

The Unicode® Standard

Version 14.0 – Core Specification

To learn about the latest version of the Unicode Standard, see <https://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.

© 2021 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 <https://www.unicode.org/reporting.html>. For information about the Unicode terms of use, please see <https://www.unicode.org/copyright.html>.

The Unicode Standard / the Unicode Consortium; edited by the Unicode Consortium. — Version 14.0.

Includes index.

ISBN 978-1-936213-29-0 (<https://www.unicode.org/versions/Unicode14.0.0/>)

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

QA268.U545 2021

ISBN 978-1-936213-29-0

Published in Mountain View, CA

September 2021

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 *Appendix B.3, Other Unicode Online Resources.*)

A

- abbreviation, Coptic 313
- abjads 258, 365
- abstract character sequences
 - definition 88
- abstract characters 29
 - definition 88
- abugidas 259, 260, 455, 653
- accent marks *see* diacritics
- accented characters
 - encoding 12
 - Latin 291
 - normalization 206
- accounting numbers, ideographic 176
- acrophonic numerals 205, 310
- Adlam 802–803
- Aegean numbers 344
- Africa
 - scripts of 781–804
- Afrikaans 296
- Ahom 648–649
- Ainu 756
- Aiton 668
- Alchemical Symbols 887
- Algonquian 808
- Ali Gali 549
- aliases
 - character name 86, 181
 - informative 942
 - normative 943
 - property 162
 - property value 162
- allocation areas 44
- allocation of encoded characters 43–51
- Alphabetic (informative property) 189
- alphabets 258
 - European 289–339
 - mathematical 841–846
- alternate format characters (deprecated) ... 193, 915–916
- Americas
 - scripts of 805–813
- Amharic 782
- Anatolian hieroglyphs 453–454
- Ancient Symbols 891
- angle brackets (U+2329 and U+232A)
 - deprecated for technical publication 872
- Annexes, Unicode Standard (UAX) xxiii, 965
 - as components of Unicode Standard 77
 - conformance 83
 - list of 83
- annotation characters 928–930
 - use in plain text discouraged 929
- ANSI/ISO C
 - wchar_t and Unicode 200
- apostrophe (U+0027) 274
- Arabic 373–398
 - digits 849
- Arabic-Indic digits 377–378
 - signs used with 379
- ArabicShaping.txt 382, 386, 404
- Aramaic 422, 455, 550, 581, 587
- areas of the Unicode Standard 44
- ARIB 882
- Armenian 321–322
- arrows 868–869
- ASCII
 - characters with multiple semantics 264
 - transparency of UTF-8 37
 - Unicode modeled on 1
 - zero extension 200, 978
- Assamese 483
- assigned code points 11, 30
- Athapascan 808
- atomic character boundaries 218
- Avestan 430–431

B

Balinese 707–712
 Bamum 797–798
 Bangla 483–489
 base characters 329
 definition 104
 multiple 59
 ordered before combining marks 220, 329
 Basic Multilingual Plane (BMP) 1, 43
 allocation areas 48
 representation in UTF-16 36
 Basque 296
 Bassa Vah 799
 Batak 718–719
 Baybayin 702
 benefits of Unicode 1
 Bengali 483–489
 Bhaiksuki 593–594
 Bidi Class (normative property) 171
 Bidi Mirrored (normative property) 178
 Bidi Mirroring Glyph (informative property) 179
 BidiMirroring.txt 179
 Bidirectional Algorithm, Unicode 52, 82
 bidirectional ordering 20
 controls 912
 bidirectional text 52, 82
 Middle Eastern scripts 365
 nonspacing marks in 223
 punctuation in 263
 big-endian 39
 definition 81
 Bihari 479
 binary comparison and sort order
 caution for UTF-16 36
 UTF differences 231, 233
 UTF-8 38
 block 44, 88, 257, 937
 headers 949
 BMP *see* Basic Multilingual Plane
 BNF (Backus-Naur Form) 959
 BOCU-1 *see* UTN #6, BOCU-1
 MIME-Compatible Unicode Compression
 Bodhi 538
 Bodo 478
 BOM (U+FEFF) 39, 66, 129–132, 926–928
 Bopomofo 752–754
 boundaries, text 60, 190, 217–218, 228
 see also UAX #14, Unicode Line Breaking Algorithm
 see also UAX #29, Unicode Text Segmentation
 boustrophedon 52, 354
 box drawing symbols 876
 Brahmi 455, 581, 583–586, 587, 655

Braille 816–817
 Breton 296
 Buginese 705–706
 Buhid 702
 Bulgarian 315
 bullets 278
 numeric 851
 Burmese *see* Myanmar
 Byelorussian 315
 byte order mark (BOM) (U+FEFF) .39, 66, 129–132,
 926–928
 byte ordering
 changing 79
 conformance 81
 byte serialization 39, 66
 Byzantine Musical Symbols 824

C

C language
 wchar_t and Unicode 200
 C0 and C1 control codes 31, 188, 900
 Cambodian *see* Khmer
 Canadian Aboriginal Syllabics 808–809
 candrabindu 481, 621
 canonical composite characters
 see canonical decomposable characters
 canonical composition algorithm 137
 canonical decomposable characters
 definition 116
 canonical decomposition 62
 definition 115
 mappings 114
 canonical equivalence
 definition 116
 nonspacing marks 225
 canonical equivalent character sequences
 conformance 79
 canonical mappings
 see canonical decomposition mappings
 canonical ordering algorithm 136
 canonical precomposed characters
 see canonical decomposable characters
 Cantonese 735
 capital letters 164, 236, 289
 Carian 348
 carriage return (U+000D) (CR) 209, 901
 carriage return and line feed (CRLF) 209
 case 297
 and text processes 12
 beyond ASCII 237
 camelcase 239
 case folding 240
 case operations (conformance) 83, 151–157

