

The Unicode® Standard

Version 13.0 – 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.

© 2020 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 13.0.

Includes index.

ISBN 978-1-936213-26-9 (<http://www.unicode.org/versions/Unicode13.0.0/>)

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

QA268.U545 2020

ISBN 978-1-936213-26-9

Published in Mountain View, CA

March 2020

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 309
- abjads 254, 357
- abstract character sequences
 - definition 88
- abstract characters 29
 - definition 88
- abugidas 255, 256, 445, 637
- accent marks *see* diacritics
- accented characters
 - encoding 12
 - Latin 287
 - normalization 204
- accounting numbers, ideographic 174
- acrophonic numerals 203, 306
- Adlam 784–785
- Aegean numbers 338
- Africa
 - scripts of 763–786
- Afrikaans 292
- Ahom 632–633
- Ainu 740
- Aiton 652
- Alchemical Symbols 868
- Algonquian 790
- Ali Gali 537
- aliases
 - character name 86, 179
 - informative 922
 - normative 923
 - property 160
 - property value 160
- allocation areas 44
- allocation of encoded characters 43–51
- Alphabetic (informative property) 187
- alphabets 254
 - European 285–334
 - mathematical 823–827
- alternate format characters (deprecated) ... 191, 896–897
- Americas
 - scripts of 787–795
- Amharic 764
- Anatolian hieroglyphs 443–444
- Ancient Symbols 872
- angle brackets (U+2329 and U+232A)
 - deprecated for technical publication 853
- Annexes, Unicode Standard (UAX) xxiv, 945
 - as components of Unicode Standard 77
 - conformance 83
 - list of 83
- annotation characters 909–911
 - use in plain text discouraged 910
- ANSI/ISO C
 - wchar_t and Unicode 198
- apostrophe (U+0027) 270
- Arabic 365–389
 - digits 830
- Arabic-Indic digits 369–370
 - signs used with 371
- ArabicShaping.txt 373, 378, 395
- Aramaic 412, 445, 538, 567, 573
- areas of the Unicode Standard 44
- ARIB 863
- Armenian 317–318
- arrows 849–850
- ASCII
 - characters with multiple semantics 260
 - transparency of UTF-8 37
 - Unicode modeled on 1
 - zero extension 198, 958
- Assamese 473
- assigned code points 11, 30
- Athapascan 790
- atomic character boundaries 216
- Avestan 420–421

B

Balinese 691–696
 Bamum 779–780
 Bangla 473–479
 base characters 325
 definition 104
 multiple 58
 ordered before combining marks 218, 325
 Basic Multilingual Plane (BMP) 1, 43
 allocation areas 48
 representation in UTF-16 36
 Basque 292
 Bassa Vah 781
 Batak 702–703
 Baybayin 686
 benefits of Unicode 1
 Bengali 473–479
 Bhaiksuki 579–580
 Bidi Class (normative property) 169
 Bidi Mirrored (normative property) 176
 Bidi Mirroring Glyph (informative property) 177
 BidiMirroring.txt 177
 Bidirectional Algorithm, Unicode 52, 82
 bidirectional ordering 20
 controls 893
 bidirectional text 52, 82
 Middle Eastern scripts 357
 nonspacing marks in 221
 punctuation in 259
 big-endian 39
 definition 81
 Bihari 469
 binary comparison and sort order
 caution for UTF-16 36
 UTF differences 229, 231
 UTF-8 38
 block 44, 88, 253, 917
 headers 929
 BMP *see* Basic Multilingual Plane
 BNF (Backus-Naur Form) 939
 BOCU-1 *see* UTN #6, BOCU-1
 MIME-Compatible Unicode Compression
 Bodhi 526
 Bodo 468
 BOM (U+FEFF) 39, 66, 128–131, 907–909
 Bopomofo 736–738
 boundaries, text 60, 188, 215–216, 226
 see also UAX #14, Unicode Line Breaking Algorithm
 see also UAX #29, Unicode Text Segmentation
 boustrophedon 52, 347
 box drawing symbols 857
 Brahmi 445, 567, 569–572, 573, 639

Braille 798–799
 Breton 292
 Buginese 689–690
 Buhid 686
 Bulgarian 311
 bullets 274
 numeric 831
 Burmese *see* Myanmar
 Byelorussian 311
 byte order mark (BOM) (U+FEFF) . 39, 66, 128–131,
 907–909
 byte ordering
 changing 79
 conformance 81
 byte serialization 39, 66
 Byzantine Musical Symbols 806

C

C language
 wchar_t and Unicode 198
 C0 and C1 control codes 31, 186, 882
 Cambodian *see* Khmer
 Canadian Aboriginal Syllabics 790–791
 candrabindu 470, 605
 canonical composite characters
 see canonical decomposable characters
 canonical composition algorithm 136
 canonical decomposable characters
 definition 116
 canonical decomposition 62
 definition 115
 mappings 114
 canonical equivalence
 definition 116
 nonspacing marks 223
 canonical equivalent character sequences
 conformance 79
 canonical mappings
 see canonical decomposition mappings
 canonical ordering algorithm 135
 canonical precomposed characters
 see canonical decomposable characters
 Cantonese 719
 capital letters 162, 234, 285
 Carian 341
 carriage return (U+000D) (CR) 207, 883
 carriage return and line feed (CRLF) 207
 case 293
 and text processes 12
 beyond ASCII 235
 camelcase 237
 case folding 238
 case operations (conformance) 83, 150–156

