

The Unicode® Standard

Version 12.0 – 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.

© 2019 Unicode, Inc.

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

The Unicode Standard / the Unicode Consortium; edited by the Unicode Consortium. — Version 12.0.

Includes index.

ISBN 978-1-936213-22-1 (<http://www.unicode.org/versions/Unicode12.0.0/>)

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

QA268.U545 2019

ISBN 978-1-936213-22-1

Published in Mountain View, CA

March 2019

I Index

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

For definitions of terms used, see the glossary on the Unicode website. To find the code points for specific characters or the code ranges for particular scripts, use the Character Index on the Unicode website. (See *Section B.3, Other Unicode Online Resources*.)

A

- abbreviation, Coptic 311
- abjads 256, 359
- abstract character sequences
 - definition 90
- abstract characters 29
 - definition 90
- abugidas 257, 258, 445, 629
- accent marks *see* diacritics
- accented characters
 - encoding 12
 - Latin 289
 - normalization 206
- accounting numbers, ideographic 176
- acrophonic numerals 205, 308
- Adlam 772–773
- Aegean numbers 340
- Africa
 - scripts of 751–774
- Afrikaans 294
- Ahom 624–625
- Ainu 731
- Aiton 644
- Alchemical Symbols 855
- Algonquian 778
- Ali Gali 532
- aliases
 - character name 88, 181
 - informative 908
 - normative 909
 - property 162
 - property value 162
- allocation areas 45
- allocation of encoded characters 44–52
- Alphabetic (informative property) 188
- alphabets 256
 - European 287–336
 - mathematical 811–815
- alternate format characters (deprecated) ... 192, 882–883
- Americas
 - scripts of 775–783
- Amharic 752
- Anatolian hieroglyphs 442–443
- Ancient Symbols 859
- angle brackets (U+2329 and U+232A)
 - deprecated for technical publication 841
- Annexes, Unicode Standard (UAX) xxiv, 929
 - as components of Unicode Standard 79
 - conformance 85
 - list of 85
- annotation characters 895–897
 - use in plain text discouraged 896
- ANSI/ISO C
 - wchar_t and Unicode 200
- apostrophe (U+0027) 272
- Arabic 367–390
 - digits 818
- Arabic-Indic digits 371–372
 - signs used with 373
- ArabicShaping.txt 375, 380, 396
- Aramaic 412, 445, 532, 563, 569
- areas of the Unicode Standard 45
- ARIB 851
- Armenian 319–320
- arrows 837–838
- ASCII
 - characters with multiple semantics 262
 - transparency of UTF-8 36
 - Unicode modeled on 1
 - zero extension 200, 942
- Assamese 472
- assigned code points 11, 30
- Athapascan 778
- atomic character boundaries 218
- Avestan 420–421

B

Balinese 683–688
 Bamum 767–768
 Bangla 472–478
 base characters 327
 definition 106
 multiple 59
 ordered before combining marks 220, 327
 Basic Multilingual Plane (BMP) 1, 44
 allocation areas 49
 representation in UTF-16 36
 Basque 294
 Bassa Vah 769
 Batak 694–695
 benefits of Unicode 1
 Bengali 472–478
 Bhaiksuki 575–576
 Bidi Class (normative property) 171
 Bidi Mirrored (normative property) 178
 Bidi Mirroring Glyph (informative property) 179
 BidiMirroring.txt 179
 Bidirectional Algorithm, Unicode 53, 84
 bidirectional ordering 20
 controls 879
 bidirectional text 53, 84
 Middle Eastern scripts 359
 nonspacing marks in 223
 punctuation in 261
 big-endian 40
 definition 83
 Bihari 468
 binary comparison and sort order
 caution for UTF-16 36
 UTF differences 231, 233
 UTF-8 39
 block 45, 90, 255, 903
 headers 916
 BMP *see* Basic Multilingual Plane
 BNF (Backus-Naur Form) 923
 BOCU-1 *see* UTN #6, BOCU-1
 MIME-Compatible Unicode Compression
 Bodhi 521
 Bodo 467
 BOM (U+FEFF) 40, 67, 130–133, 893–895
 Bopomofo 727–729
 boundaries, text 61, 189, 217–218, 228
 see also UAX #14, Unicode Line Breaking Algorithm
 see also UAX #29, Unicode Text Segmentation
 boustrophedon 53, 349
 box drawing symbols 845
 Brahmi 445, 563, 565–568, 569, 631
 Braille 786–787

Breton 294
 Buginese 681–682
 Buhid 678
 Bulgarian 313
 bullets 275
 numeric 819
 Burmese *see* Myanmar
 Byelorussian 313
 byte order mark (BOM) (U+FEFF) ..40, 67, 130–133, 893–895
 byte ordering
 changing 81
 conformance 83
 byte serialization 40, 67
 Byzantine Musical Symbols 794

C

C language
 wchar_t and Unicode 200
 C0 and C1 control codes 31, 187, 868
 Cambodian *see* Khmer
 Canadian Aboriginal Syllabics 778–779
 candrabindu 470, 600
 canonical composite characters
 see canonical decomposable characters
 canonical composition algorithm 138
 canonical decomposable characters
 definition 118
 canonical decomposition 63
 definition 117
 mappings 116
 canonical equivalence
 definition 118
 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 710
 capital letters 164, 236, 287
 Carian 343
 carriage return (U+000D) (CR) 209, 869
 carriage return and line feed (CRLF) 209
 case 295
 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 287
- exceptional Latin pairs 291, 295
- Georgian 322
- lowercase 164, 236, 287
- mapping tables 196
- mappings 152, 166, 236–238
- mappings noted in code charts 912
- titlecase 164, 236
- Turkish I 238, 291
- uppercase 164, 236, 287
- see also* default case
- Case (normative property) 164, 236
- CaseFolding.txt 166, 240
- caseless letters 295
- Catalan 293
- Caucasian Albanian 354
- cedilla 290
- CEF *see* character encoding forms
- CES *see* character encoding schemes
- Chakma 552
- Cham 669–670
- character encoding forms (CEF) 33–39, 942
- 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 930
- character literals, Unicode
- code point notation U+ 924
- character names 88, 180–186, 946
- aliases 88, 181
- conventions 921
- for CJK ideographs 917
- for control codes 185, 187
- in code charts 908
- matching 181
- character properties
- see* properties
- see also individual properties, e.g.* Combining Class
- character semantics 1, 80, 87–88, 947
- as Unicode design principle 18
- ASCII 262
- definition 87
- character sequences
- abstract *see* abstract character sequences
- canonical equivalent *see* canonical equivalent character sequences
- compatibility equivalent *see* compatibility equivalent character sequences
- conformance 81
- named 181
- character sequences, combining 106
- character shaping selectors (deprecated) 882
- character tabulation (U+0009) 869
- characters
- abstract *see* abstract characters
- arrangement in Unicode 46
- assigned 11, 30
- boundaries 217
- canonical decomposable *see* canonical decomposable characters
- classes 924
- code charts 903–920
- coded *see* encoded characters
- combining *see* combining characters
- compatibility decomposable *see* compatibility decomposable characters
- composite *see* decomposable characters
- concept of 15, 60
- conformance definitions 90–93
- confusable 245
- 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, 263, 867–883
- glyphs, relationship to 15
- graphic 30
- identity (definition) 87
- ignored in processing 248–253
- interpretation 80
- layout control 68, 871–881
- modification 81
- names list 904–916
- names *see* character names
- not encoded in Unicode 3
- number encoded in Version 12.0 3
- precomposed *see* decomposable characters
- properties *see* properties
- semantics *see* character semantics
- special 67, 867–902
- supplementary *see* supplementary characters
- transcoding 196–197
- unsupported 201