- case operations and normalization 242
- case operations, reversibility 239
- cased (definition) 152
- case-insensitive comparison 156, 231, 240
- casing context (definition) 152
- conversion 153
- detection 155
- European alphabets 289
- exceptional Latin pairs 293, 297
- Georgian 324
- lowercase 164, 236, 289
- mapping tables 196
- mappings 151, 166, 236–238
- mappings noted in code charts 946
- 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 359
- cedilla 292
- CEF *see* character encoding forms
- CES *see* character encoding schemes
- Chakma 569
- Cham 693–694
- character encoding forms (CEF) 33–38, 978
- see also* Unicode encoding forms
- character encoding model 33, 41
- see also* UTR #17, Unicode Character Encoding Model
- character encoding schemes (CES) 39–42
- see also* Unicode encoding schemes
- character encoding standards
- coverage by Unicode 3
- Character Index 966
- character literals, Unicode
- code point notation U+ 960
- character names 86, 180–187, 982
- aliases 86, 181
- conventions 957
- for CJK ideographs 951
- for control codes 185, 188
- in code charts 942
- matching 181
- character properties
- see* properties
- see also* individual properties, e.g. Combining Class
- character semantics 1, 78, 85–86, 983
- as Unicode design principle 18
- ASCII 264
- definition 85
- character sequences
- abstract *see* abstract character sequences
- canonical equivalent *see* canonical equivalent character sequences
- compatibility equivalent *see* compatibility equivalent character sequences
- conformance 79
- named 181
- character sequences, combining 104
- character shaping selectors (deprecated) 915
- character statistics 966
- character tabulation (U+0009) 901
- characters
- abstract *see* abstract characters
- arrangement in Unicode 45
- assigned 11, 30
- boundaries 217
- canonical decomposable *see* canonical decomposable characters
- classes 960
- code charts 937–955
- coded *see* encoded characters
- combining *see* combining characters
- compatibility decomposable *see* compatibility decomposable characters
- composite *see* decomposable characters
- concept of 15, 60
- conformance definitions 88–91
- 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, 67, 265, 899–916
- glyphs, relationship to 15
- graphic 30
- identity (definition) 85
- ignored in processing 248–254
- interpretation 78
- layout control 67, 903–913
- modification 79
- names list 938–950
- names *see* character names
- not encoded in Unicode 3
- number encoded in Version 14.0 3
- precomposed *see* decomposable characters
- properties *see* properties
- semantics *see* character semantics
- special 66, 899–935
- supplementary *see* supplementary characters

- transcoding 196–197
- unsupported 201
- characters, not glyphs
 - in spoofing 246
 - Unicode principle 15
- charsets
 - IANA registered names 40
- Cherokee 806
- Chinese 734–735
 - Cantonese 735
 - Hakka 753
 - Mandarin 735
 - Minnan (Hokkien/Fujian, incl. Taiwanese) . 753
 - simplified and traditional 734
- Chorasmian 432
- Chu hán 733
- Chu Nôm 990
- citations for
 - properties 75
 - Unicode algorithms 76
 - Unicode Standard 74
- CJK ideographs 260, 728–744
 - accounting numbers 176
 - CJK Compatibility Ideographs 743
 - CJK Compatibility Supplement 744
 - CJK Strokes 746, 993
 - CJK Unified Ideographs 728–743
 - CJK Unified Ideographs Extension B 743
 - code charts 951
 - compatibility ideographs in Plane 2 51
 - component structure 738
 - encoding blocks 729
 - ideographic description sequences 748–751
 - ideographic variation mark (U+303E) 750
 - Kangxi radicals 742, 745–746
 - names 951
 - numbers 849
 - numeric values 176, 205
 - order of encoding 740
 - radicals 745–746
 - source standards 992
 - unknown or unavailable 286
 - Vietnamese 726
- CJK Miscellaneous Area 49
- CJK punctuation and symbols 284
 - compatibility forms 287
 - overscores and underscores 287
 - quotation marks 272
 - sesame dots 286
 - vertical forms 287
- CJK Unified Ideographs extensions 730–731
- CJK-JRG (Chinese/Japanese/Korean Joint Research Group) 988
- CJKV Ideographs Area 49
- cluster boundaries 217
- code charts 937–955
 - representative glyphs 938
- code point sequences
 - notation 958
- code points 7, 29
 - assigned 11, 30
 - assignment 45
 - categories 30
 - default ignorable 201, 253
 - definition 88
 - designated 30
 - notation 957
 - number in Unicode Standard 1
 - private-use *see* private-use code points
 - reserved *see* reserved code points
 - semantics 32
 - surrogate *see* surrogates
 - unassigned code points
 - undesignated 30
- code positions *see* code points
- code set independence 18
- code unit sequences
 - definition 118
 - ill-formed (definition) 120
 - notation 958
 - well-formed (definition) 120
- code units
 - definition 118
 - isolated 117
- code values *see* code units
- coded character representations
 - see* coded character sequences
- coded character sequences
 - definition 90
- coded characters *see* encoded characters
- codespace *see* Unicode codespace
- coeng 669, 672
- Collation Algorithm, Unicode (UCA) 12
- collation *see* sorting
- collation tables 196
- combining character sequences 55, 104
 - defective 223
 - definition 106
 - Latin 291
 - line breaking 219
 - matching 219
 - order of base character and marks 220, 329
 - rendering 219
 - selection 217
 - truncation 220–221
- combining characters 54–59, 108–113, 219–227
 - blocking reordering 910
 - canonical ordering 61, 136, 168

- combining marks 329–330
- definition 104
- dependence 329
- display order 56
- keyboard input 220
- ligatures 59
- multiple 56
- multiple base characters 59
- normalization of 206
- ordering conventions 55
- rendering of marks 222–227
- reordrant 169
- script-specific 55
- split 169
- strikethrough 170
- subjoined 170
- typographical interaction 56, 168
- vertical stacking 56
- see also* diacritics
- Combining Class (normative property) 168
- combining classes 134, 168, 225–226
- class zero characters 168
- definition 134
- combining grapheme joiner (U+034F) 910
- combining half marks 191, 337
- combining marks *see* combining characters
- comma below 292
- Compatibility and Specials Area 26, 49
- compatibility characters 22
- compatibility composite characters 27
- see* compatibility decomposable characters
- compatibility decomposable characters 26
- definition 114
- compatibility decomposition 62
- definition 114
- compatibility decomposition mappings 114
- compatibility equivalence
- definition 115
- compatibility equivalent character sequences
- conformance 79
- compatibility mappings
- 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) 98
- compression 208
- see also* UTS #6, A Standard Compression Scheme for Unicode (SCSU)
- conferences 966
- conformance 71–157
- definitions 85–91
- examples 68
- ISO/IEC 10646 implementations 983
- requirements 77–82
- confusables 245
- conjunct consonants
- Indic 217, 463
- Myanmar 663
- selection of clusters 217
- contextual shaping
- apostrophe 274
- Arabic 373
- not used for Hebrew final forms 368
- Syriac 403
- contour tones 327
- control codes 31, 67, 900
- graphics for 871
- names 188
- properties 901
- semantics 32, 901
- specified in Unicode 901
- control sequences 900
- conversion of characters 196–197
- convertibility
- as Unicode design principle 24
- Coptic 309, 312–314
- Coptic E pact numbers 854
- corporate use subarea 921
- corrigenda 74
- CR (U+000D carriage return) 209, 901
- Creative Commons 897
- CRLF (carriage return and line feed) 209
- Croatian 296
- digraphs 296
- culturally expected sorting 12, 230
- Cuneiform
- Old Persian 443
- Sumero-Akkadian 438–441
- Ugaritic 442
- Cuneiform and Hieroglyphic Area 50
- Cuneiform and Hieroglyphs 437–454
- currency symbols block 835–838
- currency symbols encoded in other blocks .. 836
- currency symbols, other 837
- dollar sign, form and usage 836
- euro sign 837
- lari sign 837
- lira sign, compatibility usage 836
- lira sign, Turkish 837
- peso signs, usage 836
- ruble sign 837
- rupee signs, Indian, usage 837
- yen and yuan signs, usage 836