- case operations and normalization 240
- case operations, reversibility 237
- cased (definition) 151
- case-insensitive comparison 155, 229, 238
- casing context (definition) 151
- conversion 152
- detection 154
- European alphabets 285
- exceptional Latin pairs 289, 293
- Georgian 320
- lowercase 162, 234, 285
- mapping tables 194
- mappings 150, 164, 234–236
- mappings noted in code charts 926
- titlecase 162, 234
- Turkish I 236, 289
- uppercase 162, 234, 285
- see also* default case
- Case (normative property) 162, 234
- CaseFolding.txt 164, 238
- caseless letters 293
- Catalan 291
- Caucasian Albanian 352
- cedilla 288
- CEF *see* character encoding forms
- CES *see* character encoding schemes
- Chakma 557
- Cham 677–678
- character encoding forms (CEF) 33–38, 958
- 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 946
- character literals, Unicode
- code point notation U+ 940
- character names 86, 178–185, 962
- aliases 86, 179
- conventions 937
- for CJK ideographs 931
- for control codes 183, 186
- in code charts 922
- matching 179
- character properties
- see* properties
- see also* individual properties, e.g. Combining Class
- character semantics 1, 78, 85–86, 963
- as Unicode design principle 18
- ASCII 260
- 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 179
- character sequences, combining 104
- character shaping selectors (deprecated) 896
- character statistics 946
- character tabulation (U+0009) 883
- characters
- abstract *see* abstract characters
- arrangement in Unicode 45
- assigned 11, 30
- boundaries 215
- canonical decomposable *see* canonical decomposable characters
- classes 940
- code charts 917–935
- coded *see* encoded characters
- combining *see* combining characters
- compatibility decomposable *see* compatibility decomposable characters
- composite *see* decomposable characters
- concept of 15, 59
- conformance definitions 88–91
- confusable 243
- conversion 194–195
- 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 59
- format control 30, 67, 261, 881–897
- glyphs, relationship to 15
- graphic 30
- identity (definition) 85
- ignored in processing 246–251
- interpretation 78
- layout control 67, 885–895
- modification 79
- names list 918–930
- names *see* character names
- not encoded in Unicode 3
- number encoded in Version 13.0 3
- precomposed *see* decomposable characters
- properties *see* properties
- semantics *see* character semantics
- special 66, 881–916
- supplementary *see* supplementary characters
- transcoding 194–195

- unsupported 199
- characters, not glyphs
 - in spoofing 244
 - Unicode principle 15
- charsets
 - IANA registered names 40
- Cherokee 788
- Chinese 718–719
 - Cantonese 719
 - Hakka 737
 - Mandarin 719
 - Minnan (Hokkien/Fujian, incl. Taiwanese) .. 737
 - simplified and traditional 718
- Chorasmian 422
- Chu hán 717
- Chu Nôm 970
- citations for
 - properties 75
 - Unicode algorithms 76
 - Unicode Standard 74
- CJK ideographs 256, 712–728
 - accounting numbers 174
 - CJK Compatibility Ideographs 727
 - CJK Compatibility Supplement 727
 - CJK Strokes 730, 973
 - CJK Unified Ideographs 712–726
 - CJK Unified Ideographs Extension B 727
 - code charts 931
 - compatibility ideographs in Plane 2 51
 - component structure 722
 - encoding blocks 713
 - ideographic description sequences 732–735
 - ideographic variation mark (U+303E) 734
 - Kangxi radicals 725, 729–730
 - names 931
 - numbers 830
 - numeric values 174, 203
 - order of encoding 724
 - radicals 729–730
 - source standards 972
 - unknown or unavailable 282
 - Vietnamese 710
- CJK Miscellaneous Area 49
- CJK punctuation and symbols 280
 - compatibility forms 282
 - overscores and underscores 283
 - quotation marks 268
 - sesame dots 281
 - vertical forms 282
- CJK Unified Ideographs extensions 714–715
- CJK-JRG (Chinese/Japanese/Korean Joint Research Group) 968
- CJKV Ideographs Area 49
- cluster boundaries 215
- code charts 917–935
 - representative glyphs 918
- code point sequences
 - notation 938
- code points 7, 29
 - assigned 11, 30
 - assignment 45
 - categories 30
 - default ignorable 199, 250
 - definition 88
 - designated 30
 - notation 937
 - number in Unicode Standard 1
 - private-use *see* private-use code points
 - reserved *see* reserved code points
 - semantics 32
 - surrogate *see* surrogates
 - unassigned *see* unassigned code points
 - undesignated 30
- code positions *see* code points
- code set independence 18
- code unit sequences
 - definition 118
 - ill-formed (definition) 120
 - notation 938
 - 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 653, 656
- Collation Algorithm, Unicode (UCA) 12
- collation *see* sorting
- collation tables 194
- combining character sequences 55, 104
 - defective 221
 - definition 106
 - Latin 287
 - line breaking 217
 - matching 217
 - order of base character and marks 218, 325
 - rendering 217
 - selection 215
 - truncation 218–219
- combining characters 54–59, 108–113, 217–225
 - blocking reordering 892
 - canonical ordering 61, 135, 166
 - combining marks 325–326

- definition 104
- dependence 325
- display order 57
- keyboard input 218
- ligatures 58
- multiple 56
- multiple base characters 58
- normalization of 204
- ordering conventions 55
- rendering of marks 220–225
- reordrant 167
- script-specific 55
- split 167
- strikethrough 168
- subjoined 168
- typographical interaction 57, 166
- vertical stacking 57
- see also* diacritics
- Combining Class (normative property) 166
- combining classes 133, 166, 223–224
 - class zero characters 166
 - definition 133
- combining grapheme joiner (U+034F) 891
- combining half marks 189, 333
- combining marks *see* combining characters
- comma below 288
- 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 241
- composite characters
 - see* decomposable characters
- Composition Exclusion (normative property) 98
- compression 206
 - see also* UTS #6, A Standard Compression Scheme for Unicode (SCSU)
- conferences 946
- conformance 71–156
 - definitions 85–91
 - examples 68
 - ISO/IEC 10646 implementations 963
 - requirements 77–82
- confusables 243
- conjunct consonants
 - Indic 215, 453
 - Myanmar 647
 - selection of clusters 215
- contextual shaping
 - apostrophe 270
 - Arabic 365
 - not used for Hebrew final forms 360
 - quotation marks 266
 - Syriac 394
- contour tones 323
- control codes 31, 67, 882
 - graphics for 852
 - names 186
 - properties 883
 - semantics 32, 883
 - specified in Unicode 883
- control sequences 882
- conversion of characters 194–195
- convertibility
 - as Unicode design principle 24
- Coptic 305, 308–310
- Coptic Epact numbers 835
- corporate use subarea 902
- corrigenda 74
- CR (U+000D carriage return) 207, 883
- Creative Commons 878
- CRLF (carriage return and line feed) 207
- Croatian 292
 - digraphs 292
- culturally expected sorting 12, 228
- Cuneiform
 - Old Persian 433
 - Sumero-Akkadian 428–431
 - Ugaritic 432
- Cuneiform and Hieroglyphic Area 50
- Cuneiform and Hieroglyphs 427–444
- currency symbols block 817–820
 - currency symbols encoded in other blocks .. 818
 - currency symbols, other 819
 - dollar sign, form and usage 818
 - euro sign 819
 - lari sign 819
 - lira sign, compatibility usage 818
 - lira sign, Turkish 819
 - peso signs, usage 818
 - ruble sign 819
 - rupee signs, Indian, usage 819
 - yen and yuan signs, usage 818
- cursive joining 887–891

