

The Unicode Standard

Version 8.0 – 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 © 1991–2015 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 8.0

Includes bibliographical references and index.

ISBN 978-1-936213-10-8 (<http://www.unicode.org/versions/Unicode8.0.0/>)

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

QA268.U545 2015

ISBN 978-1-936213-10-8

Published in Mountain View, CA

August 2015

I General Index

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

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

A

- abbreviation, Coptic 312
- abjads 258, 359
- abstract character sequences
 - definition 90
- abstract characters 29
 - definition 90
- abugidas 259, 260, 439, 597
- accent marks *see* diacritics
- accented characters
 - encoding 12
 - Latin 291
 - normalization 206
- accounting numbers, ideographic 178
- acrophonic numerals 205, 309
- Aegean numbers 340
- Africa
 - scripts of 703–723
- Afrikaans 296
- Ahom 594–595
 - reference materials 926
- Ainu 689
- Aiton 612
- Alchemical Symbols 797
 - reference materials 926
- Algonquian 728
- Ali Gali 521
- aliases
 - character name 88, 183, 850
 - property 162
 - property value 162
- allocation areas 45
- allocation of encoded characters 44–52
- Alphabetic (informative property) 189
- alphabets 258
 - European 289–336
 - mathematical 756–760
- alternate format characters (deprecated) .. 192, 822–823
- Americas
 - scripts of 725–732
- Amharic 704
- Anatolian
 - hieroglyphs 436–437
- Anatolian Hieroglyphs
 - reference materials 926
- Ancient Symbols 800
- angle brackets (U+2329 and U+232A)
 - deprecated for technical publication 784
- Annexes, Unicode Standard (UAX) xxxiii, 867
 - as components of Unicode Standard 79
 - conformance 85
 - list of 85
- annotation characters 835–837
 - use in plain text discouraged 836
- ANSI/ISO C
 - wchar_t and Unicode 200
- apostrophe (U+0027) 274
- Arabic 367–388
 - digits 763
- Arabic-Indic digits 371–372
 - signs used with 373
- ArabicShaping.txt 375, 379, 394
- Aramaic 410, 439, 521, 545, 550
- areas of the Unicode Standard 45
- ARIB 793
- Armenian 319–320
- arrows 779–780
- ASCII
 - characters with multiple semantics 264
 - transparency of UTF-8 36
 - Unicode modeled on 1
 - zero extension 200, 882
- Assamese 463
- assigned code points 11, 30
- Athapascan 728
- atomic character boundaries 218

Avestan 418
 reference materials 927

B

Balinese 647–652
 reference materials 927
 Bamum 719–720
 reference materials 927
 Bangla 463–468
 base characters 327
 definition 105
 multiple 59
 ordered before combining marks 220, 327
 Basic Multilingual Plane (BMP) 1, 44
 allocation areas 49
 representation in UTF-16 36
 Basque 296
 Bassa Vah 721
 reference materials 927
 Batak 658
 reference materials 928
 benefits of Unicode 1
 Bengali 463–468
 Bidi Class (normative property) 173
 Bidi Mirrored (normative property) 180
 Bidi Mirroring Glyph (informative property) ... 181
 BidiMirroring.txt 181
 Bidirectional Algorithm, Unicode 53, 84
 bidirectional ordering 20
 controls 819
 bidirectional text 53, 84
 Middle Eastern scripts 359
 nonspacing marks in 223
 punctuation in 263
 big-endian 40
 definition 83
 Bihari 459
 binary comparison and sort order
 caution for UTF-16 36
 UTF differences 231, 233
 UTF-8 39
 blocks of the Unicode Standard 45, 257
 Blocks.txt 45
 BMP *see* Basic Multilingual Plane
 BNF (Backus-Naur Form) 861
 BOCU-1 *see* UTN #6, BOCU-1
 MIME-Compatible Unicode Compression
 Bodhi 509
 Bodo 458
 BOM (U+FEFF) 40, 67, 130–133, 833–835
 Bopomofo 685–687

boundaries, text 61, 190, 217–218, 228
see also UAX #14, Unicode Line Breaking Algorithm
see also UAX #29, Unicode Text Segmentation
 boustrophedon 53, 349
 box drawing symbols 788
 Brahmi 439, 545, 546–549, 550, 599
 reference materials 928
 Braille 734–735
 Breton 296
 Buginese 645–646
 Buhid 642
 Bulgarian 314
 bullets 277
 numeric 764
 Burmese *see* Myanmar
 Byelorussian 314
 byte order mark (BOM) (U+FEFF) . 40, 67, 130–133, 833–835
 byte ordering
 changing 81
 conformance 83
 byte serialization 40, 67
 Byzantine Musical Symbols 741

C

C language
 wchar_t and Unicode 200
 C0 and C1 control codes 31, 188, 808
 Cambodian *see* Khmer
 Canadian Aboriginal Syllabics 728–729
 reference materials 928
 candrabindu 461, 572
 canonical composite characters
see canonical decomposable characters
 canonical composition algorithm 138
 canonical decomposable characters
 definition 117
 canonical decomposition 63
 definition 116
 mappings 115
 canonical equivalence
 definition 117
 nonspacing marks 225
 canonical equivalent character sequences
 conformance 81
 canonical mappings
see canonical decomposition mappings
 canonical ordering algorithm 137
 canonical precomposed characters
see canonical decomposable characters
 Cantonese 669
 capital letters 164, 236, 289

- Carian 343
 - reference materials 928
- carriage return (U+000D) (CR) 209, 809
- carriage return and line feed (CRLF) 209
- case 297
 - and text processes 12
 - beyond ASCII 237
 - camelcase 239
 - case folding 240
 - case operations (conformance) 85, 152–158
 - case operations and normalization 242
 - case operations, reversibility 239
 - cased (definition) 153
 - case-insensitive comparison 157, 231, 240
 - casing context (definition) 153
 - conversion 154
 - detection 156
 - European alphabets 289
 - exceptional Latin pairs 293, 297
 - Georgian 321
 - lowercase 164, 236, 289
 - mapping tables 196
 - mappings 152, 166, 236–238
 - mappings noted in code charts 851
 - titlecase 164, 236
 - Turkish I 238, 293
 - uppercase 164, 236, 289
 - see also* default case
- Case (normative property) 164, 236
- CaseFolding.txt 166, 240
- caseless letters 297
- Catalan 295
- Caucasian Albanian 354
 - reference materials 928
- cedilla 292
- CEF *see* character encoding forms
- CES *see* character encoding schemes
- CESU-8
 - see* UTR #26, Compatibility Encoding Scheme for UTF-16: 8-Bit (CESU-8)
- Chakma 540
 - reference materials 929
- Cham 636–637
 - reference materials 929
- character encoding forms (CEF) 33–39, 882
 - see also* Unicode encoding forms
- character encoding model 33, 42
 - see also* UTR #17, Unicode Character Encoding Model
- character encoding schemes (CES) 40–43
 - see also* Unicode encoding schemes
- character encoding standards
 - coverage by Unicode 3
- Character Index 873
- character literals, Unicode
 - code point notation U+ 862
- character mapping
 - interchange format *see* UTS #22, Character Mapping Markup Language (CharMapML)
- character names 88, 182–187, 886
 - aliases 88, 183, 850
 - conventions 859
 - for CJK ideographs 855
 - for control codes 186, 188
 - in code charts 846–850
 - matching 182
- character properties
 - see* properties
 - see also* individual properties, e.g. Combining Class
- character semantics 1, 80, 87–88, 887
 - as Unicode design principle 18
 - ASCII 264
 - definition 88
- character sequences
 - abstract *see* abstract character sequences
 - canonical equivalent *see* canonical equivalent character sequences
 - compatibility equivalent *see* compatibility equivalent character sequences
 - conformance 81
 - named 183
- character sequences, combining 105
- character shaping selectors (deprecated) 822
- character tabulation (U+0009) 809
- characters
 - abstract *see* abstract characters
 - arrangement in Unicode 46
 - assigned 11, 30
 - blocks 45, 257
 - boundaries 217
 - canonical decomposable *see* canonical decomposable characters
 - classes 862
 - code charts 845–858, 873
 - coded *see* encoded characters
 - combining *see* combining characters
 - compatibility decomposable *see* compatibility decomposable characters
 - composite *see* decomposable characters
 - concept of 15, 60
 - conformance definitions 90–92
 - confusable 246
 - conversion 196–197
 - decomposable *see* decomposable characters
 - deprecated *see* deprecated characters
 - encoded *see* encoded characters
 - encoding forms *see* encoding forms
 - encoding schemes *see* encoding schemes