- cursive joining 905–909
 - Arabic 381–388
 - control characters for 192, 375–376, 553, 904
 - Mandaic 412
 - Mongolian 552–553
 - N’Ko 793
 - Phags-pa 600
 - Syriac 403–407
 - transparency 908
 - cursive scripts 365
 - Cypriot 346
 - see also* Linear B
 - Cypro-Minoan 347
 - Cyrillic 315–318
 - Czech 296
- D**
- danda, in Devanagari block 477
 - Danish 295
 - dashes 267
 - Database, Unicode Character
 - see* Unicode Character Database (UCD)
 - dead consonants, Indic 460
 - dead keys 220
 - decomposable characters 62
 - definition 114
 - normalization of 206
 - decomposition 62, 114–116
 - canonical *see* canonical decomposition
 - compatibility *see* compatibility decomposition
 - definition 114
 - in normalization 206
 - mapping, definition 114
 - mappings noted in code charts 946
 - default case
 - algorithms 83, 151–157
 - conversion 153
 - detection 155
 - folding 154
 - default caseless matching 156
 - default grapheme clusters 217
 - see also* UAX #29, Unicode Text Segmentation
 - Default Ignorable Code Point (property) 253
 - default ignorable code points 201, 253
 - default property values
 - definition 95
 - defective combining character sequences 223
 - definition 106
 - dependent vowel signs
 - Indic 459
 - Khmer 674
 - Philippine scripts 702
 - deprecated characters 72, 941
 - alternate format 193, 915–916
 - definition 90
 - Derived Age (property) 202
 - derived properties
 - definition 102
 - DerivedCoreProperties.txt 152, 164, 253
 - DerivedNormalizationProps.txt 242
 - Deseret 811–813
 - design goals of Unicode 4
 - design principles of Unicode 14–24
 - designated code points 30
 - Devanagari 457–482
 - Dhivehi 529
 - diacritics 54, 329
 - alternative glyphs 291, 329
 - Czech 292
 - display in isolation 59, 267, 330
 - double 112, 191, 331
 - German dialectology 335
 - Greek 305–306, 309
 - Latin 291–294
 - Latvian 292
 - mathematical 845
 - on i and j 293
 - rendering 222–227
 - Slovak 292
 - spacing clones of 327, 331
 - symbol 54, 336
 - see also* combining characters
 - dictionary symbols 883
 - digit form names 377
 - digits 205
 - Arabic 849
 - Arabic-Indic 377–378
 - compatibility 849
 - decimal 175
 - glyph variants 851
 - hexadecimal 849
 - Myanmar 849
 - national shapes 916
 - Shan 849
 - superscript and subscript 850
 - Tai Laing 849
 - Tai Tham 849
 - digraphs 296, 299, 301
 - dingbats 885–886
 - directionality 20, 52
 - East Asian scripts 726
 - Middle Eastern scripts 365
 - Mongolian 551
 - musical symbols 819
 - normative property 171
 - Ogham 362

Old Italic 351
 Philippine scripts 703
 Runic 354
 discussion list for Unicode 966
 Dives Akuru 646–647
 Dogra 651–652
 Dogri 478
 Domino Tiles 888
 dotless i 238, 293
 dotted circle
 in code charts 105, 330
 in fallback rendering 222
 to indicate diacritic 54
 to indicate vowel sign placement 56
 double diacritics 112, 191, 331
 Duployan 829–830
 Dutch 295, 296
 dynamic composition
 as Unicode design principle 23
 Dzongkha 538

E

East Asian scripts 725–777
 writing direction 52
 see also CJK ideographs
 Eastern Arabic-Indic digits 377
 EBCDIC
 newline function 210
 editing, text boundaries for 217–218
 efficiency
 as Unicode design principle 15
 Egyptian hieroglyphs 444–450
 format controls 446–450
 Elbasan 358
 ellipsis 276–277
 Elymaic 433
 e-mail discussion list for Unicode 966
 emoji 880, 881, 966
 animal symbols 884
 charts 966
 cultural symbols 884
 zodiacal symbols 884
 emoji modifiers 884
 emoticons 885
 Enclosed Alphanumerics 895
 enclosing marks 337
 definition 105
 encoded characters 7, 29
 allocation 43–51
 definition 90
 encoding form conversion
 definition 125

encoding forms 33–38
 ISO/IEC 10646 definitions 978
 encoding forms, Unicode
 see Unicode encoding forms
 encoding model for Unicode characters 33, 41
 see also UTR #17, Unicode Character Encoding Model
 encoding schemes 39–42
 encoding schemes, Unicode
 see Unicode encoding schemes
 endian ordering
 see byte order mark (BOM) (U+FEFF)
 end-user subarea 922
 English 295
 equivalent sequences 206
 as Unicode design principle 23
 case-insensitivity 231, 240
 combining characters in matching 219
 conformance 80
 Hangul syllables 764
 in sorting and searching 230
 language-specific 116
 security implications 245
 see also canonical equivalence
 see also compatibility equivalence
 see also encoding forms, encoding schemes
 errata xxv, 74, 968
 escape sequences 900
 not used in Unicode 1, 4
 Esperanto 296
 Estonian 296
 Ethiopic 782–785
 Etruscan 350
 European scripts 289–339
 ancient 341–363
 eyelash-RA 469

F

fallback rendering 252
 of nonspacing marks 222
 FAQ (Frequently Asked Questions) 967
 Faroese 295
 Farsi 373, 375
 featural syllabaries 259
 FF (U+000C form feed) 209, 901
 file separator (U+001C) 901
 Finnish 295
 Finno-Ugric Transcription (FUT)
 see Uralic Phonetic Alphabet (UPA)
 fixed-width Unicode encoding form (UTF-32) ... 35,
 122
 flat tables 196
 Flemish 295