- characters, not glyphs
 - in spoofing 246
 - Unicode principle 15
- charsets
 - IANA registered names 41
- Cherokee 776
- Chinese 709–711
 - Cantonese 710
 - Hakka 728
 - Mandarin 710
 - Minnan (Hokkien/Fujian, incl. Taiwanese) .. 728
 - simplified and traditional 709
- Chu hán 708
- Chu Nôm 958
- citations for
 - properties 77
 - Unicode algorithms 78
 - Unicode Standard 76
- CJK ideographs 258, 704–720
 - accounting numbers 176
 - CJK Compatibility Ideographs 719–720
 - CJK Compatibility Supplement 720
 - CJK Strokes 722, 961
 - CJK Unified Ideographs 704–718
 - CJK Unified Ideographs Extension A 706
 - CJK Unified Ideographs Extension B 718
 - CJK Unified Ideographs Extension C 719
 - CJK Unified Ideographs Extension D 719
 - CJK Unified Ideographs Extension E 719
 - CJK Unified Ideographs Extension F 719
- code charts 917
 - compatibility ideographs in Plane 2 52
 - component structure 714
 - encoding blocks 705
 - ideographic description sequences 723–726
 - ideographic variation mark (U+303E) 725
 - KangXi radicals 717, 721–722
 - names 917
 - numbers 818
 - numeric values 176, 205
 - order of encoding 716
 - radicals 721–722
 - source standards 960
 - unknown or unavailable 284
 - Vietnamese 702
- CJK Miscellaneous Area 50
- CJK punctuation and symbols 282
 - compatibility forms 284
 - overscores and underscores 284
 - quotation marks 270
 - sesame dots 283
 - vertical forms 284
- CJK-JRG (Chinese/Japanese/Korean Joint Research Group) 956
- CJKV Ideographs Area 50
- cluster boundaries 217
- code charts 903–920
 - representative glyphs 904
- code point sequences
 - notation 922
- code points 7, 29
 - assigned 11, 30
 - assignment 46
 - categories 30
 - default ignorable 201, 252
 - definition 90
 - designated 30
 - notation 921
 - 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
 - undesigned 30
- code positions *see* code points
- code set independence 18
- code unit sequences
 - definition 120
 - ill-formed (definition) 122
 - notation 922
 - well-formed (definition) 122
- code units
 - definition 120
 - isolated 119
- code values *see* code units
- coded character representations
 - see* coded character sequences
- coded character sequences
 - definition 92
- coded characters *see* encoded characters
- codespace *see* Unicode codespace
- coeng 645, 648
- Collation Algorithm, Unicode (UCA) 12
- collation *see* sorting
- collation tables 196
- combining character sequences 56, 106
 - defective 223
 - definition 108
 - Latin 289
 - line breaking 219
 - matching 219
 - order of base character and marks 220, 327
 - rendering 219
 - selection 217
 - truncation 220–221
- combining characters 55–60, 110–115, 219–227
 - blocking reordering 878

- canonical ordering 62, 137, 168
- combining marks 327–328
- definition 106
- 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 169
- strikethrough 170
- subjoined 170
- 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) 877
- combining half marks 190, 335
- combining marks *see* combining characters
- comma below 290
- Compatibility and Specials Area 26, 50
- compatibility characters 22
- compatibility composite characters 27
 - see* compatibility decomposable characters
- compatibility decomposable characters 26
 - definition 116
- compatibility decomposition 63
 - definition 116
- compatibility decomposition mappings 116
- compatibility equivalence
 - definition 117
- compatibility equivalent character sequences
 - conformance 81
 - see* compatibility decomposition mappings
- compatibility precomposed characters
 - see* compatibility decomposable characters
- compatibility variants 26
 - mapping 243
- composite characters
 - see* decomposable characters
- Composition Exclusion (normative property) ... 100
- compression 208
 - see also* UTS #6, A Standard Compression Scheme for Unicode (SCSU)
- conferences 930
- conformance 73–158
 - definitions 87–93
 - examples 69
 - ISO/IEC 10646 implementations 947
 - requirements 79–84
- confusables 245
- conjunct consonants
 - Indic 217, 453
 - Myanmar 639
 - selection of clusters 217
- contextual shaping
 - apostrophe 272
 - Arabic 367
 - not used for Hebrew final forms 362
 - quotation marks 268
 - Syriac 395
- contour tones 325
- control codes 31, 68, 868
 - graphics for 840
 - names 187
 - properties 869
 - semantics 32, 869
 - specified in Unicode 869
- control sequences 868
- conversion of characters 196–197
- convertibility
 - as Unicode design principle 24
- Coptic 307, 310–312
- Coptic Epact numbers 823
- corporate use subarea 888
- corrigenda 76
- CR (U+000D carriage return) 209, 869
- CRLF (carriage return and line feed) 209
- Croatian 294
 - digraphs 294
- culturally expected sorting 12, 230
- Cuneiform
 - Old Persian 433
 - Sumero-Akkadian 428–431
 - Ugaritic 432
- Cuneiform and Hieroglyphic Area 51
- Cuneiform and Hieroglyphs 427–443
- currency symbols block 805–808
 - currency symbols encoded in other blocks .. 806
 - currency symbols, other 807
 - dollar sign, form and usage 806
 - euro sign 807
 - lari sign 807
 - lira sign, compatibility usage 806
 - lira sign, Turkish 807
 - peso signs, usage 806
 - ruble sign 807
 - rupee signs, Indian, usage 807
 - yen and yuan signs, usage 806