- end-user perceived 60
- format control 30, 68, 265, 807–823
- glyphs, relationship to 15
- graphic 30
- identity (definition) 87
- ignored in processing 249–254
- interpretation 80
- layout control 68, 811–821
- modification 81
- names list 846–850
- names *see* character names
- not encoded in Unicode 3
- number encoded in Version 8.0 3
- precomposed *see* decomposable characters
- properties *see* properties
- semantics *see* character semantics
- special 67, 807–843
- supplementary *see* supplementary characters
- transcoding 196–197
- unsupported 201
- characters, not glyphs
 - in spoofing 247
 - Unicode principle 15
- CharMapML
 - see* UTS #22, Character Mapping Markup Language (CharMapML)
- charsets
 - IANA registered names 41
- charts, character code *see* code charts
- Cherokee 726
 - reference materials 929
- Chinese 668–669
 - Cantonese 669
 - Hakka 686
 - Mandarin 669
 - Minnan (Hokkien/Fujian, incl. Taiwanese) 686
 - simplified and traditional 668
- Chu hán 667
- Chu Nôm 900
- citations for
 - properties 77
 - Unicode algorithms 78
 - Unicode Standard 76
- CJK ideographs 260, 663–678
 - accounting numbers 178
 - CJK Compatibility Ideographs 677–678
 - CJK Compatibility Supplement 678
 - CJK Strokes 680, 907
 - CJK Unified Ideographs 663–677
 - CJK Unified Ideographs Extension A 665
 - CJK Unified Ideographs Extension B 677
 - CJK Unified Ideographs Extension C 677
 - CJK Unified Ideographs Extension D 677
 - CJK Unified Ideographs Extension E 677
 - code charts 855
 - compatibility ideographs in Plane 2 52
 - component structure 672
 - encoding blocks 664
 - ideographic description sequences 681–684
 - ideographic variation mark (U+303E) 683
 - KangXi radicals 676, 679–680
 - names 855
 - numbers 763
 - numeric values 178, 205
 - order of encoding 674
 - radicals 679–680
 - source standards 902–906
 - unknown or unavailable 285
 - Vietnamese 662
- CJK Miscellaneous Area 50
- CJK punctuation and symbols 284
 - compatibility forms 286
 - overscores and underscores 286
 - quotation marks 272
 - sesame dots 285
 - vertical forms 286
- CJK-JRG (Chinese/Japanese/Korean Joint Research Group) 898
- CJKV Ideographs Area 50
- CLDR (Unicode Common Locale Data Repository) 874
- cluster boundaries 217
- code charts 845–858, 873
 - representative glyphs 846
- code point sequences
 - notation 860
- code points 7, 29
 - assigned 11, 30
 - assignment 46
 - categories 30
 - default ignorable 201, 253
 - definition 90
 - designated 30
 - notation 859
 - number in Unicode Standard 1
 - private-use *see* private-use code points
 - reserved *see* reserved code points
 - semantics 32
 - surrogate *see* surrogates
 - unassigned *see* unassigned code points
 - undesignated 30
- code positions *see* code points
- code set independence 18
- code unit sequences
 - definition 119
 - ill-formed (definition) 121
 - notation 860
 - well-formed (definition) 121

- code units
 - definition 119
 - isolated 118
- code values *see* code units
- coded character representations
 - see* coded character sequences
- coded character sequences
 - definition 91
- coded characters *see* encoded characters
- codespace *see* Unicode codespace
- coeng 613, 616
- Collation Algorithm, Unicode (UCA) 12
- collation *see* sorting
- collation tables 196
- combining character sequences 56, 105
 - defective 223
 - definition 107
 - Latin 291
 - line breaking 219
 - matching 219
 - order of base character and marks 220, 327
 - rendering 219
 - selection 217
 - truncation 220–221
- combining characters 55–60, 109–114, 219–227
 - blocking reordering 818
 - canonical ordering 62, 137, 168
 - class zero 169
 - combining marks 327–328
 - definition 105
 - dependence 327
 - display order 58
 - keyboard input 220
 - ligatures 59
 - multiple 57
 - multiple base characters 59
 - normalization of 206
 - ordering conventions 56
 - rendering of marks 222–227
 - reordrant 169
 - script-specific 56
 - split 170
 - strikethrough 172
 - subjoined 172
 - typographical interaction 58, 168
 - vertical stacking 58
 - see also* diacritics
- Combining Class (normative property) 168
- combining classes 135, 168, 225–226
 - class zero characters 168
 - definition 135
- combining grapheme joiner (U+034F) 817
- combining half marks 191, 335
- combining marks *see* combining characters
- comma below 292
- Compatibility and Specials Area 26, 50
- compatibility characters 22
- compatibility composite characters 27
 - see* compatibility decomposable characters
- compatibility decomposable characters 26
 - definition 115
- compatibility decomposition 63
 - definition 115
- compatibility decomposition mappings 115
- Compatibility Encoding Scheme for UTF-16
 - see* UTR #26, Compatibility Encoding Scheme for UTF-16: 8-Bit (CESU-8)
- compatibility equivalence
 - definition 116
- compatibility equivalent character sequences
 - conformance 81
- compatibility mappings
 - see* compatibility decomposition mappings
- compatibility precomposed characters
 - see* compatibility decomposable characters
- compatibility variants 26
 - mapping 244
- composite characters
 - see* decomposable characters
- Composition Exclusion (normative property) 99
- compression 208
 - see also* UTS #6, A Standard Compression Scheme for Unicode (SCSU)
- conferences 873
- conformance 73–158
 - clause and definition updates 895
 - definitions 87–92
 - examples 69
 - ISO/IEC 10646 implementations 887
 - requirements 79–84
- confusables 246
- conjunct consonants
 - Indic 217, 445
 - Myanmar 607
 - selection of clusters 217
- contextual shaping
 - apostrophe 274
 - Arabic 367
 - not used for Hebrew final forms 362
 - quotation marks 270
 - Syriac 393
- contour tones 325
- control codes 31, 68, 808
 - graphics for 783
 - names 188
 - properties 809
 - semantics 32, 809
 - specified in Unicode 809