- fleurons 887
- fonts
 and Unicode characters 16
 for mathematical alphabets 844–846
 style variation for symbols 833
- form feed (U+000C) (FF) 209, 901
- format control characters 30, 67, 265, 899–916
 deprecated 915–916
 prefixed 193, 333
 stateful 913
- fraction characters 862
- fraction slash (U+2044) 275, 858
- French 296
- Frisian 296
- fullwidth forms in East Asian encodings 761
- futhark 353
- G**
- Garshuni 399
- Ge'ez 782
- General Category (normative property) 172
 list of values 172
- general punctuation 263–287
- General Scripts Area 49
- geometrical symbols 876–879
- Georgian 323–324
- German 295
- geta mark (U+3013) 286
- Glagolitic 320
- Glossary 967
- glyph selection tables 196
- glyphs 6, 15
 characters, relationship to 15
 diacritics alternative 291, 329
 Greek alternative 306–308
 Latin alternative 291
 mathematical alternative 864
 missing 253
 representative in code charts 938
 standardized variants 917
 symbols alternative 833
- golden numbers 355
- Gothic 357
- Grantha 642–645
- grapheme base 329
 definition 107
- grapheme clusters 11, 60
 see also UAX #29, Unicode Text Segmentation
 default 217
 definition 107
- grapheme extender
 definition 107
- grapheme joiner, combining (U+034F) 910
- graphic characters 30
- Greek 305–310
 acrophonic numerals 205, 310
 alternative glyphs 306–308
 ancient musical notation 826–828
 editorial marks 281
 letters as symbols 306–308, 865
 see also Cypriot, Linear B
- Greenlandic 296
- group separator (U+001D) 901
- guillemets 271
- Gujarati 495–496
- Gunjala Gondi 576–577
- Gurmukhi 490–494
- H**
- Hakka 753
- halant 455
 see also virama
- half marks, combining 191, 337
- half-consonants, Indic 464
- halfwidth forms in East Asian encodings 761
- hamza 392–394
- Han ideographs *see* CJK ideographs
- Han unification 736–743
 and language tags 215
 history 987–992
 language usage 733
 source separation rule 731, 737
 source standards 992
- hand symbols 884
- Hangul Area 49
- Hangul syllables 725, 762–765
 and combining marks 112
 canonical decomposition 143
 collation 764
 composition 145
 conjoining jamo 141–150
 equivalent sequences 764
 Hangul Compatibility Jamo 763
 Hangul Jamo 762–765
 Hangul Syllables block 764–765
 Johab set 764
 name generation 146
 normalization 763
 standard 142
- Hangzhou numerals 858
- Hanifi Rohingya 700
- Hanja *see* CJK ideographs
- Hanunóo 702
- Hanzi *see* CJK ideographs
- harakat 374
- hasant 483

- hash tables 197
 - Hatran 436
 - Hebrew 367–372
 - hentaigana 756–758
 - hieroglyphs
 - Anatolian 453–454
 - Egyptian 444–450
 - Meroitic 451–452
 - high surrogate
 - definition 117
 - high-surrogate code points 77, 923
 - high-surrogate code units 117
 - higher-level protocols
 - definition 91
 - Hindi 457
 - Hiragana 755
 - horizontal tab (U+0009) 901
 - HTML newline function 210
 - Hungarian 296
 - hyphenation 904
 - as a text process 10
 - hyphens 267, 904
- I**
- I Ching symbols 890
 - IANA charset names 40
 - Icelandic 295
 - identifiers 229
 - see also* UAX #31, Unicode Identifier and Pattern Syntax
 - Ideographic (informative property) 189
 - ideographic description sequences 749
 - Ideographic Rapporteur Group (IRG) 990
 - Ideographic Research Group (IRG) 991
 - ideographs *see also* CJK ideographs
 - IICore 731, 990
 - ill-formed
 - definition 120
 - Imperial Aramaic 422–423
 - implementation guidelines 195–255
 - in a Unicode encoding form
 - definition 121
 - in-band mechanisms 935
 - India
 - Official scripts 455–526
 - Indian rupee signs, usage 837
 - Indic scripts 455–526
 - principles, in terms of Devanagari 458–468
 - relation to ISCII standard 457
 - Indic Siyaq 856
 - Indonesia and Oceania
 - scripts of 701–723
 - Indonesian 295
 - industry character sets
 - covered in Unicode 3
 - information separators (U+001C..U+001F) 901
 - informative properties
 - definition 99
 - Inscriptional Pahlavi 428
 - Inscriptional Parthian 428
 - inside-out rule 222
 - interchange restrictions 31
 - International Phonetic Alphabet (IPA) 258, 298–299
 - Spacing Modifier Letters 326
 - see also* phonetic alphabets
 - internationalization 18
 - Internationalization & Unicode Conference 966
 - Internet protocols
 - UTF-8 as preferred encoding 37
 - Inuktitut 808
 - invisible operators 870
 - iota subscript 306
 - IPA *see* International Phonetic Alphabet
 - IRG (Ideographic Research Group) 991
 - Irish 295, 362
 - ISCII standard and Unicode 457
 - ISO/IEC 10646 969–983
 - conformance of Unicode implementations .. 983
 - encoding forms 978
 - synchrony with Unicode Standard 980
 - timeline compared to Unicode versions 972
 - Italian 295
 - ITC Zapf Dingbats 885
 - IUC *see* Internationalization & Unicode Conference
- J**
- jamos *see* Hangul syllables
 - Japanese 725
 - Japanese era names 896
 - Javanese 713–716
 - Jawi 394
 - jihvamuliya 482, 621
 - Johab 764
 - joiners 375
 - combining grapheme joiner (U+034F) 910
 - word joiner (U+2060) 903
 - zero width joiner (U+200D) 375–376, 906
 - justification 224
- K**
- Kaithi 618–620
 - Kana (Hiragana and Katakana) 755–756
 - Kanbun 744
 - Kangxi radicals 742, 745–746
 - Kanji *see* CJK ideographs
 - Kannada 514–517

Kashmiri 479
 Katakana 755–756
 Kawi 707, 709
 Kayah Li 692
 KC (normalization form)
 see Normalization Form KC
 KD (normalization form)
 see Normalization Form KD
 keytop labels 871
 Khamti Shan 666
 Kharoshthi 587–588
 Khitan Small Script 778–779
 Khmer 669–680
 characters not recommended 677
 syllable components, order of 678
 Khojki 629–630
 Khudawadi 631–632
 killer 260
 Batak 718
 Brahmi 583
 Meetei Mayek 563
 Myanmar (asat) 664
 see also virama
 Konkani 477
 Kurdish 394

