 123, 133
- Specials 550–554
- spell-checking
 - as a text process 10
- spellings, alternative
 - see* equivalent sequences
- spoofing 191
- SSP (Supplementary Special-purpose Plane) 39
- stability 88, 131, 1119–1124
 - as Unicode design principle 22
- stacked boundaries 168
- stacking sequences 50
 - nondefault 51
- Standard Compression Scheme for Unicode (SCSU)
 - see* UTS #6, A Standard Compression Scheme for Unicode
- standard Korean syllables 120
- standardized variants 453, 545
 - mathematical symbols 507
- StandardizedVariants.txt 453, 507, 545
- standards coverage 3
- stateful encoding
 - not used in Unicode 4
 - paired format controls 194
- string comparison 12
- string literals, Unicode
 - code point notation `\u1234` 1080
- strings, Unicode 37, 100
 - null termination 534
- strong directional characters 138
- styled text 18
- sublinear searching 181
- subsets, supported 61
 - conformance 71
 - ISO/IEC 10646 specification for 1097
- substitution character 58
- Sumero-Akkadian 483–486
- superscripts and subscripts 501
- supplementary characters
 - tables for 153
 - in UTF-16 strings 38
- Supplementary General Scripts Area 43
- Supplementary Ideographic Plane (SIP) 39, 46
- Supplementary Multilingual Plane (SMP) 39, 45
- supplementary planes
 - representation in UTF-8 33
 - representation in UTF-16 31
- Supplementary Private Use Areas 46, 548
- Supplementary Special-purpose Plane (SSP) 39
- supported subsets 61
 - conformance 71
- surrogate code points
 - see* surrogates
- surrogate pairs 31, 102
 - definition 97
 - processing 34, 157–158
- surrogates 27, 97–98, 548
 - interchange restrictions 28
 - isolated surrogates, handling 38
 - isolated surrogates, ill-formed 102
 - isolated surrogates, uninterpreted 98
 - support levels 157
- Surrogates Area 43, 548
- Suzhou-style numerals 500
- Swahili 230
- Swedish 230
- syllabaries 199
 - alphabetic property 144
 - featural 199
- Syloti Nagri 363–364
- symbols 489–528
 - appearance variation 489
 - arrows 506–507
 - currency 490–492
 - dingbats 515–516

Enclosed Alphanumerics 517
 fragments for mathematical typesetting 510
 geometrical 512–513
 Khmer lunar calendar 392
 letterlike 492–498
 mathematical 502–507
 mathematical alphanumeric 494–498
 miscellaneous 514
 musical 520–528
 number forms 498–500
 recycling 514
 technical 508–512
 Symbols Area 43
 symmetric swapping format characters
 (deprecated) 543
 Syriac 283–290

T

tab (U+0009 character tabulation) 533
 tab, vertical (U+000B) 162, 533
 tables of character data 152–153
 optimization 152
 supplementary characters 153
 tag characters 554–559
 use strongly discouraged 554
 Tagalog 395
 Tagbanwa 395
 tags, language 166, 555–558
 use strongly discouraged 558
 Tai Le 393–394
 Tai Xuan Jing symbols 517
 Tamil 324–330
 TCHAR in Win32 API 155
 Technical Notes (UTN) 1087
 Technical Reports (UTR) 1084
 abstracts 1086
 Technical Standards (UTS) xxxvii, 1084
 abstracts 1085
 technical symbols 508–512
 Telugu 330–331
 terminal emulation 490
 text boundaries 11, 54, 145, 168–169, 178
 see also UAX #14, Line Breaking Properties
 see also UAX #29, Text Boundaries
 text elements 6, 10, 167
 boundaries 178
 for sorting 179
 variable-width nature 34
 text processes 5, 10–13
 text rendering 6, 10, 16
 text selection, boundaries for 167–169
 Thaana 291–292
 Thai 373–376

Tibetan 343–353
 Tifinagh 457
 Tigre 445
 tilde (U+007E) 215
 titlecase 132, 184
 Todo 449
 tone letters 251–252
 tone marks
 Bopomofo spacing 431, 432
 Chinantec 252
 Chinese 252
 Tai Le 393
 Thai 373
 Vietnamese 228
 traditional Chinese 415
 trailing surrogates
 see low-surrogate code units
 transcoding 151–153
 tables 152
 triangulation in transcoding 152
 truncation
 combining character sequences 171–172
 surrogates and 158
 Turkish 231
 case mapping of I 186, 228
 two-stage tables 152