- cursive joining 873–877
 - Arabic 375–382
 - control characters for 191, 369–370, 535, 872
 - Mandaic 403
 - Mongolian 534–536
 - N’Ko 763
 - Phags-pa 582
 - Syriac 395–398
 - transparency 876
- cursive scripts 359
- Cypriot 342
 - see also* Linear B
- Cyrillic 313–316
- Czech 294
- D**
- danda, in Devanagari block 466
- Danish 293
- dashes 265
- Database, Unicode Character
 - see* Unicode Character Database (UCD)
- dead consonants, Indic 450
- dead keys 220
- decomposable characters 63
 - definition 116
 - normalization of 206
- decomposition 63, 116–118
 - canonical *see* canonical decomposition
 - compatibility *see* compatibility decomposition
 - definition 116
 - in normalization 206
 - mapping, definition 116
 - mappings noted in code charts 912
- default case
 - algorithms 85, 152–158
 - conversion 154
 - detection 156
 - folding 155
- default caseless matching 157
- default grapheme clusters 217
 - see also* UAX #29, Unicode Text Segmentation
- Default Ignorable Code Point (property) 252
- default ignorable code points 201, 252
- default property values
 - definition 97
- defective combining character sequences 223
 - definition 108
- dependent vowel signs
 - Indic 449
 - Khmer 650
 - Philippine scripts 678
- deprecated characters 74, 907
 - alternate format 192, 882–883
 - definition 92
- Derived Age (property) 201
- derived properties
 - definition 104
- DerivedCoreProperties.txt 153, 164, 253
- DerivedNormalizationProps.txt 242
- Deseret 781–783
- design goals of Unicode 4
- design principles of Unicode 14–24
- designated code points 30
- Devanagari 447–471
- Dhivehi 515
- diacritics 55, 327
 - alternative glyphs 289, 327
 - Czech 290
 - display in isolation 60, 265, 328
 - double 114, 190, 329
 - German dialectology 333
 - Greek 303–304, 307
 - Latin 289–292
 - Latvian 290
 - mathematical 814
 - on i and j 291
 - rendering 222–227
 - Slovak 290
 - spacing clones of 325, 329
 - symbol 55, 334
 - see also* combining characters
- dictionary symbols 851
- digit form names 371
- digits 205
 - Arabic 818
 - Arabic-Indic 371–372
 - compatibility 818
 - decimal 175
 - glyph variants 820
 - hexadecimal 818
 - Myanmar 818
 - national shapes 883
 - Shan 818
 - superscript and subscript 819
 - Tai Laing 818
 - Tai Tham 818
- digraphs 294, 297, 299
- dingbats 854–855
- directionality 20, 53
 - East Asian scripts 702
 - Middle Eastern scripts 359
 - Mongolian 534
 - musical symbols 789
 - normative property 171
 - Ogham 356
 - Old Italic 346
 - Philippine scripts 679
 - Runic 349

discussion list for Unicode 930
 Dogra 627–628
 Dogri 467
 Domino Tiles 856
 dotless i 238, 291
 dotted circle
 in code charts 107, 328
 in fallback rendering 222
 to indicate diacritic 55
 to indicate vowel sign placement 56
 double diacritics 114, 190, 329
 Duployan 798–799
 Dutch 293, 294
 dynamic composition
 as Unicode design principle 23
 Dzongkha 521

E

East Asian scripts 701–750
 writing direction 53
 see also CJK ideographs
 Eastern Arabic-Indic digits 371
 EBCDIC
 newline function 210
 editing, text boundaries for 217–218
 efficiency
 as Unicode design principle 15
 Egyptian hieroglyphs 434–439
 format controls 436–439
 Elbasan 353
 ellipsis 273–274
 Elymaic 422
 e-mail discussion list for Unicode 930
 emoji 849, 930
 animal symbols 853
 charts 930
 cultural symbols 853
 zodiacal symbols 853
 emoji modifiers 853
 emoticons 853
 Enclosed Alphanumerics 863
 enclosing marks 335
 definition 107
 encoded characters 7, 29
 allocation 44–52
 definition 92
 encoding form conversion
 definition 127
 encoding forms 33–39
 ISO/IEC 10646 definitions 942
 encoding forms, Unicode
 see Unicode encoding forms

encoding model for Unicode characters 33, 42
 see also UTR #17, Unicode Character Encoding Model
 encoding schemes 40–43
 encoding schemes, Unicode
 see Unicode encoding schemes
 endian ordering
 see byte order mark (BOM) (U+FEFF)
 end-user subarea 889
 English 293
 equivalent sequences 206
 as Unicode design principle 23
 case-insensitivity 231, 240
 combining characters in matching 219
 conformance 82
 Hangul syllables 737
 in sorting and searching 230
 language-specific 118
 security implications 245
 see also canonical equivalence
 see also compatibility equivalence
 see also encoding forms, encoding schemes
 errata xxvi, 76, 931
 escape sequences 868
 not used in Unicode 1, 4
 Esperanto 294
 Estonian 294
 Ethiopic 752–755
 Etruscan 345
 European scripts 287–336
 ancient 337–357
 eyelash-RA 459

F

fallback rendering 252
 of nonspacing marks 222
 FAQ (Frequently Asked Questions) 930
 Faroese 293
 Farsi 367, 370
 featural syllabaries 257
 FF (U+000C form feed) 209, 869
 file separator (U+001C) 869
 Finnish 293
 Finno-Ugric Transcription (FUT)
 see Uralic Phonetic Alphabet (UPA)
 fixed-width Unicode encoding form (UTF-32) ... 35,
 124
 flat tables 196
 Flemish 293
 fleurons 855
 fonts
 and Unicode characters 16
 for mathematical alphabets 813–815
 style variation for symbols 803

form feed (U+000C) (FF) 209, 869
 format control characters 30, 68, 263, 867–883
 deprecated 882–883
 prefixed 192, 331
 stateful 880
 fraction characters 831
 fraction slash (U+2044) 273, 827
 French 294
 Frisian 294
 fullwidth forms in East Asian encodings 734
 futhark 348

