

The Unicode Standard

Version 6.0 – Core Specification

To learn about the latest version of the Unicode Standard, see <http://www.unicode.org/versions/latest/>.

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

Unicode and the Unicode Logo are registered trademarks of Unicode, Inc., in the United States and other countries.

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

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

Copyright © 1991–2011 Unicode, Inc.

All rights reserved. This publication is protected by copyright, and permission must be obtained from the publisher prior to any prohibited reproduction. For information regarding permissions, inquire at <http://www.unicode.org/reporting.html>. For information about the Unicode terms of use, please see <http://www.unicode.org/copyright.html>.

The Unicode Standard / the Unicode Consortium ; edited by Julie D. Allen ... [et al.]. — Version 6.0.

Includes bibliographical references and index.

ISBN 978-1-936213-01-6 (<http://www.unicode.org/versions/Unicode6.0.0/>)

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

QA268.U545 2011

ISBN 978-1-936213-01-6

Published in Mountain View, CA

February 2011

I General Index

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

For definitions of terms used, 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 220
 - abjads 180, 237
 - abstract character sequences
 - definition 67
 - abstract characters 21
 - definition 66
 - abugidas 181, 182, 267, 349
 - accent marks *see* diacritics
 - accented characters
 - encoding 9
 - Latin 204
 - normalization 146
 - accounting numbers, ideographic 129
 - acrophonic numerals 146, 218
 - Aegean numbers 459
 - Afrikaans 208
 - Ainu 415
 - Aiton 359
 - Alchemical Symbols 505
 - Algonquian 443
 - Ali Gali 427
 - aliases
 - character name 66, 131, 550
 - property 119
 - property value 119
 - allocation areas 34
 - allocation of encoded characters 33–40, 577
 - Alphabetic (informative property) 135
 - alphabets 180
 - European 203–235
 - mathematical 482–485
 - Alpine 454
 - alternate format characters (deprecated) . . 136, 531–532
 - Amharic 424
 - Ancient Symbols 507
 - angle brackets (U+2329 and U+232A)
 - deprecated for technical publication 496
 - Annexes, Unicode Standard (UAX) xxxv, 562
 - as components of Unicode Standard 59
 - conformance 64
 - list of 64
 - annotation characters 539–540
 - use in plain text discouraged 540
 - ANSI/ISO C
 - wchar_t and Unicode 142
 - apostrophe 191
 - apostrophe (U+0027) 191
 - Arabic 242–256
 - Arabic-Indic digits 244–245
 - signs used with 246
 - ArabicShaping.txt 247, 250, 260
 - Aramaic 267, 314, 340, 427, 464
 - archaic scripts 451–460
 - areas of the Unicode Standard 34
 - ARIB 501
 - Armenian 224–225
 - arrows 493
 - ASCII
 - characters with multiple semantics 184
 - transparency of UTF-8 27
 - Unicode modeled on 1
 - zero extension 142, 573
 - Assamese 285
 - assigned code points 8, 23
 - Athapascan 443
 - atomic character boundaries 154
 - Avestan 468
- B**
- Balinese 380–385
 - Bamum 441–442
 - Bangla 285–289
 - base characters 230
 - definition 78
 - multiple 45
 - ordered before combining marks 155, 230
 - Basic Multilingual Plane (BMP) 1, 33
 - allocation areas 37
 - representation in UTF-16 27
 - Basque 208
 - Batak 388–389
 - benefits of Unicode 1
 - Bengali 285–289
 - Bidi Class (normative property) 126
 - Bidi Mirrored (normative property) 130
 - Bidi Mirroring Glyph (informative property) . . . 130
 - BidiMirroring.txt 130
 - Bidirectional Algorithm, Unicode 40, 63
 - bidirectional ordering 15
 - controls 136, 530
 - bidirectional text 40, 63
 - Middle Eastern scripts 237

nonspacing marks in158
 punctuation in184
 big-endian30
 definition63
 Bihari268
 binary comparison and sort order
 caution for UTF-1627
 UTF differences163, 164
 UTF-829
 blocks of the Unicode Standard34, 179
 Blocks.txt34
 BMP *see* Basic Multilingual Plane
 BNF (Backus-Naur Form)559
 BOCU-1 *see* UETN #6, BOCU-1
 MIME-Compatible Unicode Compression
 Bodhi315
 Bodo282
 BOM (U+FEFF)30, 50, 97–100, 537–538
 Bopomofo412–414
 boundaries, text8, 46, 135, 153–154, 162
 see also UAX #14, Unicode Line Breaking Algorithm
 see also UAX #29, Unicode Text Segmentation
 boustrophedon41, 455
 Brahmi267, 314, 340, 344–347, 350
 Braille511–513
 Breton208
 Buginese379–380
 Buhid378
 Bulgarian221
 bullets194
 Burmese *see* Myanmar
 Byelorussian221
 byte order mark (BOM) (U+FEFF) ..30, 50, 97–100, 537–538
 byte ordering
 changing61
 conformance62
 byte serialization30, 50
 Byzantine Musical Symbols517

C

C language
 wchar_t and Unicode142
 C0 and C1 control codes23, 134, 522
 Cambodian *see* Khmer
 camelcase169
 Canadian Aboriginal Syllabics443–444
 canonical composite characters
 see canonical decomposable characters
 canonical composition algorithm103
 canonical decomposable characters
 definition87
 canonical decomposition48
 definition87
 mappings86
 canonical equivalence
 definition87
 nonspacing marks159
 canonical equivalent character sequences
 conformance60, 61
 canonical mappings
 see canonical decomposition mappings
 canonical ordering algorithm103

canonical precomposed characters
 see canonical decomposable characters
 Cantonese399
 capital letters120, 166, 203
 Carian460
 carriage return (U+000D) (CR)148, 523
 carriage return and line feed (CRLF)148
 case209
 and text processes9
 beyond ASCII167
 camelcase169
 case folding170
 case operations (conformance)64, 111–116
 case operations and normalization171
 case operations, reversibility169
 cased (definition)112
 case-insensitive comparison ...115, 163, 164, 170
 casing context (definition)112
 conversion113
 detection114
 European alphabets203
 exceptional Latin pairs206, 209
 Georgian226
 lowercase120, 166, 203
 mapping tables140
 mappings112, 122, 166–169
 mappings noted in code charts551
 titlecase120, 167
 Turkish I168, 206
 uppercase120, 166, 203
 see also default case
 Case (normative property)120, 166
 CaseFolding.txt122, 170
 caseless letters209
 Catalan207
 cedilla205
 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)
 Cham376–377
 character encoding forms (CEF)24–29, 573
 see also Unicode encoding forms
 character encoding model24, 31
 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 Index566
 character literals, Unicode
 code point notation U+559
 character mapping
 interchange format *see* UTS #22, Character Mapping Markup Language (CharMapML)
 character names65, 130–134, 575
 aliases66, 131, 550
 conventions557
 for CJK ideographs553
 for control codes133, 134
 in code charts547–550
 matching131