L

Ladino 367
 language tags 215, 931–935
 and Han unification 215
 use strongly discouraged 931, 934
 Lanna 683
 Lao 659–661
 last-resort glyphs 253
 Latin 291–304
 alternative glyphs 291
 Basic Latin 295
 encoding blocks 44
 IPA Extensions 298–299
 Latin Extended Additional 301–304
 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 300–304
 Latin Extended-F 304
 Latin Extended-G 304
 Latvian 296, 303
 cedilla 292
 layout control characters 67, 903–913

leading surrogates
 see high-surrogate code units
 legibility criterion for plain text 19
 Lepcha 570–572
 letter spacing 904
 letterlike symbols 839–846
 LF (U+000A line feed) 209, 901
 ligatures 905–909
 Arabic 384–385
 combining characters on 59
 control characters for 192
 for nonspacing marks 226
 Latin 301
 selection 218
 Syriac 407
 Limbu 559–562
 line breaking 209–213, 903–905
 control characters 191
 in South Asian scripts 657, 665, 680
 recommendations 211
 see also UAX #14, Unicode Line Breaking Algorithm
 line feed (U+000A) (LF) 209, 901
 line separator (U+2028) (LS) 209, 905
 line tabulation (U+000B) (VT) 901
 Linear A 343
 Linear B 344–345
 see also Cypriot
 linear boundaries 218
 Lisu 770–772
 Lithuanian 296
 little-endian 39
 definition 81
 logical order
 as Unicode design principle 19
 exceptions to 169
 logograph 260
 logosyllabaries 260
 low surrogate
 definition 117
 low-surrogate code points 77, 923
 low-surrogate code units 117
 lowercase 164, 236, 289
 LS (U+2028 line separator) 209, 905
 Lycian 348
 Lydian 348

M

MacOS newline function 210
 Mahajani 627–628
 Mahjong Tiles 887
 mail discussion list for Unicode 966
 Maithili 478

major version 73
Makasar 722–723
Malay 295
Malay, Patani 658
Malayalam 518–526
 Suriyani 408, 519
Maltese 296
Manchu 550
Mandaic 411–413
Mandarin 735
Manden 790
Manichaeen 424–427
map symbols 883
mapping tables *see* tables of character data
Marathi 457, 469, 476
Marchen 602
markup languages
 and Unicode conformance 935
 line breaking 209
Masaram Gondi 574–575
Mathematical (informative property) 862
mathematical expression format characters 193
 see also UTR #25, Unicode Support for Mathematics
mathematical symbols 862–869
 alphabets 841–846
 alphanumeric 840–846
 fonts 844–846
 format characters 870
 fragments for typesetting 872
 invisible operators 870
 operators 863–866
 standardized variants 869
MathML 866
matras 168, 459
Medefaidrin 804
Meetei Mayek 563–564
Mende Kikakui 800–801
Meroitic
 cursive 451–452
 hieroglyphs 451–452
Miao 773–774
Middle Eastern scripts 365–530
 ancient 415–436
Min 735
Minnan (Hokkien/Fujian, incl. Taiwanese) 753
minor version 73
minus sign 865
 commercial (U+2052) 279
mirrored property
 see Bidi Mirrored (normative property)
mirroring of paired punctuation 269
Miscellaneous Symbols 882
missing glyphs 253

Modi 637–639
modifier letters 325–328
Modifier Letters, Spacing 301
Mongolian 549–558, 595
 writing direction 551
moon symbols 882
Mro 565
Multani 633
multibyte encodings
 compared to UTF-8 37
multistage tables 196
musical symbols 818–828
 ancient Greek 826–828
 Balinese 712
 Byzantine 824
 directionality 819
 Gregorian 823
 Kievan 823
 Persian 823
 Western 818–823
Myanmar 662–668
 digits 849
 Myanmar Extended-A 666
 Myanmar Extended-B 666

N

N’Ko 790–794
Nabataean 434
named character sequences 181
names, character *see* character names
namespace 87
Nandinagari 640–641
NEL (U+0085 next line) 209, 901
Nepali 457
neutral directional characters 171
New Tai Lue 683–685
Newa 535–537
newline function (NLF) 210, 902
newline guidelines 209–213
next line (U+0085) (NEL) 209, 901
NFC (Normalization Form C) 61
NFD (Normalization Form D) 61
NFKC (Normalization Form KC) 61
NFKD (Normalization Form KD) 61
NLF (newline function) 210, 902
no-break space (U+00A0) 903
 base for diacritic in isolation 59, 267, 330
no-break space, narrow (U+202F) 556
noncharacter code points *see* noncharacters
noncharacters 31, 924
 conformance 77
 definition 91
 handling 80

in code charts 941
 interchange restrictions 31
 semantics 32
 U+10FFFF (not a character code) 924
 U+FDD0..U+FDEF 31, 924
 U+FFFE (not a character code) 66, 925
 U+FFFF (not a character code) 31, 924
 nondecomposable characters 63
 non-joiner, zero width (U+200C) 375–376, 907
 nonlinear boundaries 218
 non-overlap principle in Unicode encoding forms 33
 nonspacing marks 329
 definition 105
 display in isolation 59, 267, 330
 positioning 226
 rendering 222–227
 see also combining characters
 see also diacritics
 normalization 61, 206–207
 and case operations 242
 canonical ordering algorithm 61, 136, 168
 conformance 82
 of private-use characters 921
 see also UAX #15, Unicode Normalization Forms
 stability 133
 Normalization Form C (NFC) 61
 Normalization Form D (NFD) 61
 Normalization Form KC (NFKC) 61
 Normalization Form KD (NFKD) 61
 normalization forms 133–140
 definition 139
 specification 135
 normative behaviors
 definition 85
 normative properties
 definition 97
 list 98
 may change 97
 Norwegian 295
 notational conventions 957–961
 notational systems 262, 815–832
 nukta 374, 396, 470
 null (U+0000)
 as Unicode string terminator 902
 number forms
 CJK ideographs 205
 numbers
 Coptic Epect 854
 handling 205
 ideographic accounting 176
 numerals 847–859
 acrophonic 310
 Chinese counting rods 860
 Coptic 314

Cuneiform 441
 Ethiopic 784
 Greek acrophonic 205
 Hangzhou 858
 Meroitic cursive 452
 old-style 276
 Roman 205, 862
 Rumi 855
 Suzhou-style 858
 numeric separators 278
 numeric shape selectors (deprecated) 916
 Numeric Type (normative property) 175
 Numeric Value (normative property) 175
 numero sign (U+2116) 839
 Nüshu 769
 Nyiakeng Puachue Hmong 697–698

O

object replacement character (U+FFFC) 930
 octet 959
 Ogham 362
 Ol Chiki 567–568
 Old Church Slavonic 315
 Old Hungarian 356
 Old Italic 350–352
 Old North Arabian 417
 Old Permic 361
 Old Persian 443
 Old Sogdian 609
 Old South Arabian 418–419
 Old Turkic 608
 Old Uyghur 611–612
 old-style numerals 276
 Oriya 497–500
 ornamental dingbats 886
 Oromo 782
 Osage 810
 Osmania 786
 Ottoman Siyaq 856
 out-of-band mechanisms 935
 overlapping encodings 33
 overscores 275