- Arabic 373–380
 - control characters for 190, 367–368, 541, 886
 - Mandaic 402
 - Mongolian 540–541
 - N’Ko 775
 - Phags-pa 586
 - Syriac 394–397
 - transparency 890
 - cursive scripts 357
 - Cypriot 340
 - see also* Linear B
 - Cyrillic 311–314
 - Czech 292
- D**
- danda, in Devanagari block 467
 - Danish 291
 - dashes 263
 - Database, Unicode Character
 - see* Unicode Character Database (UCD)
 - dead consonants, Indic 450
 - dead keys 218
 - decomposable characters 62
 - definition 114
 - normalization of 204
 - decomposition 62, 114–116
 - canonical *see* canonical decomposition
 - compatibility *see* compatibility decomposition
 - definition 114
 - in normalization 204
 - mapping, definition 114
 - mappings noted in code charts 926
 - default case
 - algorithms 83, 150–156
 - conversion 152
 - detection 154
 - folding 153
 - default caseless matching 155
 - default grapheme clusters 215
 - see also* UAX #29, Unicode Text Segmentation
 - Default Ignorable Code Point (property) 250
 - default ignorable code points 199, 250
 - default property values
 - definition 95
 - defective combining character sequences 221
 - definition 106
 - dependent vowel signs
 - Indic 449
 - Khmer 658
 - Philippine scripts 686
 - deprecated characters 72, 921
 - alternate format 191, 896–897
 - definition 90
 - Derived Age (property) 200
 - derived properties
 - definition 102
 - DerivedCoreProperties.txt 151, 162, 251
 - DerivedNormalizationProps.txt 240
 - Deseret 793–795
 - design goals of Unicode 4
 - design principles of Unicode 14–24
 - designated code points 30
 - Devanagari 447–472
 - Dhivehi 519
 - diacritics 54, 325
 - alternative glyphs 287, 325
 - Czech 288
 - display in isolation 59, 263, 326
 - double 112, 189, 327
 - German dialectology 331
 - Greek 301–302, 305
 - Latin 287–290
 - Latvian 288
 - mathematical 826
 - on i and j 289
 - rendering 220–225
 - Slovak 288
 - spacing clones of 323, 327
 - symbol 54, 332
 - see also* combining characters
 - dictionary symbols 864
 - digit form names 369
 - digits 203
 - Arabic 830
 - Arabic-Indic 369–370
 - compatibility 830
 - decimal 173
 - glyph variants 832
 - hexadecimal 830
 - Myanmar 830
 - national shapes 897
 - Shan 830
 - superscript and subscript 831
 - Tai Laing 830
 - Tai Tham 830
 - digraphs 292, 295, 297
 - dingbats 866–867
 - directionality 20, 52
 - East Asian scripts 710
 - Middle Eastern scripts 357
 - Mongolian 539
 - musical symbols 801
 - normative property 169
 - Ogham 354
 - Old Italic 344
 - Philippine scripts 687
 - Runic 347

discussion list for Unicode 946
 Dives Akuru 630–631
 Dogra 635–636
 Dogri 468
 Domino Tiles 869
 dotless i 236, 289
 dotted circle
 in code charts 105, 326
 in fallback rendering 220
 to indicate diacritic 54
 to indicate vowel sign placement 55
 double diacritics 112, 189, 327
 Duployan 810–811
 Dutch 291, 292
 dynamic composition
 as Unicode design principle 23
 Dzongkha 526

E

East Asian scripts 709–759
 writing direction 52
 see also CJK ideographs
 Eastern Arabic-Indic digits 369
 EBCDIC
 newline function 208
 editing, text boundaries for 215–216
 efficiency
 as Unicode design principle 15
 Egyptian hieroglyphs 434–440
 format controls 436–440
 Elbasan 351
 ellipsis 272
 Elymaic 423
 e-mail discussion list for Unicode 946
 emoji 861, 862, 946
 animal symbols 865
 charts 946
 cultural symbols 865
 zodiacal symbols 865
 emoji modifiers 865
 emoticons 866
 Enclosed Alphanumerics 876
 enclosing marks 333
 definition 105
 encoded characters 7, 29
 allocation 43–51
 definition 90
 encoding form conversion
 definition 125
 encoding forms 33–38
 ISO/IEC 10646 definitions 958
 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 903
 English 291
 equivalent sequences 204
 as Unicode design principle 23
 case-insensitivity 229, 238
 combining characters in matching 217
 conformance 80
 Hangul syllables 746
 in sorting and searching 228
 language-specific 116
 security implications 243
 see also canonical equivalence
 see also compatibility equivalence
 see also encoding forms, encoding schemes
 errata xxvii, 74, 948
 escape sequences 882
 not used in Unicode 1, 4
 Esperanto 292
 Estonian 292
 Ethiopic 764–767
 Etruscan 343
 European scripts 285–334
 ancient 335–355
 eyelash-RA 459

F

fallback rendering 250
 of nonspacing marks 220
 FAQ (Frequently Asked Questions) 947
 Faroese 291
 Farsi 365, 368
 featural syllabaries 255
 FF (U+000C form feed) 207, 883
 file separator (U+001C) 883
 Finnish 291
 Finno-Ugric Transcription (FUT)
 see Uralic Phonetic Alphabet (UPA)
 fixed-width Unicode encoding form (UTF-32) ... 35,
 122
 flat tables 194
 Flemish 291
 fleurons 868
 fonts
 and Unicode characters 16
 for mathematical alphabets 825–827

style variation for symbols 815
 form feed (U+000C) (FF) 207, 883
 format control characters 30, 67, 261, 881–897
 deprecated 896–897
 prefixed 191, 329
 stateful 894
 fraction characters 843
 fraction slash (U+2044) 271, 839
 French 292
 Frisian 292
 fullwidth forms in East Asian encodings 743
 futhark 346