- character properties
 - see* properties
 - see also individual properties, e.g.* Combining Class
- character semantics 1, 60, 64–65, 575
 - as Unicode design principle 14
 - ASCII 184
 - definition 65
- character sequences
 - abstract *see* abstract character sequences
 - canonical equivalent *see* canonical equivalent character sequences
 - compatibility equivalent *see* compatibility equivalent character sequences
 - conformance 60
 - named 131
- character sequences, combining 78
- character shaping selectors (deprecated) 532
- character tabulation (U+0009) 523
- characters
 - abstract *see* abstract characters
 - arrangement in Unicode 35
 - assigned 8, 23
 - blocks 34, 179
 - boundaries 153
 - canonical decomposable *see* canonical decomposable characters
 - classes 559
 - code charts 547–555, 565
 - coded *see* encoded characters
 - combining *see* combining characters
 - compatibility decomposable *see* compatibility decomposable characters
 - composite *see* decomposable characters
 - concept of 11, 46
 - conformance definitions 66–69
 - confusable 174
 - conversion 139–141
 - 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, 185, 521–545
 - glyphs, relationship to 11
 - graphic 23
 - identity (definition) 65
 - interpretation 59
 - layout control 51, 523–531
 - modification 61
 - names list 547–550
 - names *see* character names
 - not encoded in Unicode 2
 - number encoded in this and earlier versions . 577
 - number encoded in Version 6.0 2
 - precomposed *see* decomposable characters
 - properties *see* properties
 - semantics *see* character semantics
 - special 50, 521–545
 - supplementary *see* supplementary characters
 - transcoding 139–141
 - unsupported 142–144
- characters, not glyphs
 - in spoofing 174
 - Unicode principle 11
- CharMapML
 - see* UTS #22, Character Mapping Markup Language (CharMapML)
- charsets
 - IANA registered names 31
- charts, character code *see* code charts
- Cherokee 442
- Chinese 399–400
 - Cantonese 399
 - Hakka 414
 - Mandarin 399
 - Minnan (Hokkien/Fujian, incl. Taiwanese) . 414
 - simplified and traditional 399
- Chu hán 398
- Chu Nôm 586
- citations for
 - properties 58
 - Unicode algorithms 58
 - Unicode Standard 57
- CJK ideographs 182, 392–407
 - accounting numbers 129
 - CJK Compatibility Ideographs 406–407
 - CJK Compatibility Supplement 407
 - CJK Strokes 409, 589
 - CJK Unified Ideographs 392–406
 - CJK Unified Ideographs Extension A 396
 - CJK Unified Ideographs Extension B 406
 - CJK Unified Ideographs Extension C 406
 - CJK Unified Ideographs Extension D 406
 - code charts 553
 - compatibility ideographs in Plane 2 40
 - component structure 402
 - encoding blocks 395
 - ideographic description sequences 409–412
 - ideographic variation mark (U+303E) 412
 - KangXi radicals 405, 407–408
 - names 553
 - numeric values 129, 146
 - order of encoding 404
 - radicals 407–408
 - source standards 393–395
 - unknown or unavailable 200
 - Vietnamese 391
- CJK Miscellaneous Area 38
- CJK punctuation and symbols 199
 - compatibility forms 200
 - overscores and underscores 201
 - quotation marks 190
 - sesame dots 200
 - vertical forms 200
- CJK Radical (property) 410
- CJK-JRG (Chinese/Japanese/Korean Joint Research Group) 585
- CJKV Ideographs Area 38
- CLDR (Unicode Common Locale Data Repository) 566
 - cluster boundaries 153
- code charts 547–555, 565
 - representative glyphs 548
- code point sequences
 - notation 558
- code points 5, 22
 - assigned 8, 23
 - assignment 35, 577
 - categories 22

- default ignorable143, 175
- definition67
- designated23
- notation557
- number in Unicode Standard1
- private-use *see* private-use code points
- reserved *see* reserved code points
- semantics24
- surrogate *see* surrogates
- unassigned *see* unassigned code points
- undesignated23
- code positions *see* code points
- code set independence14
- code unit sequences
 - definition89
 - ill-formed (definition)90
 - notation558
 - well-formed (definition)90
- code units
 - definition89
 - isolated88
- code values *see* code units
- coded character representations
 - see* coded character sequences
- coded character sequences
 - definition67
- coded characters *see* encoded characters
- codespace *see* Unicode codespace
- coeng361, 362
- Collation Algorithm, Unicode (UCA)10
- collation *see* sorting
- collation tables140
- combining character sequences42, 78
 - defective158
 - definition80
 - Latin204
 - line breaking155
 - matching155
 - order of base character and marks155, 230
 - rendering155
 - selection153
 - truncation156–157
- combining characters41–46, 82–85, 154–162
 - blocking reordering529
 - canonical ordering47, 103, 122
 - class zero123
 - combining marks230–231
 - definition79
 - dependence230
 - display order43
 - keyboard input155
 - ligatures45
 - multiple43
 - multiple base characters45
 - normalization of147
 - ordering conventions42
 - rendering of marks157–162
 - reordrant123
 - script-specific42
 - split124
 - strikethrough126
 - subjoined125
 - typographical interaction43, 122
 - vertical stacking44
 - see also* diacritics
- Combining Class (normative property)122
- combining classes101, 122, 159–160
 - class zero characters122
 - definition101
- combining grapheme joiner (U+034F)529
- combining half marks136, 235
- combining marks *see* combining characters
- comma below205
- Compatibility and Specials Area20, 38
- compatibility characters18
- compatibility composite characters21
 - see* compatibility decomposable characters
- compatibility decomposable characters20
 - definition86
- compatibility decomposition48
 - definition86
- compatibility decomposition mappings86
- Compatibility Encoding Scheme for UTF-16
 - see* UTR #26, Compatibility Encoding Scheme for UTF-16: 8-Bit (CESU-8)
- compatibility equivalence
 - definition87
- compatibility equivalent character sequences
 - conformance61
- compatibility mappings
 - see* compatibility decomposition mappings
- compatibility precomposed characters
 - see* compatibility decomposable characters
- compatibility variants20
 - mapping172
- composite characters
 - see* decomposable characters
 - compatibility *see* compatibility decomposable characters
- Composition Exclusion (normative property)74
- compression147
 - see also* UTS #6, A Standard Compression Scheme for Unicode (SCSU)
- conferences566
- conformance55–116
 - clause and definition updates579
 - definitions64–69
 - examples51
 - ISO/IEC 10646 implementations575
 - requirements59–63
- confusables174
- conjunct consonants
 - Indic153, 271
 - Myanmar355
 - selection of clusters153
- contextual shaping
 - apostrophe192
 - Arabic243
 - not used for Hebrew final forms239
 - quotation marks189
 - Syriac260
- contour tones229
- control codes23, 51, 522
 - graphics for495
 - names134
 - properties523
 - semantics24, 523
 - specified in Unicode523
- control sequences522
- conversion of characters95, 139–141, 177