G

Garshuni 391
 Ge'ez 752
 General Category (normative property) 172
 list of values 172
 general punctuation 261–285
 General Scripts Area 50
 geometrical symbols 845–848
 Georgian 321–322
 German 293
 geta mark (U+3013) 284
 Glagolitic 318
 Glossary 930
 glyph selection tables 196
 glyphs 6, 15
 characters, relationship to 15
 diacritics alternative 289, 327
 Greek alternative 304–306
 Latin alternative 289
 mathematical alternative 833
 missing 252
 representative in code charts 904
 standardized variants 884
 symbols alternative 803
 golden numbers 350
 Gothic 352
 Grantha 621–623
 grapheme base 327
 definition 108
 grapheme clusters 11, 60–61
 see also UAX #29, Unicode Text Segmentation
 default 217
 definition 109
 grapheme extender
 definition 109
 grapheme joiner, combining (U+034F) 877
 graphic characters 30
 Greek 303–308
 acrophonic numerals 205, 308
 alternative glyphs 304–306
 ancient musical notation 795–797

editorial marks 279
 letters as symbols 304–306, 834
 see also Cypriot, Linear B
 Greenlandic 294
 group separator (U+001D) 869
 guillemets 268
 Gujarati 484–485
 Gunjala Gondi 559–560
 Gurmukhi 479–483

H

Hakka 728
 halant 445
 see also virama
 half marks, combining 190, 335
 half-consonants, Indic 454
 halfwidth forms in East Asian encodings 734
 Han ideographs *see* CJK ideographs
 Han unification 711–718
 and language tags 215
 history 955–960
 language usage 709
 source separation rule 706, 712
 source standards 960
 hand symbols 853
 Hangul Area 50
 Hangul syllables 701, 735–738
 and combining marks 114
 as grapheme clusters 61
 canonical decomposition 144
 collation 737
 composition 146
 conjoining jamo 142–151
 equivalent sequences 737
 Hangul Compatibility Jamo 736
 Hangul Jamo 735–738
 Hangul Syllables block 737–738
 Johab set 737
 name generation 147
 normalization 736
 standard 143
 Hangzhou numerals 826
 Hanifi Rohingya 676
 Hanja *see* CJK ideographs
 Hanunóo 678
 Hanzi *see* CJK ideographs
 harakat 368
 hasant 472
 hash tables 197
 Hatran 425
 Hebrew 361–366
 hentaigana 731–733

- hieroglyphs
 - Anatolian 442–443
 - Egyptian 434–439
 - Meroitic 440–441
 - high surrogate
 - definition 119
 - high-surrogate code points 79, 890
 - high-surrogate code units 119
 - higher-level protocols
 - definition 93
 - Hindi 447
 - Hiragana 730
 - horizontal tab (U+0009) 869
 - HTML newline function 210
 - Hungarian 294
 - hyphenation 872
 - as a text process 10
 - hyphens 265, 872
- I**
- I Ching symbols 858
 - IANA charset names 41
 - Icelandic 293
 - identifiers 229
 - see also* UAX #31, Unicode Identifier and Pattern Syntax
 - Ideographic (informative property) 188
 - ideographic description sequences 724
 - Ideographic Rapporteur Group (IRG) 958
 - ideographs *see also* CJK ideographs
 - IICore 707, 958
 - ill-formed
 - definition 122
 - Imperial Aramaic 412–413
 - implementation guidelines 195–254
 - in a Unicode encoding form
 - definition 123
 - in-band mechanisms 902
 - India
 - Official scripts 445–512
 - Indian rupee signs, usage 807
 - Indic scripts 445–512
 - principles, in terms of Devanagari 448–458
 - relation to ISCII standard 447
 - Indic Siyaq 825
 - Indonesia and Oceania
 - scripts of 677–699
 - Indonesian 293
 - industry character sets
 - covered in Unicode 3
 - information separators (U+001C..U+001F) 869
 - informative properties
 - definition 101
 - Inscriptional Pahlavi 418
 - Inscriptional Parthian 418
 - inside-out rule 222
 - interchange restrictions 31
 - International Phonetic Alphabet (IPA) 256, 296–297
 - Spacing Modifier Letters 324
 - see also* phonetic alphabets
 - internationalization 18
 - Internationalization & Unicode Conference 930
 - Internet protocols
 - UTF-8 as preferred encoding 37
 - Inuktitut 778
 - invisible operators 839
 - iota subscript 304
 - IPA *see* International Phonetic Alphabet
 - IRG (Ideographic Rapporteur Group) 958
 - Irish 293, 356
 - ISCII standard and Unicode 447
 - ISO/IEC 10646 933–947
 - conformance of Unicode implementations .. 947
 - encoding forms 942
 - synchrony with Unicode Standard 944
 - timeline compared to Unicode versions 935
 - Italian 293
 - ITC Zapf Dingbats 854
 - IUC *see* Internationalization & Unicode Conference
- J**
- jamos *see* Hangul syllables
 - Japanese 701
 - Javanese 689–692
 - Jawi 387
 - jihvamuliya 471, 600
 - Johab 737
 - joiners 369
 - combining grapheme joiner (U+034F) 877
 - word joiner (U+2060) 871
 - zero width joiner (U+200D) 369–370, 874
 - justification 224
- K**
- Kaithi 597–599
 - Kana (Hiragana and Katakana) 730–731
 - Kanbun 720
 - KangXi radicals 717, 721–722
 - Kanji *see* CJK ideographs
 - Kannada 502–505
 - Kashmiri 468
 - Katakana 730–731
 - Kawi 683, 685
 - Kayah Li 668
 - KC (normalization form)
 - see* Normalization Form KC