P

Pahawh Hmong 695–696
 Pahlavi, Inscriptional 428
 Pahlavi, Psalter 429
 Palmyrene 435
 Panjabi 490
 paragraph or section marks 278
 paragraph separator (U+2029) (PS) 209, 905
 Parthian, Inscriptional 428
 Pashto 373

- Patani Malay 658
- Pau Cin Hau 699
- Persian 373, 375
- Phags-pa 595–601
- Phaistos Disc symbols 891
- Phake 668
- Philippine scripts 702–704
- Phoenician 420
- phonemes 261
- phonetic alphabets 258
- IPA Extensions 298–299
 - Phonetic Extensions 300–304
 - Spacing Modifier Letters 326–328
 - Uralic Phonetic Alphabet (UPA) 279, 300
 - see also* International Phonetic Alphabet (IPA)
- Pinyin 295
- pipeline table
- proposed new characters 967
- pivot code, Unicode as 196
- plain text
- as Unicode design principle 18
 - legibility criterion 19
- planes of Unicode codespace 43
- Plane 0 (BMP) 43
 - Plane 1 (SMP) 43, 50
 - Plane 14 (SSP) 44
 - Plane 2 (SIP) 43, 51
 - Plane 3 (TIP) 43, 51
 - Planes 15–16 (Private Use) 51, 922
- Playing Cards 888
- points, Hebrew pronunciation marks 367
- policies of the Unicode Consortium 967
- Polish 296
- Portuguese 295
- precomposed characters
- see* decomposable characters
 - compatibility *see* compatibility decomposable characters
- prefixed format control characters 193
- prepended concatenation marks 254, 333
- Private Use Area (PUA) 49, 921
- Private Use planes 44, 51, 922
- private-use characters
- properties 920
 - semantics 32
- private-use code points 31, 201
- conformance 78
 - definition 103
 - high surrogates 923
- properties 18, 93–103, 159–194
- aliases 162
 - aliases (definition) 102
 - and Unicode algorithms 98
 - data tables 196
 - derived *see* derived properties
 - in Unicode Character Database (UCD) 45
 - informative *see* informative properties
 - normative references to 75, 82
 - normative *see* normative properties
 - of control codes 901
 - provisional *see* provisional properties
 - simple *see* simple properties
 - see also* individual properties, e.g. combining classes
- property values
- aliases 162
 - aliases (definition) 103
 - default 95
 - default (definition) 95
 - normative references to 82
- PropertyAliases.txt 103, 960
- PropertyValueAliases.txt 103, 960
- PropList.txt 166
- Provençal 296
- provisional properties
- definition 100
- PS (U+2029 paragraph separator) 209, 905
- Psalter Pahlavi 429
- PUA (Private Use Area) 49, 921
- punctuation 263–287
- blocks containing 257
 - CJK 284
 - doubled 276
 - ideographic 747
 - in bidirectional text 263
 - paired 269
 - small form variants 287
 - typographic forms 263
 - vertical forms 287
- Punctuation and Symbols Area 49
- Punjabi 490
- ## Q
- quotation marks 270–274
- East Asian 273
 - European 271
- ## R
- radicals, Kangxi and other CJK 745–746
- radical-stroke index 742
- record separator (U+001E) 901
- recycling symbols 883
- references 967
- referencing 82
- properties 75
 - Unicode algorithms 76
 - Unicode Standard 74

- regional indicator symbols 896
 - regular expressions 214
 - and line breaking 209
 - see also* UTS #18, Unicode Regular Expressions
 - Rejang 717
 - rendering of text 6, 10, 17
 - fallback 252
 - unsupported characters 201
 - repertoire of abstract characters 29
 - reph 468, 472, 512
 - replacement character (U+FFFD) 42, 67, 81, 930
 - reserved code points 30, 201
 - definition 91
 - in code charts 941
 - 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 52
 - East Asian scripts 726
 - Middle Eastern scripts 365
 - roadmap for script additions 45, 967
 - Roman numerals 205, 862
 - Romanian 296
 - comma below 293
 - Romany 296
 - Rong 570–572
 - Rumi numeral symbols 855
 - Runic 353–355
 - Russian 315
- S**
- Samaritan 409–410
 - Sami 296
 - Sanskrit 457
 - Saurashtra 573
 - scalar values, Unicode
 - see* Unicode scalar values
 - scripts
 - in Unicode Standard 3
 - roadmap for future additions 45, 967
 - 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 245
 - self-synchronization of encoding forms 34
 - semantics
 - see* character semantics
 - sequences
 - notation 958
 - Serbian
 - corresponding digraphs in Croatian 296
 - Shan 681
 - digits 849
 - Sharada 621–622
 - Shavian 363, 770
 - Show Hidden 79, 222, 252, 918
 - SHY (U+00AD soft hyphen) 904
 - Sibe 551
 - Siddham 625–626
 - signature for Unicode data 66, 926–928
 - simple properties
 - definition 102
 - simplified Chinese 734
 - Sindhi 373, 477
 - Sinhala 531–534
 - Sinological dot 303
 - SIP (Supplementary Ideographic Plane) 43, 51
 - Siyaq Numbers 855
 - Indic 855
 - slash, fraction (U+2044) 275
 - Slovak 296
 - Slovenian 296
 - small letters 164, 236, 289
 - SMP (Supplementary Multilingual Plane) 43, 50
 - soft hyphen (U+00AD) (SHY) 904
 - Sogdian 610
 - Somali 786
 - Sora Sompeng 650
 - Sorbian 296
 - sorting 12, 230
 - and combining grapheme joiner 911
 - as a text process 10
 - case-insensitive 231
 - culturally expected 12, 230
 - language-insensitive 230
 - see also* Unicode Collation Algorithm (UCA)
 - source separation rule 731, 737
 - South and Central Asian scripts
 - Ancient 581–610
 - Other historic 613–652
 - Other modern 527–577
 - South Asian scripts 455–562
 - Southeast Asian scripts 653–700
 - Soyombo 606–607
 - space (U+0020)
 - base for diacritic in isolation 59, 267, 330
 - space characters 266, 903–905
 - graphics for 871