- convertibility
 - as Unicode design principle 19
 - Coptic 217, 219–221
 - corporate use subarea 534
 - corrigenda 57
 - CR (U+000D carriage return) 148, 523
 - CRLF (carriage return and line feed) 148
 - Croatian 208
 - digraphs 208
 - culturally expected sorting 10, 162
 - Cuneiform
 - Old Persian 470
 - Sumero-Akkadian 471–473
 - Ugaritic 469
 - Cuneiform and Hieroglyphic Area 39
 - currency symbols 478–480
 - encoded in script blocks 479
 - cursive joining 525–529
 - Arabic 247–253
 - control characters for 136, 243–244, 429, 525
 - Mandaic 466
 - Mongolian 428–430
 - N’Ko 439
 - Syriac 260–262
 - transparency 528
 - cursive scripts 237
 - Cypriot 459–460
 - see also* Linear B
 - Cyrillic 221–223
 - Czech 208
- D**
- danda, in Devanagari block 281
 - Danish 207
 - dashes 187
 - Database, Unicode Character
 - see* Unicode Character Database (UCD)
 - dead consonants, Indic 271
 - dead keys 155
 - decomposable characters 48
 - definition 86
 - normalization of 147
 - decomposition 48, 86–88
 - canonical *see* canonical decomposition
 - compatibility *see* compatibility decomposition
 - definition 86
 - in normalization 146
 - mapping, definition 86
 - mappings noted in code charts 552
 - default case
 - algorithms 64, 111–116
 - conversion 113
 - detection 114
 - folding 113
 - default caseless matching 115
 - default grapheme clusters 153
 - see also* UAX #29, Unicode Text Segmentation
 - Default Ignorable Code Point (property) 175
 - default ignorable code points 143, 175
 - default property values 143
 - definition 71
 - defective combining character sequences 158
 - definition 80
 - dependent vowel signs
 - Indic 270
 - Khmer 364
 - Philippine scripts 378
 - deprecated characters 55, 550
 - alternate format 136, 531–532
 - definition 68
 - Derived Age (property) 144
 - derived properties
 - definition 77
 - DerivedCoreProperties.txt 112, 120, 175
 - DerivedNormalizationProps.txt 172
 - Deseret 445–446
 - design goals of Unicode 3
 - design principles of Unicode 10–19
 - designated code points 23
 - Devanagari 268–285
 - Dhivehi 264
 - diacritics 42, 230
 - alternative glyphs 204, 231
 - Czech 205
 - display in isolation 45, 186, 231
 - double 84, 136, 232
 - Greek 214–215, 218
 - Latin 204–206
 - Latvian 205
 - mathematical 485
 - on i and j 206
 - rendering 157–162
 - Slovak 205
 - spacing clones of 229, 231
 - symbol 42, 234
 - see also* combining characters
 - dictionary symbols 502
 - digit form names 245
 - digits 145
 - Arabic-Indic 244–245
 - decimal 128
 - national shapes 532
 - digraphs 208, 210, 212
 - dingbats 504
 - directionality 15, 40
 - East Asian scripts 392
 - Middle Eastern scripts 237
 - Mongolian 428
 - musical symbols 514
 - normative property 126
 - Ogham 452
 - Old Italic 454
 - Philippine scripts 379
 - Runic 455
 - discussion list for Unicode 566
 - Dogri 282
 - Domino Tiles 505
 - dotless i 168, 206
 - dotted circle
 - in code charts 79, 231
 - in fallback rendering 157
 - to indicate diacritic 41
 - to indicate vowel sign placement 43
 - double diacritics 84, 136, 232
 - Dutch 207, 208
 - dynamic composition
 - as Unicode design principle 18
 - Dzongkha 315

E

East Asian scripts 391–422
 writing direction 41
 see also CJK ideographs
 Eastern Arabic-Indic digits 244
 EBCDIC
 newline function 149
 see UTR #16, UTF-EBCDIC
 editing, text boundaries for 153–154
 efficiency
 as Unicode design principle 11
 Egyptian Hieroglyphs 473–476
 e-mail discussion list for Unicode 566
 emoji 500
 animal symbols 503
 cultural symbols 503
 zodiacal symbols 503
 Emoticons 503
 Enclosed Alphanumerics 510
 enclosing marks 235
 definition 80
 encoded characters 5, 22
 allocation 33–40, 577
 definition 67
 encoding form conversion
 definition 94
 encoding forms 24–29
 ISO/IEC 10646 definitions 573
 encoding forms, Unicode
 see Unicode encoding forms
 encoding model for Unicode characters 24, 31
 see also UTR #17, Unicode Character Encoding Model
 encoding schemes 30–32
 encoding schemes, Unicode
 see Unicode encoding schemes
 endian ordering
 see byte order mark (BOM) (U+FEFF)
 end-user subarea 535
 English 207
 equivalent sequences 146
 as Unicode design principle 18
 case-insensitivity 164, 170
 combining characters in matching 155
 conformance 61
 Hangul syllables 419
 in sorting and searching 162
 language-specific 87
 security implications 173
 see also canonical equivalence
 see also compatibility equivalence
 see also encoding forms, encoding schemes
 errata xxxvi, 57, 566
 escape sequences 522
 not used in Unicode 1, 3
 Esperanto 208
 Estonian 208
 Ethiopic 424–426
 Etruscan 453
 euro sign (U+20AC) 480
 European alphabetic scripts 203–235
 eyelash-RA 276