- KD (normalization form)
see Normalization Form KD
- keytop labels 840
- Khamti Shan 642
- Kharoshthi 569–570
- Khmer 645–656
 characters not recommended 653
 syllable components, order of 654
- Khojki 608–609
- Khudawadi 610–611
- killer 258
 Batak 694
 Brahmi 565
 Meetei Mayek 546
 Myanmar (asat) 640
see also virama
- Konkani 467
- Korean Hangul *see* Hangul
- Kurdish 387
- L**
- Ladino 361
- language tags 215, 898–902
 and Han unification 215
 use strongly discouraged 898, 901
- Lanna 659
- Lao 635–637
- last-resort glyphs 252
- Latin 289–302
 alternative glyphs 289
 Basic Latin 293
 encoding blocks 45
 IPA Extensions 296–297
 Latin Extended Additional 299–302
 Latin Extended-A 293
 Latin Extended-B 294–296
 Latin Extended-C 299
 Latin Extended-D 300
 Latin Extended-E 301
 Latin Ligatures 299
 Latin-1 Supplement 293
 Phonetic Extensions 298–302
- Latvian 294, 301
 cedilla 290
- layout control characters 68, 871–881
- leading surrogates
see high-surrogate code units
- legibility criterion for plain text 19
- Lepcha 553–555
- letter spacing 872
- letterlike symbols 809–815
- LF (U+000A line feed) 209, 869
- ligatures 873–877
 Arabic 378–379
 combining characters on 59
 control characters for 191
 for nonspacing marks 226
 Latin 299
 selection 218
 Syriac 399
- Limbu 542–545
- line breaking 209–213, 871–873
 control characters 190
 in South Asian scripts 633, 641, 656
 recommendations 211
see also UAX #14, Unicode Line Breaking Algorithm
- line feed (U+000A) (LF) 209, 869
- line separator (U+2028) (LS) 209, 873
- line tabulation (U+000B) (VT) 869
- Linear A 339
- Linear B 340–341
see also Cypriot
- linear boundaries 218
- Lisu 743–745
- Lithuanian 294
- little-endian 40
 definition 83
- logical order
 as Unicode design principle 19
 exceptions to 169
- logograph 258
- logosyllabaries 258
- low surrogate
 definition 119
 low-surrogate code points 79, 890
 low-surrogate code units 119
- lowercase 164, 236, 287
- LS (U+2028 line separator) 209, 873
- Lycian 343
- Lydian 343
- M**
- MacOS newline function 210
- Mahajani 606–607
- Mahjong Tiles 856
- mail discussion list for Unicode 930
- Maithili 467
- major version 75
- Makasar 698–699
- Malay 293
- Malay, Patani 634
- Malayalam 506–512
 Suriyani 399, 507

- Maltese 294
- Manchu 533
- Mandaic 402–404
- Mandarin 710
- Manden 760
- Manichaean 414–417
- map symbols 851
- mapping tables *see* tables of character data
- Marathi 447, 459, 466
- Marchen 584
- markup languages
 and Unicode conformance 902
 line breaking 209
- Masaram Gondi 557–558
- Mathematical (informative property) 831
- mathematical expression format characters 192
 see also UTR #25, Unicode Support for Mathematics
- mathematical symbols 831–838
 alphabets 811–815
 alphanumeric 810–815
 fonts 813–815
 format characters 839
 fragments for typesetting 841
 invisible operators 839
 operators 832–835
 standardized variants 838
- MathML 835
- matras 168, 449
- Medefaidrin 774
- Meetei Mayek 546–547
- Mende Kikakui 770–771
- Meroitic
 cursive 440–441
 hieroglyphs 440–441
- Miao 746–747
- Middle Eastern scripts 359–516
 ancient 405–425
- Min 710
- Minnan (Hokkien/Fujian, incl. Taiwanese) 728
- minor version 75
- minus sign 834
 commercial (U+2052) 276
- mirrored property
 see Bidi Mirrored (normative property)
- mirroring of paired punctuation 267
- Miscellaneous Symbols 850
- missing glyphs 252
- Modi 616–618
- modifier letters 323–326
- Modifier Letters, Spacing 299
- Mongolian 532–541, 577
 writing direction 534
- moon symbols 851
- Mro 548
- Multani 612
- multibyte encodings
 compared to UTF-8 37
- multistage tables 196
- musical symbols 788–797
 ancient Greek 795–797
 Balinese 687
 Byzantine 794
 directionality 789
 Gregorian 793
 Kievan 793
 Western 788–793
- Myanmar 638–644
 digits 818
 Myanmar Extended-A 642
 Myanmar Extended-B 642
- ## N
- N’Ko 760–764
- Nabataean 423
- named character sequences 181
- names, character *see* character names
- namespace 89
- Nandinagari 619–620
- NEL (U+0085 next line) 209, 869
- Nepali 447
- neutral directional characters 171
- New Tai Lue 659–661
- Newa 519–520
- newline function (NLF) 210, 870
- newline guidelines 209–213
- next line (U+0085) (NEL) 209, 869
- NFC (Normalization Form C) 62
- NFD (Normalization Form D) 62
- NFKC (Normalization Form KC) 62
- NFKD (Normalization Form KD) 62
- NLF (newline function) 210, 870
- no-break space (U+00A0) 871
 base for diacritic in isolation 60, 265, 328
- no-break space, narrow (U+202F) 538
- noncharacter code points *see* noncharacters
- noncharacters 31, 891
 conformance 79
 definition 93
 handling 82
 in code charts 907
 interchange restrictions 31
 semantics 32
 U+10FFFF (not a character code) 891
 U+FDD0..U+FDEF 31, 891
 U+FFFE (not a character code) 67, 892
 U+FFFF (not a character code) 31, 891

nondecomposable characters 64
 non-joiner, zero width (U+200C) 369–370, 875
 nonlinear boundaries 218
 non-overlap principle in Unicode encoding forms 33
 nonspacing marks 327
 definition 107
 display in isolation 60, 265, 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 888
 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 99
 list 100
 may change 99
 Norwegian 293
 notational conventions 921–925
 notational systems 260, 785–801
 nukta 368, 389, 460
 null (U+0000)
 as Unicode string terminator 870
 number forms
 CJK ideographs 205
 numbers
 Coptic Epact 823
 handling 205
 ideographic accounting 176
 numerals 816–828
 acrophonic 308
 Chinese counting rods 829
 Coptic 312
 Cuneiform 431
 Ethiopic 754
 Greek acrophonic 205
 Hangzhou 826
 Meroitic cursive 441
 old-style 273

 Roman 205, 831
 Rumi 824
 Suzhou-style 826
 numeric separators 276
 numeric shape selectors (deprecated) 883
 Numeric Type (normative property) 175
 Numeric Value (normative property) 175
 numero sign (U+2116) 809
 Nüshu 742
 Nyiakeng Puachue Hmong 673–674