G

Garshuni 390
 Ge'ez 764
 General Category (normative property) 170
 list of values 170
 general punctuation 259–283
 General Scripts Area 49
 geometrical symbols 857–860
 Georgian 319–320
 German 291
 geta mark (U+3013) 282
 Glagolitic 316
 Glossary 947
 glyph selection tables 194
 glyphs 6, 15
 characters, relationship to 15
 diacritics alternative 287, 325
 Greek alternative 302–304
 Latin alternative 287
 mathematical alternative 845
 missing 250
 representative in code charts 918
 standardized variants 898
 symbols alternative 815
 golden numbers 348
 Gothic 350
 Grantha 626–629
 grapheme base 325
 definition 106
 grapheme clusters 11, 59–60
 see also UAX #29, Unicode Text Segmentation
 default 215
 definition 107
 grapheme extender
 definition 107
 grapheme joiner, combining (U+034F) 891
 graphic characters 30
 Greek 301–306
 acrophonic numerals 203, 306
 alternative glyphs 302–304

ancient musical notation 807–809
 editorial marks 277
 letters as symbols 302–304, 846
 see also Cypriot, Linear B
 Greenlandic 292
 group separator (U+001D) 883
 guillemets 266
 Gujarati 485–486
 Gunjala Gondi 564–565
 Gurmukhi 480–484

H

Hakka 737
 halant 445
 see also virama
 half marks, combining 189, 333
 half-consonants, Indic 454
 halfwidth forms in East Asian encodings 743
 hamza 384–385
 Han ideographs *see* CJK ideographs
 Han unification 720–726
 and language tags 213
 history 967–972
 language usage 717
 source separation rule 715, 721
 source standards 972
 hand symbols 865
 Hangul Area 49
 Hangul syllables 709, 744–747
 and combining marks 112
 as grapheme clusters 60
 canonical decomposition 142
 collation 746
 composition 144
 conjoining jamo 140–149
 equivalent sequences 746
 Hangul Compatibility Jamo 745
 Hangul Jamo 744–747
 Hangul Syllables block 746–747
 Johab set 746
 name generation 145
 normalization 745
 standard 141
 Hangzhou numerals 839
 Hanifi Rohingya 684
 Hanja *see* CJK ideographs
 Hanunóo 686
 Hanzhi *see* CJK ideographs
 harakat 366
 hasant 473
 hash tables 195
 Hatran 426
 Hebrew 359–364

hentaigana 740–742
 hieroglyphs
 Anatolian 443–444
 Egyptian 434–440
 Meroitic 441–442
 high surrogate
 definition 117
 high-surrogate code points 77, 904
 high-surrogate code units 117
 higher-level protocols
 definition 91
 Hindi 447
 Hiragana 739
 horizontal tab (U+0009) 883
 HTML newline function 208
 Hungarian 292
 hyphenation 886
 as a text process 10
 hyphens 263, 886

I

I Ching symbols 871
 IANA charset names 40
 Icelandic 291
 identifiers 227
 see also UAX #31, Unicode Identifier and Pattern
 Syntax
 Ideographic (informative property) 187
 ideographic description sequences 733
 Ideographic Rapporteur Group (IRG) 970
 Ideographic Research Group (IRG) 971
 ideographs *see also* CJK ideographs
 IICore 715, 970
 ill-formed
 definition 120
 Imperial Aramaic 412–413
 implementation guidelines 193–252
 in a Unicode encoding form
 definition 121
 in-band mechanisms 916
 India
 Official scripts 445–515
 Indian rupee signs, usage 819
 Indic scripts 445–515
 principles, in terms of Devanagari 448–458
 relation to ISCII standard 447
 Indic Siyaq 837
 Indonesia and Oceania
 scripts of 685–707
 Indonesian 291
 industry character sets
 covered in Unicode 3
 information separators (U+001C..U+001F) 883

informative properties
 definition 99
 Inscriptional Pahlavi 418
 Inscriptional Parthian 418
 inside-out rule 220
 interchange restrictions 31
 International Phonetic Alphabet (IPA) 254, 294–295
 Spacing Modifier Letters 322
 see also phonetic alphabets
 internationalization 18
 Internationalization & Unicode Conference 946
 Internet protocols
 UTF-8 as preferred encoding 37
 Inuktitut 790
 invisible operators 851
 iota subscript 302
 IPA *see* International Phonetic Alphabet
 IRG (Ideographic Research Group) 971
 Irish 291, 354
 ISCII standard and Unicode 447
 ISO/IEC 10646 949–963
 conformance of Unicode implementations .. 963
 encoding forms 958
 synchrony with Unicode Standard 960
 timeline compared to Unicode versions 952
 Italian 291
 ITC Zapf Dingbats 866
 IUC *see* Internationalization & Unicode Conference

J

jamos *see* Hangul syllables
 Japanese 709
 Japanese era names 877
 Javanese 697–700
 Jawi 386
 jihvamuliya 472, 605
 Johab 746
 joiners 367
 combining grapheme joiner (U+034F) 891
 word joiner (U+2060) 885
 zero width joiner (U+200D) 367–368, 888
 justification 222

K

Kaithi 602–604
 Kana (Hiragana and Katakana) 739–740
 Kanbun 728
 Kangxi radicals 725, 729–730
 Kanji *see* CJK ideographs
 Kannada 504–507
 Kashmiri 469
 Katakana 739–740
 Kawi 691, 693

Kayah Li 676
 KC (normalization form)
 see Normalization Form KC
 KD (normalization form)
 see Normalization Form KD
 keytop labels 852
 Khamti Shan 650
 Kharoshthi 573–574
 Khitan Small Script 760–761
 Khmer 653–664
 characters not recommended 661
 syllable components, order of 662
 Khojki 613–614
 Khudawadi 615–616
 killer 256
 Batak 702
 Brahmi 569
 Meetei Mayek 551
 Myanmar (asat) 648
 see also virama
 Konkani 467
 Korean Hangul *see* Hangul
 Kurdish 386

L

Ladino 359
 language tags 213, 912–916
 and Han unification 213
 use strongly discouraged 912, 915
 Lanna 667
 Lao 643–645
 last-resort glyphs 250
 Latin 287–300
 alternative glyphs 287
 Basic Latin 291
 encoding blocks 44
 IPA Extensions 294–295
 Latin Extended Additional 297–300
 Latin Extended-A 291
 Latin Extended-B 292–294
 Latin Extended-C 297
 Latin Extended-D 298
 Latin Extended-E 299
 Latin Ligatures 297
 Latin-1 Supplement 291
 Phonetic Extensions 296–300
 Latvian 292, 299
 cedilla 288
 layout control characters 67, 885–895
 leading surrogates
 see high-surrogate code units
 legibility criterion for plain text 19
 Lepcha 558–560

letter spacing 886
 letterlike symbols 821–827
 LF (U+000A line feed) 207, 883
 ligatures 887–891
 Arabic 376–377
 combining characters on 58
 control characters for 190
 for nonspacing marks 224
 Latin 297
 selection 216
 Syriac 398
 Limbu 547–550
 line breaking 207–211, 885–887
 control characters 189
 in South Asian scripts 641, 649, 664
 recommendations 209
 see also UAX #14, Unicode Line Breaking Algorithm
 line feed (U+000A) (LF) 207, 883
 line separator (U+2028) (LS) 207, 887
 line tabulation (U+000B) (VT) 883
 Linear A 337
 Linear B 338–339
 see also Cypriot
 linear boundaries 216
 Lisu 752–754
 Lithuanian 292
 little-endian 39
 definition 81
 logical order
 as Unicode design principle 19
 exceptions to 167
 logograph 256
 logosyllabaries 256
 low surrogate
 definition 117
 low-surrogate code points 77, 904
 low-surrogate code units 117
 lowercase 162, 234, 285
 LS (U+2028 line separator) 207, 887
 Lycian 341
 Lydian 341