F

fallback rendering of nonspacing marks 157
 FAQ (Frequently Asked Questions) 566
 Faroese 207
 Farsi 242, 243
 featural syllabaries 181
 FF (U+000C form feed) 148, 523
 file separator (U+001C) 523
 Finnish 207
 Finno-Ugric Transcription (FUT)
 see Uralic Phonetic Alphabet (UPA)
 fixed-width Unicode encoding form (UTF-32) 26, 92
 flat tables 140
 Flemish 207
 fonts
 and Unicode characters 13
 for mathematical alphabets 484–485
 style variation for symbols 477
 form feed (U+000C) (FF) 148, 523
 format control characters 23, 51, 185, 521–545
 deprecated 531–532
 prefixed 136
 reserved ranges 144
 stateful 176
 fraction characters 485
 fraction slash (U+2044) 192, 486
 French 208
 Frisian 208
 FTP site, Unicode Consortium 565
 fullwidth forms in East Asian encodings 416–417
 futhark 456

G

Garshuni 256
 Ge'ez 424
 General Category (normative property) 126
 list of values 126
 general punctuation 183–201
 General Scripts Area 38
 geometrical symbols 498–500
 Georgian 225–227
 German 207
 geta mark (U+3013) 200
 Glagolitic 223–224
 Glossary 566
 glyph selection tables 140
 glyphs 5, 12
 characters, relationship to 11
 diacritics alternative 204, 231
 Greek alternative 215–216
 Latin alternative 204
 mathematical alternative 490
 missing 175
 representative in code charts 548
 standardized variants 533
 symbols alternative 477
 golden numbers 457
 Gothic 457–458
 grapheme base 230
 definition 80
 grapheme clusters 8, 46
 see also UAX #29, Unicode Text Segmentation
 default 153
 definition 81

- grapheme extender
 definition 81
- grapheme joiner, combining (U+034F) 529
- graphic characters 23
- Greek 214–218
 acrophonic numerals 146, 218
 alternative glyphs 215–216
 ancient musical notation 517–519
 letters as symbols 215–217, 491
see also Cypriot, Linear B
- Greenlandic 208
- group separator (U+001D) 523
- guillemets 189
- Gujarati 293–294
- Gurmukhi 289–293
- H**
- Hakka 414
- halant 267
see also virama
- half marks, combining 136, 235
- half-consonants, Indic 272
- halfwidth forms in East Asian encodings ... 416–417
- Han ideographs *see* CJK ideographs
- Han unification 400–406
 and language tags 152
 history 585–587
 language usage 398
 source separation rule 396, 401
 source standards 393–395
- Hangul Area 38
- Hangul syllables 391, 417–419
 and combining marks 85
 as grapheme clusters 46
 boundary determination 107
 canonical decomposition 110
 collation 419
 composition 109
 conjoining jamo 106–111
 equivalent sequences 419
 Hangul Compatibility Jamo 418
 Hangul Jamo 417–419
 Hangul Syllables block 418–419
 Johab set 418
 name generation 111
 normalization 418
 precomposed 107
 standard 108
- Hangzhou numerals 488
- Hanja *see* CJK ideographs
- Hanunóo 378
- Hanzi *see* CJK ideographs
- harakat, Arabic pronunciation marks 242
- hasant 285
- hash tables 140
- Hebrew 238–242
- hentaigana 415
- hieroglyphs, Egyptian 473–476
- high surrogate
 definition 88
 high-surrogate code points 59, 535
 high-surrogate code units 88
- higher-level protocols
 definition 68
- Hindi 268
- Hiragana 414–415
- historic scripts 451–460
- horizontal tab (U+0009) 523
- HTML newline function 149
- Hungarian 208
- hyphenation 525
 as a text process 8
- hyphens 187, 525
- I**
- I Ching symbols 506
- IANA charset names 31
- Icelandic 207
- identifiers 162
see also UAX #31, Unicode Identifier and Pattern
 Syntax
- Ideographic (informative property) 135
- Ideographic Rapporteur Group (IRG) 393, 586
- Ideographic Variation Database *see* UTS #37, Unicode
 Ideographic Variation Database
- ideographs *see* CJK ideographs
- IDNA *see* UTS #46, Unicode IDNA Compatibility
 Processing
- IICore 397, 586
- ill-formed
 definition 90
- Imperial Aramaic 464–465
- implementation guidelines 139–177
- in a Unicode encoding form
 definition 91
- in-band mechanisms 544
- Indian rupee sign (U+20B9) 480
- Indic scripts 267–311, 313–315
 principles, in terms of Devanagari 269–275
 relation to ISCII standard 268
- Indonesian 207
- industry character sets
 covered in Unicode 2
- information separators (U+001C..U+001F) 523
- informative properties
 definition 74
- Inscriptional Pahlavi 467
- Inscriptional Parthian 467
- inside-out rule 157
- interchange restrictions 23
- International Phonetic Alphabet (IPA) . 180, 210–211
 Spacing Modifier Letters 228
see also phonetic alphabets
- internationalization 14
- Internationalization & Unicode Conference 566
- Internet protocols
 UTF-8 as preferred encoding 28
- Inuktitut 443
- invisible operators 494
- iota subscript 215
- IPA *see* International Phonetic Alphabet
- IRG (Ideographic Rapporteur Group) 393, 586
- Irish 207, 452
- ISCII standard and Unicode 268
- ISO/IEC 10646 569–575
 conformance of Unicode implementations ... 575
 encoding forms 573