O

object replacement character (U+FFFC) 897
 octet 923
 Ogham 356
 Ol Chiki 550–551
 Old Church Slavonic 313
 Old Hungarian 351
 Old Italic 345–347
 Old North Arabian 407
 Old Permic 355
 Old Persian 433
 Old Sogdian 591
 Old South Arabian 408–409
 Old Turkic 590
 old-style numerals 273
 Oriya 486–488
 ornamental dingbats 855
 Oromo 752
 Osage 780
 Osmanya 756
 Ottoman Siyaq 825
 out-of-band mechanisms 902
 overlapping encodings 33
 overscores 273

P

Pahawh Hmong 671–672
 Pahlavi, Inscriptional 418
 Pahlavi, Psalter 419
 Palmyrene 424
 Panjabi 479
 paragraph or section marks 276
 paragraph separator (U+2029) (PS) 209, 873
 Parthian, Inscriptional 418
 Pashto 367
 Patani Malay 634
 Pau Cin Hau 675
 Persian 367, 370
 Phags-pa 577–583
 Phaistos Disc symbols 860
 Phake 644

- Philippine scripts 678–680
- Phoenician 410
- phonemes 259
- phonetic alphabets 256
- IPA Extensions 296–297
 - Phonetic Extensions 298–302
 - Spacing Modifier Letters 324–326
 - Uralic Phonetic Alphabet (UPA) 276, 298
 - see also* International Phonetic Alphabet (IPA)
- Pinyin 293
- pipeline table
- proposed new characters 931
- 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, 889
- Playing Cards 857
- points, Hebrew pronunciation marks 361
- policies of the Unicode Consortium 931
- Polish 294
- Portuguese 293
- precomposed characters
- see* decomposable characters
 - compatibility *see* compatibility decomposable characters
- prefixed format control characters 192
- prepended concatenation marks 253, 331
- Private Use Area (PUA) 50, 888
- Private Use planes 45, 52, 889
- private-use characters
- properties 887
 - semantics 32
- private-use code points 31, 201
- conformance 80
 - definition 105
 - high surrogates 890
- processing code, Unicode as 38
- properties 18, 95–105, 159–193
- aliases 162
 - aliases (definition) 104
 - and Unicode algorithms 100
 - 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 869
 - provisional *see* provisional properties
 - simple *see* simple properties
 - see also* individual properties, e.g. combining classes
- property values
- aliases 162
 - aliases (definition) 105
 - default 97
 - default (definition) 97
 - normative references to 84
- PropertyAliases.txt 104, 924
- PropertyValueAliases.txt 105, 924
- PropList.txt 166
- Provençal 294
- provisional properties
- definition 101
- PS (U+2029 paragraph separator) 209, 873
- Psalter Pahlavi 419
- PUA (Private Use Area) 50, 888
- punctuation 261–285
- blocks containing 255
 - CJK 282
 - doubled 273
 - in bidirectional text 261
 - paired 267
 - small form variants 285
 - typographic forms 261
 - vertical forms 284
- Punctuation and Symbols Area 50
- Punjabi 479
- ## Q
- quotation marks 268–271
- East Asian 270
 - European 268
- ## R
- radicals, KangXi and other CJK 721–722
- radical-stroke index 717
- record separator (U+001E) 869
- recycling symbols 852
- references 931
- referencing 84
- properties 77
 - Unicode algorithms 78
 - Unicode Standard 76
- regional indicator symbols 864
- regular expressions 214
- and line breaking 209
 - see also* UTS #18, Unicode Regular Expressions
- Rejang 693
- rendering of text 6, 10, 17
- fallback 252
 - unsupported characters 201

repertoire of abstract characters 29
 reph 458, 462, 500
 replacement character (U+FFFD) 43, 68, 83, 897
 reserved code points 30, 201
 definition 93
 in code charts 907
 preservation in interchange 31
 see also unassigned code points
 Rhaeto-Romanic 294
 rich text 18
 right single quotation mark (U+2019)
 preferred for apostrophe 272
 right-to-left text 53
 East Asian scripts 702
 Middle Eastern scripts 359
 roadmap for script additions 46, 931
 Roman numerals 205, 831
 Romanian 294
 comma below 291
 Romany 294
 Rong 553–555
 Rumi numeral symbols 824
 Runic 348–350
 Russian 313

S

Samaritan 400–401
 Sami 294
 Sanskrit 447
 Saurashtra 556
 scalar values, Unicode
 see Unicode scalar values
 scripts
 in Unicode Standard 3
 roadmap for future additions 46, 931
 types of 260
 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 276
 security issues 245
 self-synchronization of encoding forms 34
 semantics
 see character semantics
 sequences
 notation 922
 Serbian
 corresponding digraphs in Croatian 294
 Shan 657
 digits 818

Sharada 600–601
 Shavian 357, 743
 Show Hidden 81, 222, 252, 885
 SHY (U+00AD soft hyphen) 872
 Sibe 533
 Siddham 604–605
 signature for Unicode data 67, 893–895
 simple properties
 definition 104
 simplified Chinese 709
 Sindhi 367, 467
 Sinhala 517–518
 Sinological dot 301
 SIP (Supplementary Ideographic Plane) 44, 52
 Siyaq Numbers 824
 Indic 824
 slash, fraction (U+2044) 273
 Slovak 294
 Slovenian 294
 small letters 164, 236, 287
 SMP (Supplementary Multilingual Plane) 44, 51
 soft hyphen (U+00AD) (SHY) 872
 Sogdian 592
 Somali 756
 Sora Sompeng 626
 Sorbian 294
 sorting 12, 230
 and combining grapheme joiner 878
 as a text process 10
 case-insensitive 231
 culturally expected 12, 230
 language-insensitive 230
 see also Unicode Collation Algorithm (UCA)
 source separation rule 706, 712
 South and Central Asian scripts
 Ancient 563–592
 Other historic 593–628
 Other modern 513–560
 South Asian scripts 445–545
 Southeast Asian scripts 629–676
 Soyombo 588–589
 space (U+0020)
 base for diacritic in isolation 60, 265, 328
 space characters 264, 871–873
 graphics for 840
 space, zero width (U+200B) 264
 spacing clones of diacritics 325, 329
 spacing marks 327
 definition 108
 Spacing Modifier Letters 324–326
 Spanish 293
 special characters 67, 867–902
 SpecialCasing.txt 152, 166
 Specials 893–897