- space, zero width (U+200B) 266
- spacing clones of diacritics 327, 331
- spacing marks 329
 - definition 106
- Spacing Modifier Letters 326–328
- Spanish 295
- special characters 66, 899–935
- SpecialCasing.txt 151, 166
- Specials 926–930
- spell-checking
 - as a text process 11
- spellings, alternative
 - see* equivalent sequences
- spoofing 245
- SSP (Supplementary Special-purpose Plane) 44
- stability 100, 161
 - as Unicode design principle 23
- stacked boundaries 217
- stacking sequences 56
 - nondefault 58
- standardized variants 554, 917
 - in the code charts 948
 - mathematical symbols 869
- StandardizedVariants.txt 554, 869
- standards coverage 3
- starters 135
- stateful encoding
 - not used in Unicode 4
 - paired format controls 913
- string comparison 12
- string literals, Unicode
 - code point notation `\u1234` 960
- strings, Unicode 42, 119
 - null termination 902
- strong directional characters 171
- styled text 18
- sublinear searching 232
- subsets, supported 70
 - conformance 78
 - ISO/IEC 10646 specification for 981
- substitution character
 - see* replacement character
- Sumero-Akkadian 438–441
- Sundanese 720–721
- superscripts 327
 - and subscripts 860
- supplementary characters
 - in UTF-16 strings 42
 - tables for 197
- Supplementary General Scripts Area 49
- Supplementary Ideographic Plane (SIP) 43, 51
- Supplementary Multilingual Plane (SMP) 43, 50
- supplementary planes
 - representation in UTF-16 36
 - representation in UTF-8 37
- Supplementary Private Use Areas 51, 922
- Supplementary Special-purpose Plane (SSP) 44
- supported subsets 70
 - conformance 78
- supralineation 313
- surrogate code points
 - see* surrogates
- surrogate pairs 36, 123
 - definition 117
 - processing 203–204
- surrogates 31, 117, 923
 - interchange restrictions 31
 - isolated surrogates, handling 42
 - isolated surrogates, ill-formed 123
 - isolated surrogates, uninterpreted 117
 - support levels 203
- Surrogates Area 49, 923
- Sutton SignWriting 831–832
- Suzhou-style numerals 858
- svasti signs 545
- Swahili 295
- Swedish 295
- syllabaries 259
 - alphabetic property 189
 - featural 259
- Syloti Nagri 616–617
- symbols 833–898
 - animal 884
 - appearance variation 833
 - arrows 868–869
 - box drawing 876
 - cultural 884
 - currency symbols block 835–838
 - dictionary 883
 - dingbats 885–886
 - emoji 880, 881, 896
 - Enclosed Alphanumerics 895
 - fragments for mathematical typesetting 872
 - game 884
 - gender 883
 - genealogical 884
 - geometrical 876–879
 - hand 884
 - Khmer lunar calendar 680
 - letterlike 839–846
 - map 883
 - mathematical 862–869
 - mathematical alphanumeric 840–846
 - miscellaneous 882
 - musical 818–828
 - numerals 847–859

recycling 883
 regional indicator 896
 technical 871–875
 weather 882
 zodiacal 884
 symmetric swapping format characters 915
 Syriac 399–408

T

tab (U+0009 character tabulation) 901
 tab, vertical (U+000B) 209, 901
 tables of character data 196–197
 optimization 197
 supplementary characters 197
 tag characters 931–935
 Tagalog 702
 Tagbanwa 702
 tags, language 215, 931–935
 use strongly discouraged 934
 Tai Laing
 digits 849
 Tai Le 681–682
 Tai Tham 686–688
 digits 849
 Tai Viet 689–691
 Tai Xuan Jing symbols 890
 Takri 623–624
 Tamil 501–510
 Tangsa 580
 Tangut 775–777
 code charts 953
 components 776–777
 radicals 776
 tashkil 374
 tashkil, harakat, points 376
 TCHAR in Win32 API 200
 Technical Reports (UTR) 965
 Technical Standards (UTS) xxiv, 965
 abstracts 966
 technical symbols 871–875
 Telugu 511–513
 terminal emulation 834
 Tertiary Ideographic Plane (TIP) 43, 51
 text boundaries 60, 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
 text processes 6, 10–13
 text rendering 6, 10, 17
 text selection, boundaries for 217–218

Thaana 529–530
 Thai 655–658
 Tibetan 538–548
 Tifinagh 787
 Tigre 782
 tilde (U+007E) 279
 TIP (Tertiary Ideographic Plane) 43, 51
 Tirhuta 634–636
 titlecase 164, 236
 Todo 550
 tone letters 327–328
 tone marks
 Bopomofo spacing 752, 753
 Chinantec 328
 Chinese 328
 Tai Le 681
 Thai 655
 Vietnamese 294
 Toto 579
 traditional Chinese 734
 traffic signs 883
 trailing surrogates
 see low-surrogate code units
 transcoding 196–197
 tables 196
 Transport and Map Symbols 885
 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 837
 two-stage tables 197

U

U+ notation 960
 U+10FFFF (not a character code) 924
 U+FEFF (BOM) 926–928
 U+FFFE (not a character code) 925
 U+FFFF (not a character code) 924
 UAX (Unicode Standard Annex) xxiii, 965
 as component of Unicode Standard 77
 conformance 83
 list of 83
 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-2 978

- UCS-4 978
- Ugaritic 442
- Ukrainian 315
- unassigned code points 30, 77, 201
 - defined as reserved code points 91
 - handling 72
 - properties of 95
 - semantics 77
 - see also* reserved code points
- underscores 275
- undesignated code points 30
- Unicode 1.0 Name (informative property) 188
- Unicode algorithms
 - and properties 98
 - conformance 82
 - definition 91
 - normative references to 76, 82
- Unicode Bidirectional Algorithm 21, 52
 - see also* UAX #9, Unicode Bidirectional Algorithm
- Unicode Character Database (UCD) .. xxii, 161, 967
 - as component of Unicode Standard 77
 - changes 72
 - properties in 45
- Unicode character encoding model 33, 41
 - see also* UTR #17, Unicode Character Encoding Model
- Unicode character literals
 - code point notation U+ 960
- Unicode codespace
 - definition 88
 - planes 43
 - size 1, 29
- Unicode Collation Algorithm (UCA) 12
- Unicode conferences 966
- Unicode Consortium 964
 - addresses 968
 - Consortium membership in standards bodies 964
 - e-mail discussion list 966
 - membership 964
 - policies 967
 - website 966
- Unicode data signature 66, 926–928
- Unicode data types 199–200
 - for C 199–200
- Unicode encoding forms 118–125
 - conformance 34, 80
 - definition 119
 - fixed-width (UTF-32) 35, 122
 - signatures 927, 928
 - variable-width 36, 37, 123
 - see also* encoding forms
- Unicode encoding schemes
 - conformance 129–132
 - definition 129
 - endian ordering 39
 - see also* encoding schemes
- Unicode escape sequence notation \u1234 960
- Unicode scalar values
 - definition 118
- Unicode security 245
 - see also* UTS #39, Unicode Security Mechanisms
- Unicode Standard
 - allocation of encoded characters 43–51
 - architecture 10–13
 - areas 44
 - benefits 1
 - blocks 257
 - code charts 937–955
 - components 77
 - conformance 71–157
 - conformance of ISO/IEC 10646 implementations. 983
 - corrections 74
 - definitions for conformance 85–91
 - design goals 4
 - design principles 14–24
 - errata 74, 968
 - normative references to 74, 82
 - number of characters 3
 - number of code points 1, 29
 - script coverage 3
 - security issues 245
 - synchrony with ISO/IEC 10646 980
 - updates 968
 - versions *see* versions of the Unicode Standard
- Unicode Standard Annexes (UAX) xxiii, 965
 - as components of Unicode Standard 77
 - conformance 83
 - list of 83
- Unicode string literals
 - code point notation \u1234 960
- Unicode strings 42
 - definition 119
- Unicode Technical Committee (UTC) 964
- Unicode Technical Reports (UTR) 965
- Unicode Technical Standards (UTS) xxiv, 965
 - abstracts 966
- UnicodeData.txt 151, 166
- unification
 - as Unicode design principle 21
 - see also* Han unification
- Unified Repertoire and Ordering (URO) ... 737, 989
 - see also* Han unification
- Unihan Database 161, 741, 742, 967, 990
- Unihan.zip 100, 161
- UnihanCore2020 732, 990
- unit separator (U+001F) 901