control sequences	808
conversion of characters	127, 196–197, 255
convertibility	
as Unicode design principle	23
Coptic	307, 311–313
reference materials	929
Coptic Epact numbers	767
corporate use subarea	828
corrigenda	76
CR (U+000D carriage return)	209, 809
CRLF (carriage return and line feed)	209
Croatian	296
digraphs	296
culturally expected sorting	12, 230
Cuneiform	
Old Persian	429
Sumero-Akkadian	424–427
Ugaritic	428
Cuneiform and Hieroglyphic Area	51
Cuneiform and Hieroglyphs	423–437
currency symbols block	751–753
currency symbols encoded in other blocks	752
currency symbols, other	753
dollar sign, form and usage	752
euro sign	753
lari sign	753
lira sign, compatibility usage	751
lira sign, Turkish	753
peso signs, usage	752
ruble sign	753
rupee signs, Indian, usage	753
yen and yuan signs, usage	752
cursive joining	813–817
Arabic	374–381
control characters for	191, 369–370, 524, 812
Mandaic	401
Mongolian	523–525
N’Ko	715
Phags-pa	561
Syriac	393–396
transparency	816
cursive scripts	359
Cypriot	342
reference materials	936
<i>see also</i> Linear B	
Cyrillic	314–317
Czech	296

D

danda, in Devanagari block	457
Danish	295
dashes	267

Database, Unicode Character	
<i>see</i> Unicode Character Database (UCD)	
dead consonants, Indic	444
dead keys	220
decomposable characters	63
definition	115
normalization of	206
decomposition	63, 115–117
canonical <i>see</i> canonical decomposition	
compatibility <i>see</i> compatibility decomposition	
definition	115
in normalization	206
mapping, definition	115
mappings noted in code charts	852
default case	
algorithms	85, 152–158
conversion	154
detection	156
folding	155
default caseless matching	157
default grapheme clusters	217
<i>see also</i> UAX #29, Unicode Text Segmentation	
Default Ignorable Code Point (property)	253
default ignorable code points	201, 253
default property values	
definition	96
defective combining character sequences	223
definition	107
dependent vowel signs	
Indic	443
Khmer	618
Philippine scripts	642
deprecated characters	74, 849
alternate format	192, 822–823
definition	91
Derived Age (property)	201
derived properties	
definition	103
DerivedCoreProperties.txt	153, 164, 253
DerivedNormalizationProps.txt	243
Deseret	730–732
reference materials	930
design goals of Unicode	4
design principles of Unicode	14–24
designated code points	30
Devanagari	441–462
Dhivehi	505
diacritics	55, 327
alternative glyphs	291, 327
Czech	292
display in isolation	60, 267, 328
double	113, 191, 329
German dialectology	333
Greek	304–305, 308

Latin	291–294
Latvian	292
mathematical	759
on i and j	293
rendering	222–227
Slovak	292
spacing clones of	325, 329
symbol	55, 334
<i>see also</i> combining characters	
dictionary symbols	793
digit form names	371
digits	205
Arabic	763
Arabic-Indic	371–372
compatibility	763
decimal	177
glyph variants	765
hexadecimal	763
Myanmar	763
national shapes	823
Shan	763
superscript and subscript	764
Tai Laing	763
Tai Tham	763
digraphs	296, 299, 301
dingbats	795–797
directionality	20, 53
East Asian scripts	662
Middle Eastern scripts	359
Mongolian	522
musical symbols	737
normative property	173
Ogham	356
Old Italic	346
Philippine scripts	643
Runic	349
discussion list for Unicode	873
Dogri	458
Domino Tiles	798
dotless i	238, 293
dotted circle	
in code charts	106, 328
in fallback rendering	222
to indicate diacritic	55
to indicate vowel sign placement	56
double diacritics	113, 191, 329
Duployan	745–746
reference materials	930
Dutch	295, 296
dynamic composition	
as Unicode design principle	22
Dzongkha	509

E

East Asian scripts	661–702
writing direction	53
<i>see also</i> CJK ideographs	
Eastern Arabic-Indic digits	371
EBCDIC	
newline function	210
<i>see</i> UTR #16, UTF-EBCDIC	
editing, text boundaries for	217–218
efficiency	
as Unicode design principle	15
Egyptian hieroglyphs	430–433
reference materials	930
Elbasan	353
reference materials	930
ellipsis	275–276
e-mail discussion list for Unicode	873
emoji	791
animal symbols	794
cultural symbols	794
zodiacal symbols	794
emoji modifiers	795
emoticons	795
Enclosed Alphanumerics	804
enclosing marks	334
definition	106
encoded characters	7, 29
allocation	44–52
definition	90
encoding form conversion	
definition	126
encoding forms	33–39
ISO/IEC 10646 definitions	882
encoding forms, Unicode	
<i>see</i> Unicode encoding forms	
encoding model for Unicode characters	33, 42
<i>see also</i> UTR #17, Unicode Character Encoding Model	
encoding schemes	40–43
encoding schemes, Unicode	
<i>see</i> Unicode encoding schemes	
endian ordering	
<i>see</i> byte order mark (BOM) (U+FEFF)	
end-user subarea	829
English	295
equivalent sequences	206
as Unicode design principle	23
case-insensitivity	231, 240
combining characters in matching	219
conformance	82
Hangul syllables	693
in sorting and searching	230
language-specific	117