M

MacOS newline function 208
 Mahajani 611–612
 Mahjong Tiles 868
 mail discussion list for Unicode 946
 Maithili 468
 major version 73
 Makasar 706–707
 Malay 291
 Malay, Patani 642

Malayalam 508–515
 Suriyani 398, 509
 Maltese 292
 Manchu 538
 Mandaic 401–403
 Mandarin 719
 Manden 772
 Manichaean 414–417
 map symbols 864
 mapping tables *see* tables of character data
 Marathi 447, 459, 466
 Marchen 588
 markup languages
 and Unicode conformance 916
 line breaking 207
 Masaram Gondi 562–563
 Mathematical (informative property) 843
 mathematical expression format characters 191
 see also UTR #25, Unicode Support for Mathematics
 mathematical symbols 843–850
 alphabets 823–827
 alphanumeric 822–827
 fonts 825–827
 format characters 851
 fragments for typesetting 853
 invisible operators 851
 operators 844–847
 standardized variants 850
 MathML 847
 matras 166, 449
 Medefaidrin 786
 Meetei Mayek 551–552
 Mende Kikakui 782–783
 Meroitic
 cursive 441–442
 hieroglyphs 441–442
 Miao 755–756
 Middle Eastern scripts 357–520
 ancient 405–426
 Min 719
 Minnan (Hokkien/Fujian, incl. Taiwanese) 737
 minor version 73
 minus sign 846
 commercial (U+2052) 274
 mirrored property
 see Bidi Mirrored (normative property)
 mirroring of paired punctuation 265
 Miscellaneous Symbols 863
 missing glyphs 250
 Modi 621–623
 modifier letters 321–324
 Modifier Letters, Spacing 297
 Mongolian 537–546, 581

 writing direction 539
 moon symbols 863
 Mro 553
 Multani 617
 multibyte encodings
 compared to UTF-8 37
 multistage tables 194
 musical symbols 800–809
 ancient Greek 807–809
 Balinese 695
 Byzantine 806
 directionality 801
 Gregorian 805
 Kievan 805
 Western 800–805
 Myanmar 646–652
 digits 830
 Myanmar Extended-A 650
 Myanmar Extended-B 650

N

N’Ko 772–776
 Nabataean 424
 named character sequences 179
 names, character *see* character names
 namespace 87
 Nandinagari 624–625
 NEL (U+0085 next line) 207, 883
 Nepali 447
 neutral directional characters 169
 New Tai Lue 667–669
 Newa 523–525
 newline function (NLF) 208, 884
 newline guidelines 207–211
 next line (U+0085) (NEL) 207, 883
 NFC (Normalization Form C) 61
 NFD (Normalization Form D) 61
 NFKC (Normalization Form KC) 61
 NFKD (Normalization Form KD) 61
 NLF (newline function) 208, 884
 no-break space (U+00A0) 885
 base for diacritic in isolation 59, 263, 326
 no-break space, narrow (U+202F) 543
 noncharacter code points *see* noncharacters
 noncharacters 31, 905
 conformance 77
 definition 91
 handling 80
 in code charts 921
 interchange restrictions 31
 semantics 32
 U+10FFFF (not a character code) 905
 U+FD00..U+FDEF 31, 905

- U+FFFE (not a character code)66, 906
 - U+FFFF (not a character code)31, 905
 - nondecomposable characters 63
 - non-joiner, zero width (U+200C) 367–368, 889
 - nonlinear boundaries 216
 - non-overlap principle in Unicode encoding forms 33
 - nonspacing marks 325
 - definition 105
 - display in isolation 59, 263, 326
 - positioning 224
 - rendering 220–225
 - see also* combining characters
 - see also* diacritics
 - normalization 61, 204–205
 - and case operations 240
 - canonical ordering algorithm 61, 135, 166
 - conformance 82
 - of private-use characters 902
 - see also* UAX #15, Unicode Normalization Forms
 - stability 132
 - Normalization Form C (NFC) 61
 - Normalization Form D (NFD) 61
 - Normalization Form KC (NFKC) 61
 - Normalization Form KD (NFKD) 61
 - normalization forms 132–139
 - definition 138
 - specification 134
 - normative behaviors
 - definition 85
 - normative properties
 - definition 97
 - list 98
 - may change 97
 - Norwegian 291
 - notational conventions 937–941
 - notational systems 258, 797–813
 - nukta 366, 387, 460
 - null (U+0000)
 - as Unicode string terminator 884
 - number forms
 - CJK ideographs 203
 - numbers
 - Coptic Epact 835
 - handling 203
 - ideographic accounting 174
 - numerals 828–840
 - acrophonic 306
 - Chinese counting rods 841
 - Coptic 310
 - Cuneiform 431
 - Ethiopic 766
 - Greek acrophonic 203
 - Hangzhou 839
 - Meroitic cursive 442
 - old-style 271
 - Roman 203, 843
 - Rumi 836
 - Suzhou-style 839
 - numeric separators 274
 - numeric shape selectors (deprecated) 897
 - Numeric Type (normative property) 173
 - Numeric Value (normative property) 173
 - numero sign (U+2116) 821
 - Nüshu 751
 - Nyiakeng Puachue Hmong 681–682
- ## O
- object replacement character (U+FFFC) 911
 - octet 939
 - Ogham 354
 - Ol Chiki 555–556
 - Old Church Slavonic 311
 - Old Hungarian 349
 - Old Italic 343–345
 - Old North Arabian 407
 - Old Permic 353
 - Old Persian 433
 - Old Sogdian 595
 - Old South Arabian 408–409
 - Old Turkic 594
 - old-style numerals 271
 - Oriya 487–490
 - ornamental dingbats 867
 - Oromo 764
 - Osage 792
 - Osmanya 768
 - Ottoman Siyaq 837
 - out-of-band mechanisms 916
 - overlapping encodings 33
 - overscores 271
- ## P
- Pahawh Hmong 679–680
 - Pahlavi, Inscriptional 418
 - Pahlavi, Psalter 419
 - Palmyrene 425
 - Panjabi 480
 - paragraph or section marks 274
 - paragraph separator (U+2029) (PS) 207, 887
 - Parthian, Inscriptional 418
 - Pashto 365
 - Patani Malay 642
 - Pau Cin Hau 683
 - Persian 365, 368
 - Phags-pa 581–587
 - Phaistos Disc symbols 872
 - Phake 652