- Universal Character Set (UCS)
 - see* ISO/IEC 10646
 - universality
 - as Unicode design principle 14
 - Unix
 - newline function 210
 - UTF-8 in 18
 - unsupported characters 201
 - upadhmānīya 482, 621
 - update version 73
 - uppercase 164, 236, 289
 - Uralic Phonetic Alphabet (UPA) 279, 300
 - Urdu 373
 - URO (Unified Repertoire and Ordering) .. 737, 989
 - see also* Han unification
 - UTF, Unicode Transformation Formats 33, 119
 - as encoding form or scheme 132
 - binary comparison and sort order differences 231, 233
 - in APIs 200
 - UTF-16 36, 123, 979
 - binary comparison and sort order caution 36
 - bit distribution (table) 123
 - BOM in 130, 926
 - encoding form (definition) 123
 - encoding scheme (definition) 130
 - encoding schemes 39
 - in ISO/IEC 10646 979
 - in UTF-8 order 234
 - surrogates and string handling 42, 203
 - UTF-16BE (Big-endian) 927
 - encoding scheme 40
 - encoding scheme (definition) 129
 - UTF-16LE (Little-endian) 927
 - encoding scheme 40
 - encoding scheme (definition) 129
 - UTF-32 35, 122
 - BOM in 131
 - encoding form (definition) 122
 - encoding scheme (definition) 131
 - encoding schemes 39
 - UTF-32BE (Big-endian)
 - encoding scheme 40
 - encoding scheme (definition) 131
 - UTF-32LE (Little-endian)
 - encoding scheme 40
 - encoding scheme (definition) 131
 - UTF-8 37, 123, 979
 - ASCII transparency 37
 - binary comparison and sort order 38
 - bit distribution (table) 124
 - BOM in 129, 132, 927
 - byte ranges 124
 - compared to multibyte encodings 37
 - encoding form (definition) 123
 - encoding scheme 39
 - encoding scheme (definition) 129
 - in Unix 18
 - in UTF-16 order 233
 - non-shortest form is invalid 123, 245
 - preferred encoding for Internet protocols 37
 - security and 245
 - signature 129, 132, 927
 - UTR (Unicode Technical Report) 965
 - UTS (Unicode Technical Standard) xxiv, 965
 - abstracts 966
 - Uyghur 373, 595
- ## V
- Vai 795–796
 - valid (synonym for well-formed) 121
 - variable-width Unicode encoding form 36, 37, 123
 - variants
 - compatibility 26
 - fullwidth and halfwidth 287
 - mathematical symbols 869
 - small form 287
 - standardized 917
 - variation selectors 194, 917
 - ideographic variation mark (U+303E) 750
 - Mongolian free variation selectors 554
 - variation sequences 917
 - for Phags-pa 599–601
 - Version 14.0 77
 - number of characters 3
 - versions of the Unicode Standard .xxiii, 72, 968, 985–986
 - backward compatibility 72
 - compared to ISO/IEC 10646 editions 985
 - content 73
 - interaction in implementations 202
 - numbering 73
 - property changes 72
 - stability 72
 - updates 968
 - vertical tab (U+000B) 209, 901
 - vertical text 52, 264, 287
 - East Asian scripts 726
 - Mongolian 551
 - Vietnamese 294, 301
 - ideographs 726
 - virama 260, 455
 - definition 460
 - Kharoshthi 591
 - Khmer 672
 - Myanmar 663

Philippine scripts 703
 virama-like characters 192
 visual order used for Thai and Lao 21
 Vithkuqi 360
 vowel harmony
 Mongolian 555
 vowel marks, Middle Eastern scripts 366
 vowel separator
 Mongolian 556
 vowel signs
 Indic 56, 459
 Khmer 674
 Philippine scripts 702

W

Wancho 578
 Warang Citi 566
 wchar_t
 in C language 200
 weak directional characters 171
 weather symbols 882
 website, Unicode Consortium 966
 Weierstrass elliptic function symbol 840
 well-formed
 definition 120
 Welsh 296
 Where Is My Character? 968
 wide characters
 data type in C 200
 wiggly fence (U+29DB) 867
 Windows newline function 210
 word breaks 219, 903–905
 in South Asian scripts 657, 665, 680
 word joiner (U+2060) 903
 writing direction *see* directionality
 writing systems 258–262
 Wu (Shanghainese) 735

X

Xibe 551
 Xishuangbanna Dai 683

Y

Yezidi 414
 Yi 766–768
 Yiddish 367
 Yijing Hexagram Symbols 890
 ypogegrammeni 306

Z

Zanabazar Square 603–605
 Zapf Dingbats 885

zero extension relation among encodings 978
 zero width joiner (U+200D) 375–376, 906
 zero width no-break space (U+FEFF) 66, 82, 903
 initial 132, 927
 zero width non-joiner (U+200C) 375–376, 907
 zero width space (U+200B) 904
 for word breaks in South Asian scripts . 657, 665, 680
 zero-width space characters 904
 Znamenny Musical Notation 825
 ZWJ *see* zero width joiner (U+200D)
 ZWNBSP *see* zero width no-break space (U+FEFF)
 ZWNJ *see* zero width non-joiner (U+200C)
 ZWSP *see* zero width space (U+200B)