- security implications 246
see also canonical equivalence
see also compatibility equivalence
see also encoding forms, encoding schemes
- errata xxxvi, 76, 874
- escape sequences 809
 not used in Unicode 1, 4
- Esperanto 296
- Estonian 296
- Ethiopic 704–707
 reference materials 931
- Etruscan 345
- European scripts 289–336
 ancient 337–357
- eyelash-RA 450
- F**
- fallback rendering 253
 of nonspacing marks 222
- FAQ (Frequently Asked Questions) 873
- Faroese 295
- Farsi 367, 370
- featural syllabaries 259
- FF (U+000C form feed) 209, 809
- file separator (U+001C) 809
- Finnish 295
- Finno-Ugric Transcription (FUT)
see Uralic Phonetic Alphabet (UPA)
- fixed-width Unicode encoding form (UTF-32) ... 35,
 123
- flat tables 196
- Flemish 295
- fleurons 797
- fonts
 and Unicode characters 16
 for mathematical alphabets 758–760
 style variation for symbols 749
- form feed (U+000C) (FF) 209, 809
- format control characters 30, 68, 265, 807–823
 deprecated 822–823
 prefixed 192
 stateful 820
- fraction characters 774
- fraction slash (U+2044) 275, 770
- French 296
- Frisian 296
- FTP site, Unicode Consortium 873
- fullwidth forms in East Asian encodings 690
- futhark 348
- G**
- Garshuni 389
- Ge'ez 704
- General Category (normative property) 174
 list of values 174
- general punctuation 263–287
- General Scripts Area 50
- geometrical symbols 788–790
- Georgian 321–322
- German 295
- geta mark (U+3013) 285
- Glagolitic 318
 reference materials 931
- Glossary 873
- glyph selection tables 196
- glyphs 6, 15
 characters, relationship to 15
 diacritics alternative 291, 327
 Greek alternative 305–307
 Latin alternative 291
 mathematical alternative 775
 missing 253
 representative in code charts 846
 standardized variants 824
 symbols alternative 749
- golden numbers 350
- Gothic 352
 reference materials 931
- Grantha 591–593
 reference materials 931
- grapheme base 327
 definition 107
- grapheme clusters 11, 60–61
see also UAX #29, Unicode Text Segmentation
 default 217
 definition 108
- grapheme extender
 definition 108
- grapheme joiner, combining (U+034F) 817
- graphic characters 30
- Greek 304–309
 acrophonic numerals 205, 309
 alternative glyphs 305–307
 ancient musical notation 742–744
 editorial marks 280
 letters as symbols 305–307, 777
see also Cypriot, Linear B
- Greek editorial marks
 reference materials 931
- Greenlandic 296
- group separator (U+001D) 809
- guillemets 270
- Gujarati 474
- Gurmukhi 469–473

H

- Hakka 686
 - halant 439
 - see also* virama
 - half marks, combining 191, 335
 - half-consonants, Indic 446
 - halfwidth forms in East Asian encodings 690
 - Han ideographs *see* CJK ideographs
 - Han unification 670–677
 - and language tags 215
 - history 897–906
 - language usage 667
 - source separation rule 665, 671
 - source standards 902–906
 - hand symbols 794
 - Hangul Area 50
 - Hangul syllables 661, 691–694
 - and combining marks 113
 - as grapheme clusters 61
 - canonical decomposition 144
 - collation 693
 - composition 146
 - conjoining jamo 142–151
 - equivalent sequences 693
 - Hangul Compatibility Jamo 692
 - Hangul Jamo 691–694
 - Hangul Syllables block 693–694
 - Johab set 693
 - name generation 147
 - normalization 692
 - standard 143
 - Hangzhou numerals 769
 - Hanja *see* CJK ideographs
 - Hanunóo 642
 - Hanzi *see* CJK ideographs
 - harakat 368
 - hasant 463
 - hash tables 197
 - Hatran 422
 - reference materials 932
 - Hebrew 361–366
 - hentaigana 689
 - hieroglyphs
 - Anatolian 436–437
 - Egyptian 430–433
 - Meroitic 434–435
 - high surrogate
 - definition 118
 - high-surrogate code points 79, 830
 - high-surrogate code units 118
 - higher-level protocols
 - definition 92
 - Hindi 441
 - Hiragana 688
 - horizontal tab (U+0009) 809
 - HTML newline function 210
 - Hungarian 296
 - hyphenation 812
 - as a text process 10
 - hyphens 267, 812
- I**
- I Ching symbols 799
 - IANA charset names 41
 - Icelandic 295
 - identifiers 229
 - see also* UAX #31, Unicode Identifier and Pattern Syntax
 - Ideographic (informative property) 189
 - ideographic description sequences 682
 - Ideographic Rapporteur Group (IRG) 900
 - Ideographic Variation Database *see* UTS #37, Unicode Ideographic Variation Database
 - ideographs *see also* CJK ideographs
 - IDNA *see* UTS #46, Unicode IDNA Compatibility Processing
 - IICore 665, 900
 - ill-formed
 - definition 121
 - Imperial Aramaic 410–411
 - reference materials 932
 - implementation guidelines 195–255
 - in a Unicode encoding form
 - definition 122
 - in-band mechanisms 842
 - India
 - Official scripts 439–501
 - Indian rupee signs, usage 753
 - Indic scripts 439–501
 - principles, in terms of Devanagari 442–449
 - relation to ISCII standard 441
 - Indonesia and Oceania
 - scripts of 641–660
 - Indonesian 295
 - industry character sets
 - covered in Unicode 3
 - information separators (U+001C..U+001F) 809
 - informative properties
 - definition 99
 - Inscriptional Pahlavi 416
 - Inscriptional Parthian 416
 - inside-out rule 222
 - interchange restrictions 31
 - International Phonetic Alphabet (IPA) 258, 298–299
 - reference materials 933

Spacing Modifier Letters	324
<i>see also</i> phonetic alphabets	
internationalization	18
Internationalization & Unicode Conference	873
Internet protocols	
UTF-8 as preferred encoding	37
Inuktitut	728
invisible operators	782
iota subscript	305
IPA <i>see</i> International Phonetic Alphabet	
IRG (Ideographic Rapporteur Group)	900
Irish	295, 356
ISCI standard and Unicode	441
ISO/IEC 10646	875–887
conformance of Unicode implementations ..	887
encoding forms	882
synchrony with Unicode Standard	884
timeline compared to Unicode versions	877
Italian	295
ITC Zapf Dingbats	795
IUC <i>see</i> Internationalization & Unicode Conference	

J

jamos <i>see</i> Hangul syllables	
Japanese	661
Javanese	653–656
reference materials	933
Jawi	385
jihvamuliya	462, 572
Johab	693
joiners	369
combining grapheme joiner (U+034F)	817
word joiner (U+2060)	811
zero width joiner (U+200D)	369–370, 814
justification	224

K

Kaithi	569–571
reference materials	933
Kana (Hiragana and Katakana)	688–689
Kanbun	678
KangXi radicals	676, 679–680
Kanji <i>see</i> CJK ideographs	
Kannada	491–494
Kashmiri	459
Katakana	688–689
Kawi	647, 649
Kayah Li	635
reference materials	934
KC (normalization form)	
<i>see</i> Normalization Form KC	
KD (normalization form)	
<i>see</i> Normalization Form KD	

keytop labels	783
Khamti Shan	610
Kharoshthi	550–551
reference materials	934
Khmer	613–623
characters not recommended	620
syllable components, order of	621
Khojki	580–581
reference materials	934
Khudawadi	582–583
reference materials	935
killer	260
Batak	658
Brahmi	546
Meetei Mayek	534
Myanmar (asat)	608
<i>see also</i> virama	
Konkani	458
Korean Hangul <i>see</i> Hangul	
Kurdish	385

L

Ladino	361
language tags	215, 838–843
and Han unification	215
use strongly discouraged	838, 842
Lanna	626
Lao	603–605
last-resort glyphs	253
Latin	291–303
alternative glyphs	291
Basic Latin	295
encoding blocks	45
IPA Extensions	298–299
Latin Extended Additional	301–303
Latin Extended-A	295
Latin Extended-B	296–298
Latin Extended-C	301
Latin Extended-D	302
Latin Extended-E	303
Latin Ligatures	301
Latin-1 Supplement	295
Phonetic Extensions	300–303
Latvian	296, 303
cedilla	292
layout control characters	68, 811–821
leading surrogates	
<i>see</i> high-surrogate code units	
legibility criterion for plain text	19
Lepcha	541–543
reference materials	935
letter spacing	812
letterlike symbols	754–760