U

U+ notation 1080
 U+10FFFF (not a character code) 549
 U+FEFF (BOM) 550–552
 U+FFFE (not a character code) 549
 U+FFFF (not a character code) 549
 UAX (Unicode Standard Annex) xxxvi, 1084
 as component of Unicode Standard 70
 conformance 75
 list of 75
 UCA *see* Unicode Collation Algorithm
 UCD *see* Unicode Character Database
 UCS (Universal Character Set)
 see ISO/IEC 10646
 UCS-2 1095
 UCS-4 1095
 Ugaritic 482–483
 Uighur 354, 449
 Ukrainian 245
 unassigned code points 27, 70, 155
 defined as reserved code points 80
 handling 66
 properties of 156
 semantics 71
 see also reserved code points
 underscores 212

- undesignated code points 27
- Unicode 1.0 Name (informative property) 144
- Unicode algorithms
 - conformance 75
 - definition 80
 - normative references to 69, 75
 - and properties 85
- Unicode Bidirectional Algorithm 20, 47
see also UAX #9, Bidirectional Algorithm
- Unicode Character Database (UCD) ... xxxvii, 130, 1089
 - changes 66
 - as component of Unicode Standard 70
 - properties in 40
- Unicode character encoding model 28, 37
see also UTR #17, Character Encoding Model
- Unicode character literals
 - code point notation U+ 1080
- Unicode codespace
 - allocation numbers 1100
 - definition 79
 - planes 39
 - same as ISO/IEC 10646 1095
 - size 2, 25
- Unicode Collation Algorithm (UCA) 12
see also UTS #10, Unicode Collation Algorithm
- Unicode conferences 1088
- Unicode Consortium 1083
 - addresses 1089
 - Consortium membership in standards bodies ..
1083
 - e-mail discussion list 1088
 - FTP site 1088
 - liaison members 1083
 - membership 1083
 - policies 1089
 - Web site xxxvii, 1088
- Unicode data signature 58, 550–552
- Unicode data types 153–155
 - for C 154–155
- Unicode encoding forms 98–105
 - advantages of each 33
 - conformance 30, 73
 - definition 99
 - fixed-width (UTF-32) 31, 102
 - signatures 551, 552
 - variable-width (UTF-16) 32, 102
 - variable-width (UTF-8) 32, 103
 - see also* encoding forms
- Unicode encoding schemes
 - conformance 105–109
 - definition 105
 - endian ordering 35
 - see also* encoding schemes
- Unicode escape sequence notation \u1234 1080
- Unicode Regular Expression Guidelines *see* UTS #18,
Unicode Regular Expression Guidelines
- Unicode scalar values
 - definition 98
- Unicode Security Mechanisms *see* UTS #39, Unicode
Security Mechanisms
- Unicode Standard
 - allocation of encoded characters 38–46
 - application areas 4
 - architecture 9–13
 - areas 39
 - benefits 1
 - blocks 39, 197
 - code charts 563–570
 - components 70
 - conformance 65–126
 - conformance of ISO/IEC 10646 implementations ..
1098
 - corrections 67
 - definitions for conformance 76–80
 - design goals 4
 - design principles 13–23
 - errata 67, 1089, 1111
 - normative references to 68, 75
 - number of characters 2, 1100
 - number of code points 2, 25
 - online code charts 1088
 - script coverage 2
 - security issues 190
 - synchrony with ISO/IEC 10646 1097
 - updates 1089
 - user community 4
 - versions *see* versions of the Unicode Standard
see also Version 5.0
- Unicode Standard Annexes (UAX) xxxvi, 1084
 - as components of Unicode Standard 70
 - conformance 75
 - list of 75
- Unicode string literals
 - code point notation \u1234 1080
- Unicode strings 37
 - definition 100
- Unicode Technical Committee (UTC) 1084
- Unicode Technical Notes (UTN) 1087
- Unicode Technical Reports (UTR) 1084
 - abstracts 1086
- Unicode Technical Standards (UTS) ... xxxvii, 1084
 - abstracts 1085
- UnicodeData.txt 123, 133
- unification
 - as Unicode design principle 21
 - see also* Han unification
- Unified CJK Ideograph (property) 428