- synchrony with Unicode Standard 574
 - timeline compared to Unicode versions 570
 - Italian 207
 - ITC Zapf Dingbats 504
 - IUC *see* Internationalization & Unicode Conference
- J**
- Jamo.txt 111
 - jamos *see* Hangul syllables
 - Japanese 391
 - Javanese 385–387
 - Jawi 254
 - Johab 418
 - joiners 243
 - combining grapheme joiner (U+034F) 529
 - word joiner (U+2060) 524
 - zero width joiner (U+200D) 243–244, 526
 - justification 159
- K**
- Kaithi 335–337
 - Kana (Hiragana and Katakana) 414–416
 - Kanbun 407
 - KangXi radicals 405, 407–408
 - Kanji *see* CJK ideographs
 - Kannada 304–306
 - Kashmiri 283
 - Katakana 415–416
 - Kawi 380, 382
 - Kayah Li 375–376
 - KC (normalization form)
 - see* Normalization Form KC
 - KD (normalization form)
 - see* Normalization Form KD
 - keytop labels 495
 - Khamti Shan 358
 - Kharoshthi 340–341
 - Khmer 360–369
 - characters not recommended 366
 - syllable components, order of 367
 - killer 181
 - Batak 388
 - Brahmi 345
 - Meetei Mayek 338
 - Myanmar (asat) 356
 - see also* virama
 - Konkani 282
 - Korean Hangul *see* Hangul
 - Kurdish 242
- L**
- Ladino 238
 - language tags 152, 541–544
 - and Han unification 152
 - use strongly discouraged 544
 - Lanna 370
 - Lao 352–354
 - last-resort glyphs 175
 - Latin 204–214
 - alternative glyphs 204
 - Basic Latin 207
 - encoding blocks 34
 - IPA Extensions 210–211
 - Latin Extended Additional 212–214
 - Latin Extended-A 207
 - Latin Extended-B 208–209
 - Latin Extended-C 212
 - Latin Extended-D 213
 - Latin Ligatures 212
 - Latin-1 Supplement 207
 - Phonetic Extensions 211–212
 - Latvian 208, 213
 - cedilla 205
 - layout control characters 51, 523–531
 - leading surrogates
 - see* high-surrogate code units
 - legibility criterion for plain text 15
 - Lepcha 324–326
 - letter spacing 524
 - letterlike symbols 480–485
 - LF (U+000A line feed) 148, 523
 - ligatures 525–529
 - Arabic 249–250
 - combining characters on 45
 - control characters for 136
 - for nonspacing marks 160
 - Latin 212
 - selection 154
 - Syriac 262
 - Limbu 331–334
 - line breaking 148–151, 524–525
 - control characters 138
 - in South Asian scripts 352, 357, 369
 - recommendations 150
 - see also* UAX #14, Unicode Line Breaking Algorithm
 - line feed (U+000A) (LF) 148, 523
 - line separator (U+2028) (LS) 148, 525
 - line tabulation (U+000B) (VT) 523
 - Linear B 458–459
 - see also* Cypriot
 - linear boundaries 154
 - Lisu 447–449
 - Lithuanian 208
 - little-endian 30
 - definition 63
 - Locale Data Markup Language
 - see* UTS #35, Unicode Locale Data Markup Language (LDML)
 - logical order
 - as Unicode design principle 15
 - exceptions to 124
 - logosyllabaries 182
 - low surrogate
 - definition 88
 - low-surrogate code points 59, 535
 - low-surrogate code units 88
 - lowercase 120, 166, 203
 - LS (U+2028 line separator) 148, 525
 - Lycian 460
 - Lydian 460
- M**
- MacOS newline function 149
 - Mahjong Tiles 505
 - mail discussion list for Unicode 566
 - Maithili 282

- major version 56
- Malay 207
- Malayalam 307–311
- Maltese 208
- Manchu 427
- Mandaic 465–467
- Mandarin 399
- Manden 436
- map symbols 502
- mapping tables *see* tables of character data
- Marathi 268, 276, 280
- markup languages
 - and Unicode conformance 544
 - line breaking 148
 - see also* UTR #20, Unicode in XML and Other Markup Languages
- Mathematical (informative property) 489
- mathematical expression format characters 136
 - see also* UTR #25, Unicode Support for Mathematics
- mathematical symbols 489–494
 - alphabets 482–485
 - alphanumeric 481–485
 - fonts 484–485
 - format characters 494
 - fragments for typesetting 496
 - invisible operators 494
 - operators 490–491
 - standardized variants 494
- MathML 491
- matras 123, 270
- Meetei Mayek 338–339
- Middle Eastern scripts 237–265
- Min 399
- Minnan (Hokkien/Fujian, incl. Taiwanese) 414
- minor version 56
- minus sign 491
 - commercial (U+2052) 195
- mirrored property
 - see* Bidi Mirrored (normative property)
- mirroring of paired punctuation 189
- Miscellaneous Symbols 501
- missing glyphs 175
- modifier letters 227–230
- Modifier Letters, Spacing 212
- Mongolian 326, 426–433
 - writing direction 428
- multibyte encodings
 - compared to UTF-8 28
- multistage tables 140
- musical symbols 513–519
 - ancient Greek 517–519
 - Balinese 384
 - Byzantine 517
 - directionality 514
 - Gregorian 514
 - Western 513–516
- Myanmar 354–359
 - Myanmar Extended-A 357
- N**
- N’Ko 436–440
- named character sequences 131
- names, character *see* character names
- namespace 66
- NEL (U+0085 next line) 148, 523
- Nepali 268
- neutral directional characters 126
- New Tai Lue 370–371
- newline function (NLF) 149, 523
- newline guidelines 148–151
- next line (U+0085) (NEL) 148, 523
- NFC (Normalization Form C) 47
- NFD (Normalization Form D) 47
- NFKC (Normalization Form KC) 47
- NFKD (Normalization Form KD) 47
- NLF (newline function) 149, 523
- no-break space (U+00A0) 524
 - base for diacritic in isolation 45, 186, 231
- no-break space, narrow (U+202F) 431
- noncharacter code points *see* noncharacters
- noncharacters 23, 50, 536
 - conformance 59
 - definition 68
 - handling 61
 - in code charts 550
 - interchange restrictions 24
 - semantics 24
 - U+10FFFF (not a character code) 536
 - U+FDD0..U+FDEF 23, 536
 - U+FFFE (not a character code) 50, 536
 - U+FFFF (not a character code) 23, 536
- nondecomposable characters 48
- non-joiner, zero width (U+200C) 243–244, 527
- nonlinear boundaries 154
- non-overlap principle in Unicode encoding forms 24
- nonspacing marks 230
 - definition 79
 - display in isolation 45, 186, 231
 - positioning 160
 - rendering 157–162
 - see also* combining characters
 - see also* diacritics
- normalization 47, 146–147
 - and case operations 171
 - canonical ordering algorithm 47, 103, 122
 - conformance 63
 - of private-use characters 534
 - see also* 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 100–106
 - definition 105
 - specification 102
- normalization stability 101
- normative behaviors
 - definition 65
- normative properties
 - definition 73
 - list 73
 - may change 73
- Norwegian 207
- notational conventions 557–560
- notational systems 183
- nukta 255, 277
- null (U+0000)
 - as Unicode string terminator 523