- LF (U+000A line feed) 209, 809
- ligatures 813–817
- Arabic 377–379
 - combining characters on 59
 - control characters for 191
 - for nonspacing marks 226
 - Latin 301
 - selection 218
 - Syriac 396
- Limbu 530–533
- reference materials 935
- line breaking 209–213, 811–813
- control characters 193
 - in South Asian scripts 601, 609, 623
 - recommendations 211
 - see also* UAX #14, Unicode Line Breaking Algorithm
- line feed (U+000A) (LF) 209, 809
- line separator (U+2028) (LS) 209, 813
- line tabulation (U+000B) (VT) 809
- Linear A 339
- reference materials 936
- Linear B 340–341
- reference materials 936
 - see also* Cypriot
- linear boundaries 218
- Lisu 698–700
- reference materials 937
- Lithuanian 296
- little-endian 40
- definition 83
- Locale Data Markup Language
- see* UTS #35, Unicode Locale Data Markup Language (LDML)
- logical order
- as Unicode design principle 19
 - exceptions to 170
- logograph 260
- logosyllabaries 260
- low surrogate
- definition 118
 - low-surrogate code points 79, 830
 - low-surrogate code units 118
- lowercase 164, 236, 289
- LS (U+2028 line separator) 209, 813
- Lycian 343
- reference materials 937
- Lydian 343
- reference materials 937
- M**
- MacOS newline function 210
- Mahajani 578–579
- reference materials 938
- Mahjong Tiles 798
- mail discussion list for Unicode 873
- Maithili 458
- major version 75
- Malay 295
- Malay, Patani 602
- Malayalam 495–501
- Maltese 296
- Manchu 522
- Mandaic 400–401
- reference materials 938
- Mandarin 669
- Manden 712
- Manichaean 412–415
- reference materials 938
- map symbols 793
- mapping tables *see* tables of character data
- Marathi 441, 450, 456
- markup languages
- and Unicode conformance 842
 - line breaking 209
 - see also* UTR #20, Unicode in XML and Other Markup Languages
- Mathematical (informative property) 774
- mathematical expression format characters 191
- see also* UTR #25, Unicode Support for Mathematics
- mathematical symbols 774–781
- alphabets 756–760
 - alphanumeric 755–760
 - fonts 758–760
 - format characters 782
 - fragments for typesetting 784
 - invisible operators 782
 - operators 775–778
 - reference materials 938
 - standardized variants 781
- MathML 778
- matras 168, 443
- Meetei Mayek 534–535
- reference materials 938
- Mende Kikakui 722–723
- reference materials 939
- Meroitic
- cursive 434–435
 - hieroglyphs 434–435
 - reference materials 939
- Miao 701–702
- reference materials 939
- Middle Eastern scripts 359–506
- ancient 403–422
- Min 669

- Minnan (Hokkien/Fujian, incl. Taiwanese) 686
 minor version 75
 minus sign 777
 commercial (U+2052) 278
 mirrored property
 see Bidi Mirrored (normative property)
 mirroring of paired punctuation 269
 Miscellaneous Symbols 792
 missing glyphs 253
 Modi 588–590
 reference materials 939
 modifier letters 323–326
 Modifier Letters, Spacing 300
 Mongolian 521–529, 556
 writing direction 522
 Mro 536
 reference materials 940
 Multani 584
 reference materials 940
 multibyte encodings
 compared to UTF-8 37
 multistage tables 196
 musical symbols 736–744
 ancient Greek 742–744
 Balinese 651
 Byzantine 741
 directionality 737
 Gregorian 739
 Kievan 740
 reference materials 940
 Western 736–740
 Myanmar 606–612
 digits 763
 Myanmar Extended-A 610
 Myanmar Extended-B 610
 reference materials 941
- ## N
- N’Ko 712–716
 reference materials 941
 Nabataean 420
 reference materials 941
 named character sequences 183
 names, character *see* character names
 namespace 89
 NEL (U+0085 next line) 209, 809
 Nepali 441
 neutral directional characters 173
 New Tai Lue 626–628
 newline function (NLF) 210, 810
 newline guidelines 209–213
 next line (U+0085) (NEL) 209, 809
 NFC (Normalization Form C) 62
 NFD (Normalization Form D) 62
 NFKC (Normalization Form KC) 62
 NFKD (Normalization Form KD) 62
 NLF (newline function) 210, 810
 no-break space (U+00A0) 811
 base for diacritic in isolation 60, 267, 328
 no-break space, narrow (U+202F) 527
 noncharacter code points *see* noncharacters
 noncharacters 31, 831
 conformance 79
 definition 92
 handling 82
 in code charts 849
 interchange restrictions 31
 semantics 32
 U+10FFFF (not a character code) 831
 U+FDD0..U+FDEF 31, 831
 U+FFFE (not a character code) 67, 832
 U+FFFF (not a character code) 31, 831
 nondecomposable characters 64
 non-joiner, zero width (U+200C) 369–370, 815
 nonlinear boundaries 218
 non-overlap principle in Unicode encoding forms 33
 nonspacing marks 327
 definition 106
 display in isolation (U+202F) 60, 267, 328
 positioning 226
 rendering 222–227
 see also combining characters
 see also diacritics
 normalization 62, 206–207
 and case operations 242
 canonical ordering algorithm 62, 137, 168
 conformance 84
 of private-use characters 828
 see also UAX #15, Unicode Normalization Forms
 stability 134
 Normalization Form C (NFC) 62
 Normalization Form D (NFD) 62
 Normalization Form KC (NFKC) 62
 Normalization Form KD (NFKD) 62
 normalization forms 134–141
 definition 140
 specification 136
 normative behaviors
 definition 87
 normative properties
 definition 98
 list 99
 may change 98
 Norwegian 295
 notational conventions 859–863
 notational systems 261, 733–748
 nukta 368, 387, 451

null (U+0000)	
as Unicode string terminator	810
number forms	
CJK ideographs	205
numbers	
Coptic Epact	767
handling	205
ideographic accounting	178
numerals	761–771
acrophonic	309
Chinese counting rods	772
Coptic	313
Cuneiform	427
Ethiopic	706
Greek acrophonic	205
Hangzhou	769
Meroitic cursive	435
old-style	275
Roman	205, 774
Rumi	768
Suzhou-style	769
numeric separators	278
numeric shape selectors (deprecated)	823
Numeric Type (normative property)	177
Numeric Value (normative property)	177
numero sign (U+2116)	754