- Unified Repertoire and Ordering (URO) . . .418, 1116
 see also Han unification
 - Unihan Database 131, 422, 569, 1089, 1117
 - Unihan.txt87, 131
 - unit separator (U+001F)533
 - Universal Character Set (UCS)
 see ISO/IEC 10646
 - universality
 as Unicode design principle14
 - Unix
 newline function163
 UTF-8 in18
 UTF-32 in31
 and UTFs34
 - unsupported characters 155–156
 - update version67
 - uppercase 132, 184, 225
 - Uralic Phonetic Alphabet (UPA)215, 234
 - Urdu269
 - URO (Unified Repertoire and Ordering) . . .418, 1116
 see also Han unification
 - UTF, Unicode Transformation Formats29, 99
 advantages of each33
 in APIs155
 binary comparison and sort order differences . . .
 180, 182
 in ISO/IEC 106461096
 - UTF-832, 103
 ASCII transparency32
 binary comparison and sort order35
 bit distribution (table)103
 BOM in 105, 109, 551
 byte ranges104
 compared to multibyte encodings33
 encoding form (definition)103
 encoding scheme36
 encoding scheme (definition)105
 in ISO/IEC 106461096
 non-shortest form is invalid103, 191
 preferred encoding for Internet protocols33
 security and191
 signature 105, 109, 551
 in Unix18
 in UTF-16 order182
 - UTF-1631, 102
 binary comparison and sort order caution . . .32
 bit distribution (table)102
 BOM in106, 550
 encoding form (definition)102
 encoding scheme (definition)106
 encoding schemes36
 in ISO/IEC 106461096
 surrogates and string handling37, 157
 in UTF-8 order183
 - UTF-16BE (big-endian)551
 encoding scheme36
 encoding scheme (definition)105
 - UTF-16LE (little-endian)551
 encoding scheme36
 encoding scheme (definition)106
 - UTF-3231, 102
 BOM in107
 encoding form (definition)102
 encoding scheme (definition)107
 encoding schemes36
 in Unix31
 - UTF-32BE (big-endian)
 encoding scheme36
 encoding scheme (definition)107
 - UTF-32LE (little-endian)
 encoding scheme36
 encoding scheme (definition)107
 - UTF-EBCDIC
 see UTR #16, UTF-EBCDIC
 - UTN (Unicode Technical Note)1087
 - UTR (Unicode Technical Report)1084
 abstracts1086
 - UTS (Unicode Technical Standard)xxxvii, 1084
 abstracts1085
- V**
- valid (synonym for well-formed)101
 - variable-width Unicode encoding form (UTF-8) . . .32,
 103
 - variable-width Unicode encoding form (UTF-16) . . .32,
 102
 - variants
 compatibility23
 fullwidth and halfwidth221
 mathematical symbols507
 small form221
 standardized545
 - variation selectors147, 545
 ideographic variation mark (U+303E)430
 Mongolian free variation selectors452
 - variation sequences545
 for Phags-pa358–359
 - Version 5.070
 additions3
 correlation with ISO/IEC 106461094
 number of characters2, 1100
 - versions of the Unicode Standard . . .xxxvii, 65, 1089,
 1100–1101
 backward compatibility66
 compared to ISO/IEC 10646 editions1100
 content67
 interaction in implementations156

numbering	66
property changes	66
stability	66
updates	1089
vertical tab (U+000B)	162, 533
vertical text	47, 203, 221
East Asian scripts	408
Mongolian	450
Vietnamese	228, 236
ideographs	408
virama	200, 295
definition	299
Kharoshthi	369
Khmer	385
Myanmar	380
Philippine scripts	396
virama-like characters	147
visual order used for Thai and Lao	20
vowel harmony	
Mongolian	454
vowel marks, Middle Eastern scripts	263
vowel separator	
Mongolian	455
vowel signs	
Indic	50, 298
Khmer	387
Philippine scripts	396
W	
wchar_t	
in C language	154
and Unicode encoding forms	33
weak directional characters	138
Web site, Unicode Consortium	xxxvii, 1088
Weierstrass elliptic function symbol	494
well-formed	
definition	100
Welsh	231
Where Is My Character?	1089
wide characters	
data type in C	154
wiggly fence (U+29DB)	506
Windows newline function	163
word breaks	170, 534–536
in South Asian scripts	376, 381, 392
word joiner (U+2060)	534
writing direction <i>see</i> directionality	
writing systems	198–201
Wu (Shanghainese)	416
X	
Xibe	450
Xishuang Banna Dai	394

XML	
<i>see</i> UTR #20, Unicode in XML and Other Markup Languages	
Y	
yen currency sign	492
Yi	438–441
Yiddish	264
Yijing Hexagram Symbols	516
ypogegrammeni	238
yuan currency sign	492
Z	
Zapf Dingbats	515
zero extension relation among encodings	1095
zero width joiner (U+200D)	270–271, 537
zero width no-break space (U+FEFF)	58, 74, 535
initial	108, 551
zero width non-joiner (U+200C)	270–271, 538
zero width space (U+200B)	535
for word breaks in South Asian scripts	376, 381, 392
zero-width space characters	535
ZWJ <i>see</i> zero width joiner (U+200D)	
ZWNBSP <i>see</i> zero width no-break space (U+FEFF)	
ZWNJ <i>see</i> zero width non-joiner (U+200C)	
ZWSP <i>see</i> zero width space (U+200B)	