- number forms 485–488
 - CJK ideographs 146
 - numbers
 - handling 145
 - ideographic accounting 129
 - numerals
 - acrophonic 218
 - Chinese counting rods 487
 - Coptic 221
 - Cuneiform 473
 - Ethiopic 425
 - Greek acrophonic 146
 - Hangzhou 488
 - old-style 193
 - Roman 146, 486
 - Rumi 487
 - Suzhou-style 488
 - numeric separators 195
 - numeric shape selectors (deprecated) 532
 - Numeric Type (normative property) 128
 - Numeric Value (normative property) 128
 - numero sign (U+2116) 480
- O**
- object replacement character (U+FFFC) 540
 - octet 559
 - Ogham 452–453
 - Ol Chiki 339–340
 - Old Italic 453–455
 - Old Persian 470
 - Old South Arabian 461–463
 - Old Turkic 458
 - old-style numerals 193
 - Oriya 294–296
 - Oromo 424
 - Osmanya 433
 - out-of-band mechanisms 544
 - overlapping encodings 24
 - overscores 192
- P**
- Pahlavi, Inscriptional 467
 - Panjabi 289
 - paragraph or section marks 194
 - paragraph separator (U+2029) (PS) 148, 525
 - Parthian, Inscriptional 467
 - Pashto 242
 - Persian 242, 243
 - Phags-pa 326–331
 - Phaistos Disc symbols 507
 - Phake 359
 - Philippine scripts 378–379
 - Phoenician 463
 - phonemes 182
 - phonetic alphabets 180
 - IPA Extensions 210–211
 - Phonetic Extensions 211–212
 - Spacing Modifier Letters 228–230
 - Uralic Phonetic Alphabet (UPA) 195, 211
 - see also* International Phonetic Alphabet (IPA)
 - phonetic extensions 212, 213
 - Pinyin 207
 - pivot code, Unicode as 140
 - 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, 40
 - Planes 15-16 (Private Use) 40, 535
 - Playing Cards 506
 - points, Hebrew pronunciation marks 238
 - policies of the Unicode Consortium 566
 - Polish 208
 - Portuguese 207
 - precomposed characters
 - see* decomposable characters
 - compatibility *see* compatibility decomposable characters
 - prefixed format control characters 136
 - Private Use Area (PUA) 38, 534
 - Private Use planes 34, 40, 535
 - private-use characters
 - semantics 24
 - private-use code points 23, 142
 - conformance 60
 - definition 78
 - high surrogates 535
 - processing code, choice of Unicode encoding form 28
 - properties 14, 69–78, 117–138
 - aliases 119
 - aliases (definition) 77
 - and Unicode algorithms 73
 - data tables 140
 - 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 523
 - provisional *see* provisional properties
 - simple *see* simple properties
 - see also individual properties, e.g. combining classes*
 - property values
 - aliases 119
 - aliases (definition) 77
 - default 143, 534
 - default (definition) 71
 - normative references to 63
 - PropertyAliases.txt 77, 559
 - PropertyValueAliases.txt 77, 559
 - PropList.txt 122
 - Provençal 208
 - provisional properties
 - definition 75
 - PS (U+2029 paragraph separator) 148, 525
 - PUA (Private Use Area) 38, 534
 - pulli* 296
 - punctuation 183–201
 - blocks containing 179
 - CJK 199
 - doubled 192
 - in bidirectional text 184
 - paired 189
 - small form variants 201

typographic forms	184
vertical forms	200
Punctuation and Symbols Area	38
Punjabi	289

Q

quotation marks	189–191
East Asian	190
European	189

R

radicals, KangXi and other CJK	407–408
radical-stroke index	405
record separator (U+001E)	523
recycling symbols	502
referencing	63
properties	58
Unicode algorithms	58
Unicode Standard	57
regional indicator symbols	511
regular expressions	151
and line breaking	148
<i>see also</i> UTS #18, Unicode Regular Expressions	
Rejang	387–388
rendering of text	5, 8, 13
unsupported characters	143
repertoire of abstract characters	22
replacement character (U+FFFD)	32, 51, 62, 95, 177, 541
reserved code points	23, 142
definition	68
in code charts	550
preservation in interchange	24
<i>see also</i> unassigned code points	
Rhaeto-Romanic	208
rich text	14
right single quotation mark (U+2019)	
preferred for apostrophe	191
right-to-left text	40
East Asian scripts	392
Middle Eastern scripts	237
roadmap for script additions	34
Roman numerals	146, 486
Romanian	208
comma below	205
Romany	208
Rumi numeral forms	487
Runic	455–457
rupee sign, Indian (U+20B9)	480
Russian	221

S

Samaritan	263–264
Sami	208
Sanskrit	268
Saurashtra	337–338
scalar values, Unicode	
<i>see</i> Unicode scalar values	
scripts	
in Unicode Standard	2
roadmap for future additions	34
types of	183
<i>see also</i> UAX #24, Unicode Script Property	

SCSU

see UTS #6, A Standard Compression Scheme for Unicode

searching	162–164
as a text process	8
case-insensitive	164, 170
section or paragraph marks	194
security issues	173
self-synchronization of encoding forms	25
semantics	
<i>see</i> character semantics	
sequences	
notation	558
Serbian	
corresponding digraphs in Croatian	208
Shan	369
Shavian	447
Show Hidden	61, 157, 176, 533
SHY (U+00AD soft hyphen)	525
Sibe	428
signature for Unicode data	51, 537–538
simple properties	
definition	77
simplified Chinese	399
Sindhi	242, 282
Sinhala	314–315
SIP (Supplementary Ideographic Plane)	33, 40
slash, fraction (U+2044)	192
Slovak	208
Slovenian	208
small letters	120, 166, 203
SMP (Supplementary Multilingual Plane)	33, 39
soft hyphen (U+00AD) (SHY)	525
Somali	433
Sorbian	208
sorting	10, 162
and combining grapheme joiner	530
as a text process	8
case-insensitive	163
culturally expected	10, 162
language-insensitive	163
<i>see also</i> Unicode Collation Algorithm (UCA)	
source separation rule	396, 401
South Asian scripts	267–311, 313–334
Southeast Asian scripts	349–379
space (U+0020)	
base for diacritic in isolation	46, 186, 231
space characters	186, 524–525
graphics for	495
space, zero width (U+200B)	187
spacing clones of diacritics	229, 231
spacing marks	230
definition	80
Spacing Modifier Letters	228–230
Spanish	207
special characters	50, 521–545
SpecialCasing.txt	112, 122
Specials	537–541
spell-checking	
as a text process	8
spellings, alternative	
<i>see</i> equivalent sequences	
spoofing	174
SSP (Supplementary Special-purpose Plane)	33