O

object replacement character (U+FFFC)	837
octet	861
Ogham	356
reference materials	942
Ol Chiki	538–539
reference materials	942
Old Church Slavonic	314
Old Hungarian	351
Old Italic	345–347
reference materials	942
Old North Arabian	405
reference materials	942
Old Permic	355
reference materials	942
Old Persian	429
reference materials	943
Old South Arabian	406–407
reference materials	943
Old Turkic	563
reference materials	943
old-style numerals	275
Oriya	476–478
ornamental dingbats	797
Oromo	704

Osmanya	708
reference materials	943
out-of-band mechanisms	842
overlapping encodings	33
overscores	275

P

Pahawh Hmong	638–639
reference materials	944
Pahlavi, Inscriptional	416
reference materials	932
Pahlavi, Psalter	417
Palmyrene	421
reference materials	944
Panjabi	469
paragraph or section marks	278
paragraph separator (U+2029) (PS)	209, 813
Parthian, Inscriptional	416
reference materials	932
Pashto	367
Patani Malay	602
Pau Cin Hau	640
reference materials	944
Persian	367, 370
Phags-pa	556–562
reference materials	944
Phaistos Disc symbols	801
Phake	612
Philippine scripts	642–644
reference materials	945
Phoenician	408
reference materials	946
phonemes	261
phonetic alphabets	258
IPA Extensions	298–299
Phonetic Extensions	300–303
Spacing Modifier Letters	324–326
Uralic Phonetic Alphabet (UPA)	278, 300
<i>see also</i> International Phonetic Alphabet (IPA)	
Pinyin	295
pivot code, Unicode as	196
plain text	
as Unicode design principle	18
legibility criterion	19
planes of Unicode codespace	44
Plane 0 (BMP)	44
Plane 1 (SMP)	44, 51
Plane 14 (SSP)	45
Plane 2 (SIP)	44, 52
Planes 15–16 (Private Use)	52, 829
Playing Cards	798
points, Hebrew pronunciation marks	361
policies of the Unicode Consortium	874

- Polish 296
- Portuguese 295
- precomposed characters
see decomposable characters
 compatibility *see* compatibility decomposable
 characters
- prefixed format control characters 192
- Private Use Area (PUA) 50, 828
- Private Use planes 45, 52, 829
- private-use characters
 properties 827
 semantics 32
- private-use code points 31, 201
 conformance 80
 definition 104
 high surrogates 830
- processing code, Unicode as 38
- properties 18, 94–104, 159–193
 aliases 162
 aliases (definition) 103
 and Unicode algorithms 98
 data tables 196
 derived *see* derived properties
 in Unicode Character Database (UCD) 46
 informative *see* informative properties
 normative references to 77, 84
 normative *see* normative properties
 of control codes 809
 provisional *see* provisional properties
 simple *see* simple properties
see also individual properties, e.g. combining
 classes
- property values
 aliases 162
 aliases (definition) 104
 default 96
 default (definition) 96
 normative references to 84
- PropertyAliases.txt 103, 862
- PropertyValueAliases.txt 104, 862
- PropList.txt 166
- Provençal 296
- provisional properties
 definition 100
- PS (U+2029 paragraph separator) 209, 813
- Psalter Pahlavi 417
 reference materials 946
- PUA (Private Use Area) 50, 828
- pulli* 479
- punctuation 263–287
 blocks containing 257
 CJK 284
 doubled 275
 in bidirectional text 263
 paired 269
 small form variants 286
 typographic forms 263
 vertical forms 286
- Punctuation and Symbols Area 50
- Punjabi 469
- ## Q
- quotation marks 270–273
 East Asian 272
 European 270
- ## R
- radicals, KangXi and other CJK 679–680
- radical-stroke index 676
- record separator (U+001E) 809
- recycling symbols 793–794
- referencing 84
 properties 77
 Unicode algorithms 78
 Unicode Standard 76
- regional indicator symbols 805
- regular expressions 214
 and line breaking 209
see also UTS #18, Unicode Regular Expressions
- Rejang 657
 reference materials 946
- rendering of text 6, 10, 17
 fallback 253
 unsupported characters 201
- repertoire of abstract characters 29
- replacement character (U+FFFD) ... 43, 68, 83, 127,
 255, 837
- reserved code points 30, 201
 definition 92
 in code charts 849
 preservation in interchange 31
see also unassigned code points
- Rhaeto-Romanic 296
- rich text 18
- right single quotation mark (U+2019)
 preferred for apostrophe 274
- right-to-left text 53
 East Asian scripts 662
 Middle Eastern scripts 359
- roadmap for script additions 46
- Roman numerals 205, 774
- Romanian 296
 comma below 293
- Romany 296
- Rong 541–543
- Rumi numeral forms 768

- Runic 348–350
 reference materials 946
- Russian 314
- S**
- Samaritan 398–399
 reference materials 947
- Sami 296
- Sanskrit 441
- Saurashtra 544
 reference materials 947
- scalar values, Unicode
see Unicode scalar values
- scripts
 in Unicode Standard 3
 roadmap for future additions 46
 types of 262
see also UAX #24, Unicode Script Property
- SCSU
see UTS #6, A Standard Compression Scheme for Unicode
- searching 230–232
 as a text process 10
 case-insensitive 231, 240
- section or paragraph marks 278
- security issues 246
- self-synchronization of encoding forms 34
- semantics
see character semantics
- sequences
 notation 860
- Serbian
 corresponding digraphs in Croatian 296
- Shan 624
 digits 763
- Sharada 572–573
 reference materials 947
- Shavian 357, 698
 reference materials 947
- Show Hidden 81, 222, 253, 825
- SHY (U+00AD soft hyphen) 812
- Sibe 522
- Siddham 576–577
 reference materials 948
- signature for Unicode data 67, 833–835
- simple properties
 definition 103
- simplified Chinese 668
- Sindhi 367, 458
- Sinhala 507–508
 reference materials 948
- Sinological dot 302
- SIP (Supplementary Ideographic Plane) 44, 52
- slash, fraction (U+2044) 275
- Slovak 296
- Slovenian 296
- small letters 164, 236, 289
- SMP (Supplementary Multilingual Plane) 44, 51
- soft hyphen (U+00AD) (SHY) 812
- Somali 708
- Sora Sompeng 596
 reference materials 948
- Sorbian 296
- sorting 12, 230
 and combining grapheme joiner 818
 as a text process 10
 case-insensitive 231
 culturally expected 12, 230
 language-insensitive 230
see also Unicode Collation Algorithm (UCA)
- source separation rule 665, 671
- South and Central Asian scripts
 Ancient 545–563
 Other historic 565–596
 Other modern 503–544
- South Asian scripts 439–533
- Southeast Asian scripts 597–640
- space (U+0020)
 base for diacritic in isolation 60, 267, 328
- space characters 266, 811–813
 graphics for 783
- space, zero width (U+200B) 266
- spacing clones of diacritics 325, 329
- spacing marks 327
 definition 107
- Spacing Modifier Letters 324–326
- Spanish 295
- special characters 67, 807–843
- SpecialCasing.txt 152, 166
- Specials 833–837
- spell-checking
 as a text process 11
- spellings, alternative
see equivalent sequences
- spoofing 246
- SSP (Supplementary Special-purpose Plane) 45
- stability 101, 161
 as Unicode design principle 23
- stacked boundaries 217
- stacking sequences 57
 nondefault 58
- Standard Compression Scheme for Unicode (SCSU)
see UTS #6, A Standard Compression Scheme for Unicode
- standardized variants 525, 824
 in the code charts 853
 mathematical symbols 781