- Philippine scripts 686–688
- Phoenician 410
- phonemes 257
- phonetic alphabets 254
- IPA Extensions 294–295
 - Phonetic Extensions 296–300
 - Spacing Modifier Letters 322–324
 - Uralic Phonetic Alphabet (UPA) 274, 296
 - see also* International Phonetic Alphabet (IPA)
- Pinyin 291
- pipeline table
- proposed new characters 947
- pivot code, Unicode as 194
- 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
 - Planes 15–16 (Private Use) 51, 903
- Playing Cards 869
- points, Hebrew pronunciation marks 359
- policies of the Unicode Consortium 947
- Polish 292
- Portuguese 291
- precomposed characters
- see* decomposable characters
 - compatibility *see* compatibility decomposable characters
- prefixed format control characters 191
- prepended concatenation marks 251, 329
- Private Use Area (PUA) 49, 902
- Private Use planes 44, 51, 903
- private-use characters
- properties 901
 - semantics 32
- private-use code points 31, 199
- conformance 78
- definition 103
- high surrogates 904
- properties 18, 93–103, 157–192
- aliases 160
 - aliases (definition) 102
 - and Unicode algorithms 98
 - data tables 194
 - 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 883
 - provisional *see* provisional properties
 - simple *see* simple properties
 - see also* individual properties, e.g. combining classes
- property values
- aliases 160
 - aliases (definition) 103
 - default 95
 - default (definition) 95
 - normative references to 82
- PropertyAliases.txt 103, 940
- PropertyValueAliases.txt 103, 940
- PropList.txt 164
- Provençal 292
- provisional properties
- definition 100
- PS (U+2029 paragraph separator) 207, 887
- Psalter Pahlavi 419
- PUA (Private Use Area) 49, 902
- punctuation 259–283
- blocks containing 253
 - CJK 280
 - doubled 271
 - ideographic 731
 - in bidirectional text 259
 - paired 265
 - small form variants 283
 - typographic forms 259
 - vertical forms 282
- Punctuation and Symbols Area 49
- Punjabi 480
- ## Q
- quotation marks 266–269
- East Asian 268
 - European 266
- ## R
- radicals, Kangxi and other CJK 729–730
- radical-stroke index 725
- record separator (U+001E) 883
- recycling symbols 864
- references 947
- referencing 82
- properties 75
 - Unicode algorithms 76
 - Unicode Standard 74
- regional indicator symbols 877
- regular expressions 212
- and line breaking 207
 - see also* UTS #18, Unicode Regular Expressions
- Rejang 701
- rendering of text 6, 10, 17
- fallback 250

- unsupported characters 199
 - repertoire of abstract characters 29
 - reph 458, 462, 502
 - replacement character (U+FFFD) 42, 67, 81, 911
 - reserved code points 30, 199
 - definition 91
 - in code charts 921
 - preservation in interchange 31
 - see also* unassigned code points
 - Rhaeto-Romanic 292
 - rich text 18
 - right single quotation mark (U+2019)
 - preferred for apostrophe 270
 - right-to-left text 52
 - East Asian scripts 710
 - Middle Eastern scripts 357
 - roadmap for script additions 45, 947
 - Roman numerals 203, 843
 - Romanian 292
 - comma below 289
 - Romany 292
 - Rong 558–560
 - Rumi numeral symbols 836
 - Runic 346–348
 - Russian 311
- S**
- Samaritan 399–400
- Sami 292
- Sanskrit 447
- Saurashtra 561
- scalar values, Unicode
 - see* Unicode scalar values
- scripts
 - in Unicode Standard 3
 - roadmap for future additions 45, 947
 - types of 258
 - see also* UAX #24, Unicode Script Property
- SCSU
 - see* UTS #6, A Standard Compression Scheme for Unicode
- searching 228–230
 - as a text process 10
 - case-insensitive 229, 238
- section or paragraph marks 274
- security issues 243
- self-synchronization of encoding forms 34
- semantics
 - see* character semantics
- sequences
 - notation 938
- Serbian
 - corresponding digraphs in Croatian 292
- Shan 665
 - digits 830
- Sharada 605–606
- Shavian 355, 752
- Show Hidden 79, 220, 250, 899
- SHY (U+00AD soft hyphen) 886
- Sibe 539
- Siddham 609–610
- signature for Unicode data 66, 907–909
- simple properties
 - definition 102
 - simplified Chinese 718
- Sindhi 365, 467
- Sinhala 521–522
- Sinological dot 299
- SIP (Supplementary Ideographic Plane) 43, 51
- Siyaq Numbers 836
 - Indic 836
- slash, fraction (U+2044) 271
- Slovak 292
- Slovenian 292
- small letters 162, 234, 285
- SMP (Supplementary Multilingual Plane) 43, 50
- soft hyphen (U+00AD) (SHY) 886
- Sogdian 596
- Somali 768
- Sora Sompeng 634
- Sorbian 292
- sorting 12, 228
 - and combining grapheme joiner 892
 - as a text process 10
 - case-insensitive 229
 - culturally expected 12, 228
 - language-insensitive 228
 - see also* Unicode Collation Algorithm (UCA)
- source separation rule 715, 721
- South and Central Asian scripts
 - Ancient 567–596
 - Other historic 597–636
 - Other modern 517–565
- South Asian scripts 445–550
- Southeast Asian scripts 637–684
- Soyombo 592–593
- space (U+0020)
 - base for diacritic in isolation 59, 263, 326
- space characters 262, 885–887
 - graphics for 852
 - space, zero width (U+200B) 262
- spacing clones of diacritics 323, 327
- spacing marks 325
 - definition 106
- Spacing Modifier Letters 322–324
- Spanish 291
- special characters 66, 881–916