- stability75, 118
 - as Unicode design principle18
 - stacked boundaries153
 - stacking sequences43
 - nondefault44
 - Standard Compression Scheme for Unicode (SCSU)
 - see UTS #6, A Standard Compression Scheme for Unicode
 - standard Korean syllables108
 - standardized variants430, 533
 - mathematical symbols494
 - StandardizedVariants.txt430, 494
 - standards coverage2
 - starters102
 - stateful encoding
 - not used in Unicode3
 - paired format controls176
 - string comparison10
 - string literals, Unicode
 - code point notation `\u1234`559
 - strings, Unicode32, 90
 - null termination523
 - strong directional characters126
 - styled text14
 - sublinear searching164
 - subsets, supported53
 - conformance60
 - ISO/IEC 10646 specification for574
 - substitution character
 - see replacement character
 - Sumero-Akkadian471–473
 - Sundanese389–390
 - superscripts228
 - and subscripts488
 - supplementary characters
 - in UTF-16 strings32
 - tables for140
 - Supplementary General Scripts Area38
 - Supplementary Ideographic Plane (SIP)33, 40
 - Supplementary Multilingual Plane (SMP)33, 39
 - supplementary planes
 - representation in UTF-1627
 - representation in UTF-828
 - Supplementary Private Use Areas40, 535
 - Supplementary Special-purpose Plane (SSP)33
 - supported subsets53
 - conformance60
 - supralineation220
 - surrogate code points
 - see surrogates
 - surrogate pairs27, 93
 - definition88
 - processing28, 144–145
 - surrogates23, 88, 535
 - interchange restrictions23
 - isolated surrogates, handling32
 - isolated surrogates, ill-formed93
 - isolated surrogates, uninterpreted88
 - support levels145
 - Surrogates Area38, 535
 - Suzhou-style numerals488
 - svasti signs321
 - Swahili207
 - Swedish207
 - syllabaries180
 - alphabetic property135
 - featural181
 - Syloti Nagri334–335
 - symbols477–519
 - animal503
 - appearance variation477
 - arrows493
 - cultural503
 - currency478–480
 - dictionary502
 - dingbats504
 - emoji500, 511
 - Enclosed Alphanumerics510
 - fragments for mathematical typesetting496
 - game502
 - gender502
 - genealogical502
 - geometrical498–500
 - Khmer lunar calendar369
 - letterlike480–485
 - map502
 - mathematical489–494
 - mathematical alphanumeric481–485
 - miscellaneous501
 - musical513–519
 - number forms485–488
 - recycling502
 - regional indicator511
 - technical495–498
 - weather501
 - zodiacal503
 - symmetric swapping format characters (deprecated)531
 - Syriac256–262
- ## T
- tab (U+0009 character tabulation)523
 - tab, vertical (U+000B)148, 523
 - tables of character data139–141
 - optimization140
 - supplementary characters140
 - tag characters541–545
 - Tagalog378
 - Tagbanwa378
 - tags, language152, 541–544
 - use strongly discouraged544
 - Tai Le369–370
 - Tai Tham371–373
 - Tai Viet373–375
 - Tai Xuan Jing symbols506
 - Tamil296–303
 - TCHAR in Win32 API142
 - Technical Notes (UTN)565
 - Technical Reports (UTR)562
 - abstracts563
 - Technical Standards (UTS)xxxvi, 562
 - abstracts562
 - technical symbols495–498
 - Telugu303–304
 - terminal emulation478

- text boundaries 8, 46, 135, 153–154, 162
 - see also* UAX #14, Unicode Line Breaking Algorithm
 - see also* UAX #29, Unicode Text Boundaries
 - text elements 5, 8, 153
 - boundaries 162
 - for sorting 163
 - variable-width nature 29
 - text processes 4, 8–10
 - text rendering 5, 8, 13
 - text selection, boundaries for 153–154
 - Thaana 264–265
 - Thai 350–352
 - Tibetan 315–324
 - Tifinagh 434
 - Tigre 424
 - tilde (U+007E) 195
 - titlecase 120, 167
 - Todo 427
 - tone letters 229–230
 - tone marks
 - Bopomofo spacing 413, 414
 - Chinantec 230
 - Chinese 229
 - Tai Le 369
 - Thai 350
 - Vietnamese 206
 - traditional Chinese 399
 - traffic signs 502
 - trailing surrogates
 - see* low-surrogate code units
 - transcoding 139–141
 - tables 140
 - Transport and Map Symbols 503
 - triangulation in transcoding 140
 - tries 140
 - truncation
 - combining character sequences 156–157
 - surrogates and 145
 - Turkish 208
 - case mapping of I 168, 206
 - cedilla 205
 - two-stage tables 140
- U**
- U+ notation 559
 - U+10FFFF (not a character code) 536
 - U+FEFF (BOM) 537–538
 - U+FFFE (not a character code) 536
 - U+FFFF (not a character code) 536
 - UAX (Unicode Standard Annex) xxxv, 562
 - as component of Unicode Standard 59
 - conformance 64
 - list of 64
 - UCA *see* Unicode Collation Algorithm
 - UCD *see* Unicode Character Database
 - UCS (Universal Character Set)
 - see* ISO/IEC 10646
 - UCS-2 573
 - UCS-4 573
 - Ugaritic 469–470
 - Uighur 326, 427
 - Ukrainian 221
 - unassigned code points 23, 59, 143
 - defined as reserved code points 68
 - handling 55
 - properties of 143
 - semantics 59
 - see also* reserved code points
 - underscores 192
 - undesignated code points 23
 - Unicode 1.0 Name (informative property) 134
 - 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) . xxxvi, 118, 566
 - as component of Unicode Standard 59
 - changes 56
 - properties in 34
 - Unicode character encoding model 24, 31
 - see also* UTR #17, Unicode Character Encoding Model
 - Unicode character literals
 - code point notation U+ 559
 - Unicode codespace
 - allocation numbers 577
 - definition 67
 - planes 33
 - size 1, 22
 - Unicode Collation Algorithm (UCA) 10
 - see also* UTS #10, Unicode Collation Algorithm
 - Unicode Common Locale Data Repository (CLDR) 566
 - Unicode conferences 566
 - Unicode Consortium 561
 - addresses 567
 - Consortium membership in standards bodies 561
 - e-mail discussion list 566
 - FTP site 565
 - membership 561
 - policies 566
 - Web site 565
 - Unicode data signature 51, 537–538
 - Unicode data types 141–142
 - for C 141–142
 - Unicode encoding forms 88–94
 - advantages of each 28
 - conformance 26, 62
 - definition 89
 - fixed-width (UTF-32) 26, 92
 - signatures 538, 539
 - variable-width 27, 92, 93
 - see also* encoding forms
 - Unicode encoding schemes
 - conformance 97–100
 - definition 97
 - endian ordering 30
 - see also* encoding schemes
 - Unicode escape sequence notation \u1234 559
 - 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 security173
 - Unicode Standard
 - allocation of encoded characters33–40
 - architecture7–10
 - areas34
 - benefits1
 - blocks34, 179
 - code charts547–555, 565
 - components59
 - conformance55–116
 - conformance of ISO/IEC 10646 implementations
 -575
 - corrections57
 - definitions for conformance64–69
 - design goals3
 - design principles10–19
 - errata57, 566
 - normative references to57, 63
 - number of characters2, 577
 - number of code points1, 22
 - script coverage2
 - security issues173
 - synchrony with ISO/IEC 10646574
 - updates566
 - versions *see* versions of the Unicode Standard
 - see also* Version 6.0
 - Unicode Standard Annexes (UAX)xxxv, 562
 - as components of Unicode Standard59
 - conformance64
 - list of64
 - Unicode string literals
 - code point notation `\u1234`559
 - Unicode strings32
 - definition90
 - Unicode Technical Committee (UTC)561
 - Unicode Technical Notes (UTN)565
 - Unicode Technical Reports (UTR)562
 - abstracts563
 - Unicode Technical Standards (UTS)xxxvi, 562
 - abstracts562
 - UnicodeData.txt112, 122
 - unification
 - as Unicode design principle17
 - see also* Han unification
 - Unified CJK Ideograph (property)410
 - Unified Repertoire and Ordering (URO) ...401, 586
 - see also* Han unification
 - Unihan Database118, 404, 405, 553, 566, 587
 - Unihan.zip75, 118
 - unit separator (U+001F)523
 - Universal Character Set (UCS)
 - see* ISO/IEC 10646
 - universality
 - as Unicode design principle10
 - Unix
 - and UTFs29
 - newline function149
 - UTF-32 in27
 - UTF-8 in14
 - unsupported characters142–144
 - update version57
 - uppercase120, 166, 203
 - Uralic Phonetic Alphabet (UPA)195, 211
 - Urdu242
 - URO (Unified Repertoire and Ordering) ...401, 586
 - see also* Han unification
 - UTF, Unicode Transformation Formats24, 89
 - advantages of each28
 - as encoding form or scheme99
 - binary comparison and sort order differences ...
 - 163,164
 - in APIs142
 - UTF-1627, 92, 574
 - binary comparison and sort order caution ...27
 - bit distribution (table)93
 - BOM in98, 537
 - encoding form (definition)92
 - encoding scheme (definition)98
 - encoding schemes30
 - in ISO/IEC 10646574
 - in UTF-8 order165
 - surrogates and string handling32, 144
 - UTF-16BE (Big-endian)538
 - encoding scheme30
 - encoding scheme (definition)97
 - UTF-16LE (Little-endian)538
 - encoding scheme30
 - encoding scheme (definition)97
 - UTF-3226, 92
 - BOM in99
 - encoding form (definition)92
 - encoding scheme (definition)99
 - encoding schemes30
 - in Unix27
 - UTF-32BE (Big-endian)
 - encoding scheme30
 - encoding scheme (definition)98
 - UTF-32LE (Little-endian)
 - encoding scheme30
 - encoding scheme (definition)98
 - UTF-827, 93, 573
 - ASCII transparency27
 - binary comparison and sort order29
 - bit distribution (table)93
 - BOM in97, 100, 537
 - byte ranges93
 - compared to multibyte encodings28
 - encoding form (definition)93
 - encoding scheme30
 - encoding scheme (definition)97
 - in Unix14
 - in UTF-16 order165
 - non-shortest form is invalid93, 174
 - preferred encoding for Internet protocols ...28
 - security and174
 - signature97, 100, 537
 - UTF-EBCDIC
 - see* UTR #16, UTF-EBCDIC
 - UTN (Unicode Technical Note)565
 - UTR (Unicode Technical Report)562
 - abstracts563
 - UTS (Unicode Technical Standard)xxxvi, 562
 - abstracts562
- ## V
- Vai440–441
 - valid (synonym for well-formed)91