StandardizedVariants.txt	525, 781
standards coverage	3
starters	136
stateful encoding	
not used in Unicode	4
paired format controls	820
string comparison	12
string literals, Unicode	
code point notation <code>\u1234</code>	862
strings, Unicode	43, 120
null termination	810
strong directional characters	173
styled text	18
sublinear searching	231
subsets, supported	71
conformance	80
ISO/IEC 10646 specification for	885
substitution character	
<i>see</i> replacement character	
Sumero-Akkadian	424–427
Sumero-Akkadian Cuneiform	
reference materials	948
Sundanese	659–660
reference materials	949
superscripts	325
and subscripts	772
supplementary characters	
in UTF-16 strings	43
tables for	197
Supplementary General Scripts Area	50
Supplementary Ideographic Plane (SIP)	44, 52
Supplementary Multilingual Plane (SMP)	44, 51
supplementary planes	
representation in UTF-16	36
representation in UTF-8	37
Supplementary Private Use Areas	52, 829
Supplementary Special-purpose Plane (SSP)	45
supported subsets	71
conformance	80
supralineation	312
surrogate code points	
<i>see</i> surrogates	
surrogate pairs	36, 124
definition	118
processing	38, 203–204
surrogates	31, 118, 830
interchange restrictions	31
isolated surrogates, handling	43
isolated surrogates, ill-formed	124
isolated surrogates, uninterpreted	118
support levels	203
Surrogates Area	50, 830
Sutton SignWriting	747–748
reference materials	949
Suzhou-style numerals	769
svasti signs	516
Swahili	295
Swedish	295
syllabaries	259
alphabetic property	189
featural	259
Syloti Nagri	567–568
symbols	749–806
animal	794
appearance variation	749
arrows	779–780
box drawing	788
cultural	794
currency symbols block	751–753
dictionary	793
dingbats	795–797
emoji	791, 805
Enclosed Alphanumerics	804
fragments for mathematical typesetting	784
game	794
gender	794
genealogical	794
geometrical	788–790
hand	794
Khmer lunar calendar	623
letterlike	754–760
map	793
mathematical	774–781
mathematical alphanumeric	755–760
miscellaneous	792
musical	736–744
numerals	761–771
recycling	793–794
regional indicator	805
technical	783–787
weather	793
zodiacal	794
symmetric swapping format characters	822
Syriac	389–396
reference materials	950

T

tab (U+0009 character tabulation)	809
tab, vertical (U+000B)	209, 809
tables of character data	196–197
optimization	197
supplementary characters	197
tag characters	838–843
Tagalog	642
Tagbanwa	642
tags, language	215, 838–843
use strongly discouraged	842

- Tai Laing
 digits 763
- Tai Le 624–625
 reference materials 950
- Tai Tham 629–631
 digits 763
 reference materials 950
- Tai Viet 632–634
- Tai Xuan Jing symbols 800
- Takri 574–575
 reference materials 950
- Tamil 479–487
- tashkil 368
- tashkil, harakat, points 370
- TCHAR in Win32 API 200
- Technical Notes (UTN) 872
- Technical Reports (UTR) 867
 abstracts 870
- Technical Standards (UTS) xxxv, 867
 abstracts 868
- technical symbols 783–787
- Telugu 488–490
- terminal emulation 750
- text boundaries 61, 190, 217–218, 228
see also UAX #14, Unicode Line Breaking Algorithm
see also UAX #29, Unicode Text Boundaries
- text elements 6, 10, 217
 boundaries 228
 for sorting 230
 variable-width nature 38
- text processes 6, 10–13
- text rendering 6, 10, 17
- text selection, boundaries for 217–218
- Thaana 505–506
 reference materials 950
- Thai 599–602
- Tibetan 509–520
- Tifinagh 709
- Tigre 704
- tilde (U+007E) 278
- Tirhuta 585–587
 reference materials 951
- titlecase 164, 236
- Todo 522
- tone letters 325–326
- tone marks
 Bopomofo spacing 685, 686
 Chinantec 326
 Chinese 326
 Tai Le 624
 Thai 599
 Vietnamese 294
- traditional Chinese 668
- traffic signs 793
- trailing surrogates
see low-surrogate code units
- transcoding 196–197
 tables 196
- Transport and Map Symbols 795
- triangulation in transcoding 196
- tries 196
- truncation
 combining character sequences 220–221
 surrogates and 204
- Turkish 296
 case mapping of I 238, 293
 cedilla 293
 lira sign 753
- two-stage tables 197
- ## U
- U+ notation 862
- U+10FFFF (not a character code) 831
- U+FEFF (BOM) 833–835
- U+FFFE (not a character code) 832
- U+FFFF (not a character code) 831
- UAX (Unicode Standard Annex) xxxiii, 867
 as component of Unicode Standard 79
 conformance 85
 list of 85
- UCA *see* Unicode Collation Algorithm
- UCD *see* Unicode Character Database
- UCS (Universal Character Set)
see ISO/IEC 10646
- UCS-2 882
- UCS-4 882
- Ugaritic 428
 reference materials 951
- Uighur 521, 556
- Ukrainian 314
- unassigned code points 30, 79, 201
 defined as reserved code points 92
 handling 74
 properties of 96
 semantics 79
see also reserved code points
- underscores 275
- undesignated code points 30
- Unicode 1.0 Name (informative property) 188
- Unicode algorithms
 and properties 98
 conformance 84
 definition 92
 normative references to 78, 84
- Unicode Bidirectional Algorithm 20, 53
see also UAX #9, Unicode Bidirectional Algorithm