- SpecialCasing.txt 150, 164
- Specials 907–911
- spell-checking
 - as a text process 11
- spellings, alternative
 - see* equivalent sequences
- spoofing 243
- SSP (Supplementary Special-purpose Plane) 44
- stability 100, 159
 - as Unicode design principle 23
- stacked boundaries 215
- stacking sequences 56
 - nondefault 57
- standardized variants 542, 898
 - in the code charts 928
 - mathematical symbols 850
- StandardizedVariants.txt 542, 850
- standards coverage 3
- starters 134
- stateful encoding
 - not used in Unicode 4
 - paired format controls 894
- string comparison 12
- string literals, Unicode
 - code point notation `\u1234` 940
- strings, Unicode 42, 119
 - null termination 884
- strong directional characters 169
- styled text 18
- sublinear searching 230
- subsets, supported 70
 - conformance 78
 - ISO/IEC 10646 specification for 961
- substitution character
 - see* replacement character
- Sumero-Akkadian 428–431
- Sundanese 704–705
- superscripts 323
 - and subscripts 841
- supplementary characters
 - in UTF-16 strings 42
 - tables for 195
- 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, 903
- Supplementary Special-purpose Plane (SSP) 44
- supported subsets 70
 - conformance 78
- supralineation 309
- surrogate code points
 - see* surrogates
- surrogate pairs 36, 123
 - definition 117
 - processing 201–202
- surrogates 31, 117, 904
 - interchange restrictions 31
 - isolated surrogates, handling 42
 - isolated surrogates, ill-formed 123
 - isolated surrogates, uninterpreted 117
 - support levels 201
- Surrogates Area 49, 904
- Sutton SignWriting 812–813
- Suzhou-style numerals 839
- svasti signs 533
- Swahili 291
- Swedish 291
- syllabaries 255
 - alphabetic property 187
 - featural 255
- Syloti Nagri 600–601
- symbols 815–879
 - animal 865
 - appearance variation 815
 - arrows 849–850
 - box drawing 857
 - cultural 865
 - currency symbols block 817–820
 - dictionary 864
 - dingbats 866–867
 - emoji 861, 862, 877
 - Enclosed Alphanumerics 876
 - fragments for mathematical typesetting 853
 - game 865
 - gender 864
 - genealogical 865
 - geometrical 857–860
 - hand 865
 - Khmer lunar calendar 664
 - letterlike 821–827
 - map 864
 - mathematical 843–850
 - mathematical alphanumeric 822–827
 - miscellaneous 863
 - musical 800–809
 - numerals 828–840
 - recycling 864
 - regional indicator 877
 - technical 852–856
 - weather 863
 - zodiacal 865
- symmetric swapping format characters 896
- Syriac 390–398

T

tab (U+0009 character tabulation) 883
 tab, vertical (U+000B) 207, 883
 tables of character data 194–195
 optimization 195
 supplementary characters 195
 tag characters 912–916
 Tagalog 686
 Tagbanwa 686
 tags, language 213, 912–916
 use strongly discouraged 915
 Tai Laing
 digits 830
 Tai Le 665–666
 Tai Tham 670–672
 digits 830
 Tai Viet 673–675
 Tai Xuan Jing symbols 871
 Takri 607–608
 Tamil 491–500
 Tangut 757–759
 components 758–759
 radicals 758
 tashkil 366
 tashkil, harakat, points 368
 TCHAR in Win32 API 198
 Technical Reports (UTR) 945
 Technical Standards (UTS) xxvi, 945
 abstracts 946
 technical symbols 852–856
 Telugu 501–503
 terminal emulation 816
 text boundaries 60, 188, 215–216, 226
 see also UAX #14, Unicode Line Breaking Algorithm
 see also UAX #29, Unicode Text Boundaries
 text elements 6, 10, 215
 boundaries 226
 for sorting 228
 text processes 6, 10–13
 text rendering 6, 10, 17
 text selection, boundaries for 215–216
 Thaana 519–520
 Thai 639–642
 Tibetan 526–536
 Tifinagh 769
 Tigre 764
 tilde (U+007E) 274
 Tirhuta 618–620
 titlecase 162, 234
 Todo 538
 tone letters 323–324

tone marks
 Bopomofo spacing 736, 737
 Chinantec 324
 Chinese 324
 Tai Le 665
 Thai 639
 Vietnamese 290
 traditional Chinese 718
 traffic signs 864
 trailing surrogates
 see low-surrogate code units
 transcoding 194–195
 tables 194
 Transport and Map Symbols 866
 triangulation in transcoding 194
 tries 194
 truncation
 combining character sequences 218–219
 surrogates and 202
 Turkish 292
 case mapping of I 236, 289
 cedilla 289
 lira sign 819
 two-stage tables 195

U

U+ notation 940
 U+10FFFF (not a character code) 905
 U+FEFF (BOM) 907–909
 U+FFFE (not a character code) 906
 U+FFFF (not a character code) 905
 UAX (Unicode Standard Annex) xxiv, 945
 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-2 958
 UCS-4 958
 Ugaritic 432
 Ukrainian 311
 unassigned code points 30, 77, 199
 defined as reserved code points 91
 handling 72
 properties of 95
 semantics 77
 see also reserved code points
 underscores 271
 undesignated code points 30
 Unicode 1.0 Name (informative property) 186