- variable-width Unicode encoding form . . . 27, 92, 93
 - variants
 - compatibility 20
 - fullwidth and halfwidth 201
 - mathematical symbols 494
 - small form 201
 - standardized 533
 - variation selectors 137, 532
 - ideographic variation mark (U+303E) 412
 - Mongolian free variation selectors 430
 - variation sequences 532
 - for Phags-pa 330–331
 - Version 6.0 59
 - correlation with ISO/IEC 10646 572
 - number of characters 2, 577
 - versions of the Unicode Standard .xxxvi, 55, 566, 577–578
 - backward compatibility 55
 - compared to ISO/IEC 10646 editions 577
 - content 56
 - interaction in implementations 144
 - numbering 56
 - property changes 56
 - stability 56
 - updates 566
 - vertical tab (U+000B) 148, 523
 - vertical text 41, 184, 200
 - East Asian scripts 392
 - Mongolian 428
 - Vietnamese 206, 212
 - ideographs 391
 - virama 181, 267
 - definition 270
 - Kharoshthi 344
 - Khmer 362
 - Myanmar 355
 - Philippine scripts 378
 - virama-like characters 137
 - visual order used for Thai and Lao 16
 - vowel harmony
 - Mongolian 431
 - vowel marks, Middle Eastern scripts 237
 - vowel separator
 - Mongolian 432
 - vowel signs
 - Indic 43, 270
 - Khmer 364
 - Philippine scripts 378
- W**
- wchar_t
 - and Unicode encoding forms 28
 - in C language 142
 - weak directional characters 126
 - weather symbols 501
 - Web site, Unicode Consortium 565
 - Weierstrass elliptic function symbol 481
 - well-formed
 - definition 90
 - Welsh 208
 - Where Is My Character? 567
 - wide characters
 - data type in C 142
 - wiggly fence (U+29DB) 492
 - Windows newline function 149
 - word breaks 155, 524–525
 - in South Asian scripts 352, 357, 369
 - word joiner (U+2060) 524
 - writing direction *see* directionality
 - writing systems 180–183
 - Wu (Shanghainese) 399
- X**
- Xibe 428
 - Xishuang Banna Dai 370
 - XML
 - see* UTR #20, Unicode in XML and Other Markup Languages
- Y**
- yen currency sign 479
 - Yi 420–422
 - Yiddish 238
 - Yijing Hexagram Symbols 506
 - ypogegrammeni 215
 - yuan currency sign 479
- Z**
- Zapf Dingbats 504
 - zero extension relation among encodings 573
 - zero width joiner (U+200D) 243–244, 526
 - zero width no-break space (U+FEFF) 50, 63, 524
 - initial 100, 538
 - zero width non-joiner (U+200C) 243–244, 527
 - zero width space (U+200B) 524
 - for word breaks in South Asian scripts 352, 357, 369
 - zero-width space characters 525
 - ZWJ *see* zero width joiner (U+200D)
 - ZWNBSP *see* zero width no-break space (U+FEFF)
 - ZWNJ *see* zero width non-joiner (U+200C)
 - ZWSP *see* zero width space (U+200B)

