

The Unicode® Standard

Version 9.0 – 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.

© 2016 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 9.0.

Includes bibliographical references and index.

ISBN 978-1-936213-13-9 (<http://www.unicode.org/versions/Unicode9.0.0/>)

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

QA268.U545 2016

ISBN 978-1-936213-13-9

Published in Mountain View, CA

July 2016

I Index

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

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

A

- abbreviation, Coptic 314
- abjads 260, 361
- abstract character sequences
 - definition 90
- abstract characters 29
 - definition 90
- abugidas 261, 262, 443, 605
- accent marks *see* diacritics
- accented characters
 - encoding 12
 - Latin 293
 - normalization 208
- accounting numbers, ideographic 178
- acrophonic numerals 207, 311
- Adlam 734–735
 - reference materials 959
- Aegean numbers 342
- Africa
 - scripts of 713–735
- Afrikaans 298
- Ahom 602–603
 - reference materials 942
- Ainu 697
- Aiton 620
- Alchemical Symbols 811
 - reference materials 942
- Algonquian 740
- Ali Gali 527
- aliases
 - character name 88, 183
 - informative 862
 - normative 863
 - property 162
 - property value 162
- allocation areas 45
- allocation of encoded characters 44–52
- Alphabetic (informative property) 190
- alphabets 260
 - European 291–338
 - mathematical 770–774
- alternate format characters (deprecated) . . 194, 836–837
- Americas
 - scripts of 737–745
- Amharic 714
- Anatolian hieroglyphs 441–442
 - reference materials 942
- Ancient Symbols 814
- angle brackets (U+2329 and U+232A)
 - deprecated for technical publication 798
- Annexes, Unicode Standard (UAX) xxxiii, 883
 - as components of Unicode Standard 79
 - conformance 85
 - list of 85
- annotation characters 849–851
 - use in plain text discouraged 850
- ANSI/ISO C
 - wchar_t and Unicode 202
- apostrophe (U+0027) 276
- Arabic 369–391
 - digits 777
- Arabic-Indic digits 373–374
 - signs used with 375
- ArabicShaping.txt 377, 382, 397
- Aramaic 414, 443, 527, 551, 556
- areas of the Unicode Standard 45
- ARIB 807
- Armenian 321–322
- arrows 794–795
- ASCII
 - characters with multiple semantics 266
 - transparency of UTF-8 36
 - Unicode modeled on 1
 - zero extension 202, 900
- Assamese 467
- assigned code points 11, 30
- Athapascan 740

- atomic character boundaries 220
- Avestan 422
 - reference materials 943
- B**
- Balinese 655–660
 - reference materials 943
- Bamum 729–730
 - reference materials 943
- Bangla 467–472
- base characters 329
 - definition 105
 - multiple 59
 - ordered before combining marks 222, 329
- Basic Multilingual Plane (BMP) 1, 44
 - allocation areas 49
 - representation in UTF-16 36
- Basque 298
- Bassa Vah 731
 - reference materials 943
- Batak 666
 - reference materials 944
- benefits of Unicode 1
- Bengali 467–472
- Bhaiksuki 562–563
 - reference materials 944
- Bidi Class (normative property) 173
- Bidi Mirrored (normative property) 180
- Bidi Mirroring Glyph (informative property) 181
- BidiMirroring.txt 181
- Bidirectional Algorithm, Unicode 53, 84
- bidirectional ordering 20
 - controls 833
- bidirectional text 53, 84
 - Middle Eastern scripts 361
 - nonspacing marks in 225
 - punctuation in 265
- big-endian 40
 - definition 83
- Bihari 463
- binary comparison and sort order
 - caution for UTF-16 36
 - UTF differences 233, 235
 - UTF-8 39
- blocks of the Unicode Standard 45, 259
- Blocks.txt 45
- BMP *see* Basic Multilingual Plane
- BNF (Backus-Naur Form) 877
- BOCU-1 *see* ITN #6, BOCU-1
 - MIME-Compatible Unicode Compression
- Bodhi 515
- Bodo 462
- BOM (U+FEFF) 40, 67, 130–133, 847–849
- Bopomofo 693–695
- boundaries, text 61, 191, 219–220, 230
 - see also* UAX #14, Unicode Line Breaking Algorithm
 - see also* UAX #29, Unicode Text Segmentation
- boustrophedon 53, 351
- box drawing symbols 802
- Brahmi 443, 551, 552–555, 556, 607
 - reference materials 944
- Braille 748–749
- Breton 298
- Buginese 653–654
- Buhid 650
- Bulgarian 316
- bullets 279
 - numeric 778
- Burmese *see* Myanmar
- Byelorussian 316
- byte order mark (BOM) (U+FEFF) . 40, 67, 130–133, 847–849
- byte ordering
 - changing 81
 - conformance 83
- byte serialization 40, 67
- Byzantine Musical Symbols 755
- C**
- C language
 - wchar_t and Unicode 202
- C0 and C1 control codes 31, 189, 822
- Cambodian *see* Khmer
- Canadian Aboriginal Syllabics 740–741
 - reference materials 945
- candrabindu 465, 580
- canonical composite characters
 - see* canonical decomposable characters
- canonical composition algorithm 138
- canonical decomposable characters
 - definition 117
- canonical decomposition 63
 - definition 116
 - mappings 115
- canonical equivalence
 - definition 117
 - nonspacing marks 227
- canonical equivalent character sequences
 - conformance 81
- canonical mappings
 - see* canonical decomposition mappings
- canonical ordering algorithm 137
- canonical precomposed characters
 - see* canonical decomposable characters
- Cantonese 677