- 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) ..xxiv, 159, 947
 - 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+ 940
- Unicode codespace
 - definition 88
 - planes 43
 - size 1, 29
- Unicode Collation Algorithm (UCA) 12
- Unicode conferences 946
- Unicode Consortium 944
 - addresses 948
 - Consortium membership in standards bodies 944
 - e-mail discussion list 946
 - membership 944
 - policies 947
 - website 946
- Unicode data signature 66, 907–909
- Unicode data types 197–198
 - for C 197–198
- Unicode encoding forms 118–125
 - conformance 34, 80
 - definition 119
 - fixed-width (UTF-32) 35, 122
 - signatures 908, 909
 - variable-width 36, 37, 123
 - see also* encoding forms
- Unicode encoding schemes
 - conformance 128–131
 - definition 128
 - endian ordering 39
 - see also* encoding schemes
- Unicode escape sequence notation \u1234 940
- Unicode scalar values
 - definition 118
- Unicode security 243
see also UTS #39, Unicode Security Mechanisms
- Unicode Standard
 - allocation of encoded characters 43–51
 - architecture 10–13
 - areas 44
 - benefits 1
 - blocks 253
 - code charts 917–935
 - components 77
 - conformance 71–156
 - conformance of ISO/IEC 10646 implementations 963
 - corrections 74
 - definitions for conformance 85–91
 - design goals 4
 - design principles 14–24
 - errata 74, 948
 - normative references to 74, 82
 - number of characters 3
 - number of code points 1, 29
 - script coverage 3
 - security issues 243
 - synchrony with ISO/IEC 10646 960
 - updates 948
 - versions *see* versions of the Unicode Standard
see also Version 13.0
- Unicode Standard Annexes (UAX)xxiv, 945
 - as components of Unicode Standard 77
 - conformance 83
 - list of 83
- Unicode string literals
 - code point notation \u1234 940
- Unicode strings 42
 - definition 119
- Unicode Technical Committee (UTC) 944
- Unicode Technical Reports (UTR) 945
- Unicode Technical Standards (UTS)xxvi, 945
 - abstracts 946
- UnicodeData.txt 150, 164
- unification
 - as Unicode design principle 21
 - see also* Han unification
- Unified Repertoire and Ordering (URO) ... 721, 969
see also Han unification
- Unihan Database 159, 725, 726, 947, 970
- Unihan.zip 100, 159
- UnihanCore2020 715, 970
- unit separator (U+001F) 883
- Universal Character Set (UCS)
see ISO/IEC 10646
- universality
 - as Unicode design principle 14
- Unix
 - newline function 208
 - UTF-8 in 18
- unsupported characters 199
- upadhmaniya 472, 605
- update version 73
- uppercase 162, 234, 285
- Uralic Phonetic Alphabet (UPA) 274, 296

- Urdu 365
 - URO (Unified Repertoire and Ordering) .. 721, 969
 see also Han unification
 - UTF, Unicode Transformation Formats 33, 119
 - as encoding form or scheme 131
 - binary comparison and sort order differences ...
 229, 231
 - in APIs 198
 - UTF-16 36, 123, 959
 - binary comparison and sort order caution ... 36
 - bit distribution (table) 123
 - BOM in 129, 907
 - encoding form (definition) 123
 - encoding scheme (definition) 129
 - encoding schemes 39
 - in ISO/IEC 10646 959
 - in UTF-8 order 232
 - surrogates and string handling 42, 201
 - UTF-16BE (Big-endian) 908
 - encoding scheme 40
 - encoding scheme (definition) 128
 - UTF-16LE (Little-endian) 908
 - encoding scheme 40
 - encoding scheme (definition) 128
 - UTF-32 35, 122
 - BOM in 130
 - encoding form (definition) 122
 - encoding scheme (definition) 130
 - encoding schemes 39
 - UTF-32BE (Big-endian)
 - encoding scheme 40
 - encoding scheme (definition) 130
 - UTF-32LE (Little-endian)
 - encoding scheme 40
 - encoding scheme (definition) 130
 - UTF-8 37, 123, 959
 - ASCII transparency 37
 - binary comparison and sort order 38
 - bit distribution (table) 124
 - BOM in 128, 131, 908
 - byte ranges 124
 - compared to multibyte encodings 37
 - encoding form (definition) 123
 - encoding scheme 39
 - encoding scheme (definition) 128
 - in Unix 18
 - in UTF-16 order 231
 - non-shortest form is invalid 123, 243
 - preferred encoding for Internet protocols 37
 - security and 243
 - signature 128, 131, 908
 - UTR (Unicode Technical Report) 945
 - UTS (Unicode Technical Standard) xxvi, 945
 - abstracts 946
 - Uyghur 365, 581
- ## V
- Vai 777–778
 - valid (synonym for well-formed) 121
 - variable-width Unicode encoding form ... 36, 37, 123
 - variants
 - compatibility 26
 - fullwidth and halfwidth 283
 - mathematical symbols 850
 - small form 283
 - standardized 898
 - variation selectors 192, 898
 - ideographic variation mark (U+303E) 734
 - Mongolian free variation selectors 541
 - variation sequences 898
 - for Phags-pa 585–587
 - Version 13.0 77
 - number of characters 3
 - versions of the Unicode Standard .xxiv, 72, 948, 965–966
 - backward compatibility 72
 - compared to ISO/IEC 10646 editions 965
 - content 73
 - interaction in implementations 200
 - numbering 73
 - property changes 72
 - stability 72
 - updates 948
 - vertical tab (U+000B) 207, 883
 - vertical text 52, 260, 282
 - East Asian scripts 710
 - Mongolian 539
 - Vietnamese 290, 297
 - ideographs 710
 - virama 256, 445
 - definition 450
 - Kharoshthi 577
 - Khmer 656
 - Myanmar 647
 - Philippine scripts 686
 - virama-like characters 190
 - visual order used for Thai and Lao 21
 - vowel harmony
 - Mongolian 543
 - vowel marks, Middle Eastern scripts 358
 - vowel separator
 - Mongolian 544
 - vowel signs
 - Indic 55, 449
 - Khmer 658
 - Philippine scripts 686

W

- Wancho 566
- Warang Citi 554
- wchar_t
 - in C language 198
- weak directional characters 169
- weather symbols 863
- website, Unicode Consortium 946
- Weierstrass elliptic function symbol 822
- well-formed
 - definition 120
- Welsh 292
- Where Is My Character? 948
- wide characters
 - data type in C 198
- wiggly fence (U+29DB) 848
- Windows newline function 208
- word breaks 217, 885–887
 - in South Asian scripts 641, 649, 664
- word joiner (U+2060) 885
- writing direction *see* directionality
- writing systems 254–258
- Wu (Shanghainese) 719

X

- Xibe 539
- Xishuangbanna Dai 667

Y

- Yezidi 404
- Yi 748–750
- Yiddish 359
- Yijing Hexagram Symbols 871
- ypogegrammeni 302

Z

- Zanabazar Square 589–591
- Zapf Dingbats 866
- zero extension relation among encodings 958
- zero width joiner (U+200D) 367–368, 888
- zero width no-break space (U+FEFF) ... 66, 82, 885
 - initial 131, 908
- zero width non-joiner (U+200C) 367–368, 889
- zero width space (U+200B) 886
 - for word breaks in South Asian scripts . 641, 649, 664
- zero-width space characters 886
- ZWJ *see* zero width joiner (U+200D)
- ZWNBS *see* zero width no-break space (U+FEFF)
- ZWNJ *see* zero width non-joiner (U+200C)
- ZWSP *see* zero width space (U+200B)