- spell-checking
 - as a text process 11
- spellings, alternative
 - see* equivalent sequences
- spoofing 245
- SSP (Supplementary Special-purpose Plane) 45
- stability 102, 161
 - as Unicode design principle 23
- stacked boundaries 217
- stacking sequences 57
 - nondefault 58
- standardized variants 536, 884
 - in the code charts 914
 - mathematical symbols 838
- StandardizedVariants.txt 536, 838
- standards coverage 3
- starters 136
- stateful encoding
 - not used in Unicode 4
 - paired format controls 880
- string comparison 12
- string literals, Unicode
 - code point notation `\u1234` 924
- strings, Unicode 43, 121
 - null termination 870
- strong directional characters 171
- styled text 18
- sublinear searching 232
- subsets, supported 71
 - conformance 80
 - ISO/IEC 10646 specification for 945
- substitution character
 - see* replacement character
- Sumero-Akkadian 428–431
- Sundanese 696–697
- superscripts 325
 - and subscripts 829
- 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, 889
- Supplementary Special-purpose Plane (SSP) 45
- supported subsets 71
 - conformance 80
- supralineation 311
- surrogate code points
 - see* surrogates
- surrogate pairs 36, 125
 - definition 119
 - processing 38, 203–204
- surrogates 31, 119, 890
 - interchange restrictions 31
 - isolated surrogates, handling 43
 - isolated surrogates, ill-formed 125
 - isolated surrogates, uninterpreted 119
 - support levels 203
- Surrogates Area 50, 890
- Sutton SignWriting 800–801
- Suzhou-style numerals 826
- svasti signs 528
- Swahili 293
- Swedish 293
- syllabaries 257
 - alphabetic property 188
 - featural 257
- Syloti Nagri 595–596
- symbols 803–865
 - animal 853
 - appearance variation 803
 - arrows 837–838
 - box drawing 845
 - cultural 853
 - currency symbols block 805–808
 - dictionary 851
 - dingbats 854–855
 - emoji 849, 864
 - Enclosed Alphanumerics 863
 - fragments for mathematical typesetting 841
 - game 852
 - gender 852
 - genealogical 852
 - geometrical 845–848
 - hand 853
 - Khmer lunar calendar 656
 - letterlike 809–815
 - map 851
 - mathematical 831–838
 - mathematical alphanumeric 810–815
 - miscellaneous 850
 - musical 788–797
 - numerals 816–828
 - recycling 852
 - regional indicator 864
 - technical 840–844
 - weather 851
 - zodiacal 853
- symmetric swapping format characters 882
- Syriac 391–399

T

tab (U+0009 character tabulation) 869
 tab, vertical (U+000B) 209, 869
 tables of character data 196–197
 optimization 197
 supplementary characters 197
 tag characters 898–902
 Tagalog 678
 Tagbanwa 678
 tags, language 215, 898–902
 use strongly discouraged 901
 Tai Laing
 digits 818
 Tai Le 657–658
 Tai Tham 662–664
 digits 818
 Tai Viet 665–667
 Tai Xuan Jing symbols 859
 Takri 602–603
 Tamil 489–498
 Tangut 748–750
 components 749–750
 radicals 749
 tashkil 368
 tashkil, harakat, points 370
 TCHAR in Win32 API 200
 Technical Reports (UTR) 929
 Technical Standards (UTS) xxvi, 929
 abstracts 930
 technical symbols 840–844
 Telugu 499–501
 terminal emulation 804
 text boundaries 61, 189, 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 515–516
 Thai 631–634
 Tibetan 521–531
 Tifinagh 757
 Tigre 752
 tilde (U+007E) 276
 Tirthuta 613–615
 titlecase 164, 236
 Todo 533
 tone letters 325–326

tone marks
 Bopomofo spacing 727, 728
 Chinantec 326
 Chinese 326
 Tai Le 657
 Thai 631
 Vietnamese 292
 traditional Chinese 709
 traffic signs 851
 trailing surrogates
 see low-surrogate code units
 transcoding 196–197
 tables 196
 Transport and Map Symbols 854
 triangulation in transcoding 196
 tries 196
 truncation
 combining character sequences 220–221
 surrogates and 204
 Turkish 294
 case mapping of I 238, 291
 cedilla 291
 lira sign 807
 two-stage tables 197

U

U+ notation 924
 U+10FFFF (not a character code) 891
 U+FEFF (BOM) 893–895
 U+FFFE (not a character code) 892
 U+FFFF (not a character code) 891
 UAX (Unicode Standard Annex) xxiv, 929
 as component of Unicode Standard 79
 conformance 85
 list of 85
 UCA *see* Unicode Collation Algorithm and *see also*
 UTS #10, Unicode Collation Algorithm
 UCD *see* Unicode Character Database
 UCS (Universal Character Set)
 see ISO/IEC 10646
 UCS-2 942
 UCS-4 942
 Ugaritic 432
 Ukrainian 313
 unassigned code points 30, 79, 201
 defined as reserved code points 93
 handling 74
 properties of 97
 semantics 79
 see also reserved code points
 underscores 273
 undesignated code points 30
 Unicode 1.0 Name (informative property) 187