- Unicode Character Database (UCD) . xxxv, 161, 873
 - as component of Unicode Standard 79
 - changes 74
 - properties in 46
- Unicode character encoding model 33, 42
 - see also* UTR #17, Unicode Character Encoding Model
- Unicode character literals
 - code point notation U+ 862
- Unicode codespace
 - allocation numbers 890
 - definition 90
 - planes 44
 - size 1, 29
- Unicode Collation Algorithm (UCA) 12
 - see also* UTS #10, Unicode Collation Algorithm
- Unicode Common Locale Data Repository (CLDR) 874
- Unicode conferences 873
- Unicode Consortium 866
 - addresses 874
 - Consortium membership in standards bodies . 866
 - e-mail discussion list 873
 - FTP site 873
 - membership 866
 - policies 874
 - website 873
- Unicode data signature 67, 833–835
- Unicode data types 199–200
 - for C 199–200
- Unicode encoding forms 119–126
 - advantages of each 38
 - conformance 34, 82
 - definition 120
 - fixed-width (UTF-32) 35, 123
 - signatures 834, 835
 - variable-width 36, 124
 - see also* encoding forms
- Unicode encoding schemes
 - conformance 130–133
 - definition 130
 - endian ordering 40
 - see also* encoding schemes
- Unicode escape sequence notation \u1234 862
- Unicode Regular Expressions *see* UTS #18, Unicode Regular Expressions
- Unicode scalar values
 - definition 119
- Unicode security 246
 - see also* UTS #39, Unicode Security Mechanisms
- Unicode Standard
 - allocation of encoded characters 44–52
 - architecture 10–13
 - areas 45
 - benefits 1
 - blocks 45, 257
 - code charts 845–858, 873
 - components 79
 - conformance 73–158
 - conformance of ISO/IEC 10646 implementations . 87
 - corrections 76
 - definitions for conformance 87–92
 - design goals 4
 - design principles 14–24
 - errata 76, 874
 - normative references to 76, 84
 - number of characters 3
 - number of code points 1, 29
 - script coverage 3
 - security issues 246
 - synchrony with ISO/IEC 10646 884
 - updates 874
 - versions *see* versions of the Unicode Standard
 - see also* Version 8.0
- Unicode Standard Annexes (UAX) xxxiii, 867
 - as components of Unicode Standard 79
 - conformance 85
 - list of 85
- Unicode string literals
 - code point notation \u1234 862
- Unicode strings 43
 - definition 120
- Unicode Technical Committee (UTC) 866
- Unicode Technical Notes (UTN) 872
- Unicode Technical Reports (UTR) 867
 - abstracts 870
- Unicode Technical Standards (UTS) xxxv, 867
 - abstracts 868
- UnicodeData.txt 152, 166
- unification
 - as Unicode design principle 21
 - see also* Han unification
- Unified Repertoire and Ordering (URO) . . . 671, 899
 - see also* Han unification
- Unihan Database 161, 675, 676, 855, 874, 900
- Unihan.zip 101, 161
- unit separator (U+001F) 809
- Universal Character Set (UCS)
 - see* ISO/IEC 10646
- universality
 - as Unicode design principle 14
- Unix
 - and UTFs 38
 - newline function 210
 - UTF-32 in 35
 - UTF-8 in 18
- unsupported characters 201

- upadhamaniya 462, 572
 - update version 75
 - uppercase 164, 236, 289
 - Uralic Phonetic Alphabet (UPA) 278, 300
 - Urdu 367
 - URO (Unified Repertoire and Ordering) ... 671, 899
see also Han unification
 - UTF, Unicode Transformation Formats 33, 120
 - advantages of each 38
 - as encoding form or scheme 133
 - binary comparison and sort order differences ..
 231, 233
 - in APIs 200
 - UTF-16 36, 124, 883
 - binary comparison and sort order caution ... 36
 - bit distribution (table) 124
 - BOM in 131, 833
 - encoding form (definition) 124
 - encoding scheme (definition) 131
 - encoding schemes 40
 - in ISO/IEC 10646 883
 - in UTF-8 order 234
 - surrogates and string handling 43, 203
 - UTF-16BE (Big-endian) 834
 - encoding scheme 41
 - encoding scheme (definition) 130
 - UTF-16LE (Little-endian) 834
 - encoding scheme 41
 - encoding scheme (definition) 130
 - UTF-32 35, 123
 - as processing code 38
 - BOM in 132
 - encoding form (definition) 123
 - encoding scheme (definition) 132
 - encoding schemes 40
 - in Unix 35
 - UTF-32BE (Big-endian)
 - encoding scheme 41
 - encoding scheme (definition) 131
 - UTF-32LE (Little-endian)
 - encoding scheme 41
 - encoding scheme (definition) 132
 - UTF-8 36, 124, 883
 - ASCII transparency 36
 - binary comparison and sort order 39
 - bit distribution (table) 125
 - BOM in 130, 133, 834
 - byte ranges 125
 - compared to multibyte encodings 37
 - encoding form (definition) 124
 - encoding scheme 40
 - encoding scheme (definition) 130
 - in Unix 18
 - in UTF-16 order 233
 - non-shortest form is invalid 124, 246
 - preferred encoding for Internet protocols 37
 - security and 246
 - signature 130, 133, 834
 - UTF-EBCDIC
see UTR #16, UTF-EBCDIC
 - UTN (Unicode Technical Note) 872
 - UTR (Unicode Technical Report) 867
 - abstracts 870
 - UTS (Unicode Technical Standard) xxxv, 867
 - abstracts 868
 - Uyghur 367
- ## V
- Vai 717–718
 - reference materials 951
 - valid (synonym for well-formed) 122
 - variable-width Unicode encoding form 36, 124
 - variants
 - compatibility 26
 - fullwidth and halfwidth 287
 - mathematical symbols 781
 - small form 286
 - standardized 824
 - variation selectors 192, 824
 - ideographic variation mark (U+303E) 683
 - Mongolian free variation selectors 525
 - variation sequences 824
 - for Phags-pa 560–562
 - Version 7.0 79
 - Version 8.0
 - number of characters 3
 - versions of the Unicode Standard xxxv, 74, 874, 890–891
 - backward compatibility 74
 - compared to ISO/IEC 10646 editions 890
 - content 75
 - interaction in implementations 201
 - numbering 75
 - property changes 74
 - stability 74
 - updates 874
 - vertical tab (U+000B) 209, 809
 - vertical text 53, 264, 286
 - East Asian scripts 662
 - Mongolian 522
 - Vietnamese 294, 301
 - ideographs 662
 - virama 260, 439
 - definition 444
 - Kharoshthi 554
 - Khmer 616
 - Myanmar 607

Philippine scripts	642
virama-like characters	192
visual order used for Thai and Lao	21
vowel harmony	
Mongolian	526
vowel marks, Middle Eastern scripts	359
vowel separator	
Mongolian	527
vowel signs	
Indic	56, 443
Khmer	618
Philippine scripts	642

W

Warang Citi	537
reference materials	951
wchar_t	
and Unicode encoding forms	38
in C language	200
weak directional characters	173
weather symbols	793
website, Unicode Consortium	873
Weierstrass elliptic function symbol	755
well-formed	
definition	121
Welsh	296
Where Is My Character?	874
wide characters	
data type in C	200
wiggly fence (U+29DB)	779
Windows newline function	210
word breaks	219, 811–813
in South Asian scripts	601, 609, 623
word joiner (U+2060)	811
writing direction <i>see</i> directionality	
writing systems	258–262
Wu (Shanghainese)	669

X

Xibe	522
Xishuangbanna Dai	626
XML	
<i>see</i> UTR #20, Unicode in XML and Other Markup Languages	

Y

Yi	695–697
reference materials	952
Yiddish	361
Yijing Hexagram Symbols	799
ypogegrammeni	305

Z

Zapf Dingbats	795
zero extension relation among encodings	882
zero width joiner (U+200D)	369–370, 814
zero width no-break space (U+FEFF) ...	67, 83, 811
initial	133, 834
zero width non-joiner (U+200C)	369–370, 815
zero width space (U+200B)	812
for word breaks in South Asian scripts ..	601, 609, 623
zero-width space characters	812
ZWJ <i>see</i> zero width joiner (U+200D)	
ZWNBSpace <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)	