- Unicode algorithms
 - and properties 100
 - conformance 84
 - definition 93
 - normative references to 78, 84
- Unicode Bidirectional Algorithm 21, 53
see also UAX #9, Unicode Bidirectional Algorithm
- Unicode Character Database (UCD) .. xxiv, 161, 931
 - 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+ 924
- Unicode codespace
 - allocation numbers 950
 - definition 90
 - planes 44
 - size 1, 29
- Unicode Collation Algorithm (UCA) 12
- Unicode conferences 930
- Unicode Consortium 928
 - addresses 932
 - Consortium membership in standards bodies 928
 - e-mail discussion list 930
 - membership 928
 - policies 931
 - website 930
- Unicode data signature 67, 893–895
- Unicode data types 199–200
 - for C 199–200
- Unicode encoding forms 120–127
 - advantages of each 38
 - conformance 34, 82
 - definition 121
 - fixed-width (UTF-32) 35, 124
 - signatures 894, 895
 - variable-width 36, 125*see also* encoding forms
- Unicode encoding schemes
 - conformance 130–133
 - definition 130
 - endian ordering 40*see also* encoding schemes
- Unicode escape sequence notation \u1234 924
- Unicode scalar values
 - definition 120
- Unicode security 245
see also UTS #39, Unicode Security Mechanisms
- Unicode Standard
 - allocation of encoded characters 44–52
 - architecture 10–13
 - areas 45
 - benefits 1
 - blocks 255
 - code charts 903–920
 - components 79
 - conformance 73–158
 - conformance of ISO/IEC 10646 implementations .. 947
 - corrections 76
 - definitions for conformance 87–93
 - design goals 4
 - design principles 14–24
 - errata 76, 931
 - normative references to 76, 84
 - number of characters 3
 - number of code points 1, 29
 - script coverage 3
 - security issues 245
 - synchrony with ISO/IEC 10646 944
 - updates 931
 - versions *see* versions of the Unicode Standard
 - see also* Version 12.0
- Unicode Standard Annexes (UAX) xxiv, 929
 - as components of Unicode Standard 79
 - conformance 85
 - list of 85
- Unicode string literals
 - code point notation \u1234 924
- Unicode strings 43
 - definition 121
- Unicode Technical Committee (UTC) 928
- Unicode Technical Reports (UTR) 929
- Unicode Technical Standards (UTS) xxvi, 929
 - abstracts 930
- UnicodeData.txt 152, 166
- unification
 - as Unicode design principle 21
 - see also* Han unification
- Unified Repertoire and Ordering (URO) ... 713, 957
see also Han unification
- Unihan Database 161, 717, 718, 931, 958
- Unihan.zip 102, 161
- unit separator (U+001F) 869
- 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
- upadhmuniya 471, 600

- update version 75
 - uppercase 164, 236, 287
 - Uralic Phonetic Alphabet (UPA) 276, 298
 - Urdu 367
 - URO (Unified Repertoire and Ordering) .. 713, 957
 see also Han unification
 - UTF, Unicode Transformation Formats 33, 121
 - advantages of each 38
 - as encoding form or scheme 133
 - binary comparison and sort order differences ...
 231, 233
 - in APIs 200
 - UTF-16 36, 125, 943
 - binary comparison and sort order caution ... 36
 - bit distribution (table) 125
 - BOM in 131, 893
 - encoding form (definition) 125
 - encoding scheme (definition) 131
 - encoding schemes 40
 - in ISO/IEC 10646 943
 - in UTF-8 order 234
 - surrogates and string handling 43, 203
 - UTF-16BE (Big-endian) 894
 - encoding scheme 41
 - encoding scheme (definition) 130
 - UTF-16LE (Little-endian) 894
 - encoding scheme 41
 - encoding scheme (definition) 130
 - UTF-32 35, 124
 - as processing code 38
 - BOM in 132
 - encoding form (definition) 124
 - encoding scheme (definition) 132
 - encoding schemes 40
 - in Unix 35
 - UTF-32BE (Big-endian)
 - encoding scheme 41
 - encoding scheme (definition) 132
 - UTF-32LE (Little-endian)
 - encoding scheme 41
 - encoding scheme (definition) 132
 - UTF-8 36, 125, 943
 - ASCII transparency 36
 - binary comparison and sort order 39
 - bit distribution (table) 126
 - BOM in 130, 133, 894
 - byte ranges 126
 - compared to multibyte encodings 37
 - encoding form (definition) 125
 - encoding scheme 40
 - encoding scheme (definition) 130
 - in Unix 18
 - in UTF-16 order 233
 - non-shortest form is invalid 125, 245
 - preferred encoding for Internet protocols 37
 - security and 245
 - signature 130, 133, 894
 - UTR (Unicode Technical Report) 929
 - UTS (Unicode Technical Standard) xxvi, 929
 - abstracts 930
 - Uyghur 367, 577
- ## V
- Vai 765–766
 - valid (synonym for well-formed) 123
 - variable-width Unicode encoding form 36, 125
 - variants
 - compatibility 26
 - fullwidth and halfwidth 285
 - mathematical symbols 838
 - small form 285
 - standardized 884
 - variation selectors 193, 884
 - ideographic variation mark (U+303E) 725
 - Mongolian free variation selectors 536
 - variation sequences 884
 - for Phags-pa 581–583
 - Version 12.0 79
 - number of characters 3
 - versions of the Unicode Standard . xxiv, 74, 931, 949–951
 - backward compatibility 74
 - compared to ISO/IEC 10646 editions 949
 - content 75
 - interaction in implementations 201
 - numbering 75
 - property changes 74
 - stability 74
 - updates 931
 - vertical tab (U+000B) 209, 869
 - vertical text 53, 262, 284
 - East Asian scripts 702
 - Mongolian 534
 - Vietnamese 292, 299
 - ideographs 702
 - virama 258, 445
 - definition 450
 - Kharoshthi 573
 - Khmer 648
 - Myanmar 639
 - Philippine scripts 678
 - virama-like characters 191
 - visual order used for Thai and Lao 21
 - vowel harmony
 - Mongolian 538
 - vowel marks, Middle Eastern scripts 359
 - vowel separator
 - Mongolian 539

vowel signs	
Indic	56, 449
Khmer	650
Philippine scripts	678

W

Wancho	561
Warang Citi	549
wchar_t	
and Unicode encoding forms	38
in C language	200
weak directional characters	171
weather symbols	851
website, Unicode Consortium	930
Weierstrass elliptic function symbol	810
well-formed	
definition	122
Welsh	294
Where Is My Character?	932
wide characters	
data type in C	200
wiggly fence (U+29DB)	836
Windows newline function	210
word breaks	219, 871–873
in South Asian scripts	633, 641, 656
word joiner (U+2060)	871
writing direction <i>see</i> directionality	
writing systems	256–260
Wu (Shanghainese)	710

X

Xibe	533
Xishuangbanna Dai	659

Y

Yi	739–741
Yiddish	361
Yijing Hexagram Symbols	858
ypogegrammeni	304

Z

Zanabazar Square	585–587
Zapf Dingbats	854
zero extension relation among encodings	942
zero width joiner (U+200D)	369–370, 874
zero width no-break space (U+FEFF) ...	67, 83, 871
initial	133, 894
zero width non-joiner (U+200C)	369–370, 875
zero width space (U+200B)	872
for word breaks in South Asian scripts	633, 641,
656	
zero-width space characters	872

ZWJ <i>see</i> zero width joiner (U+200D)
ZWNBSP <i>see</i> zero width no-break space (U+FEFF)
ZWNJ <i>see</i> zero width non-joiner (U+200C)
ZWSP <i>see</i> zero width space (U+200B)