- capital letters 164, 238, 291
- Carian 345
 - reference materials 945
- carriage return (U+000D) (CR) 211, 823
- carriage return and line feed (CRLF) 211
- case 299
 - and text processes 12
 - beyond ASCII 239
 - camelcase 241
 - case folding 242
 - case operations (conformance) 85, 152–158
 - case operations and normalization 244
 - case operations, reversibility 241
 - cased (definition) 153
 - case-insensitive comparison 157, 233, 242
 - casing context (definition) 153
 - conversion 154
 - detection 156
 - European alphabets 291
 - exceptional Latin pairs 295, 299
 - Georgian 323
 - lowercase 164, 238, 291
 - mapping tables 198
 - mappings 152, 166, 238–240
 - mappings noted in code charts 866
 - titlecase 164, 238
 - Turkish I 240, 295
 - uppercase 164, 238, 291
 - see also* default case
- Case (normative property) 164, 238
- CaseFolding.txt 166, 242
- caseless letters 299
- Catalan 297
- Caucasian Albanian 356
 - reference materials 945
- cedilla 294
- CEF *see* character encoding forms
- CES *see* character encoding schemes
- CESU-8
 - see* UTR #26, Compatibility Encoding Scheme for UTF-16: 8-Bit (CESU-8)
- Chakma 546
 - reference materials 945
- Cham 644–645
 - reference materials 945
- character encoding forms (CEF) 33–39, 900
 - see also* Unicode encoding forms
- character encoding model 33, 42
 - see also* UTR #17, Unicode Character Encoding Model
- character encoding schemes (CES) 40–43
 - see also* Unicode encoding schemes
- character encoding standards
 - coverage by Unicode 3
- Character Index 889
- character literals, Unicode
 - code point notation U+ 878
- character mapping
 - interchange format *see* UTS #22, Character Mapping Markup Language (CharMapML)
- character names 88, 182–188, 904
 - aliases 88, 183
 - conventions 875
 - for CJK ideographs 870
 - for control codes 187, 189
 - in code charts 862
 - matching 182
- character properties
 - see* properties
 - see also individual properties, e.g.* Combining Class
- character semantics 1, 80, 87–88, 905
 - as Unicode design principle 18
 - ASCII 266
 - definition 87
- character sequences
 - abstract *see* abstract character sequences
 - canonical equivalent *see* canonical equivalent character sequences
 - compatibility equivalent *see* compatibility equivalent character sequences
 - conformance 81
 - named 183
- character sequences, combining 105
- character shaping selectors (deprecated) 836
- character tabulation (U+0009) 823
- characters
 - abstract *see* abstract characters
 - arrangement in Unicode 46
 - assigned 11, 30
 - blocks 45, 259
 - boundaries 219
 - canonical decomposable *see* canonical decomposable characters
 - classes 878
 - code charts 857–873, 889
 - coded *see* encoded characters
 - combining *see* combining characters
 - compatibility decomposable *see* compatibility decomposable characters
 - composite *see* decomposable characters
 - concept of 15, 60
 - conformance definitions 90–92
 - confusable 248
 - conversion 198–199
 - decomposable *see* decomposable characters
 - deprecated *see* deprecated characters
 - encoded *see* encoded characters
 - encoding forms *see* encoding forms

- encoding schemes *see* encoding schemes
- end-user perceived 60
- format control 30, 68, 267, 821–837
- glyphs, relationship to 15
- graphic 30
- identity (definition) 87
- ignored in processing 251–256
- interpretation 80
- layout control 68, 825–835
- modification 81
- names list 858–869
- names *see* character names
- not encoded in Unicode 3
- number encoded in Version 9.0 3
- precomposed *see* decomposable characters
- properties *see* properties
- semantics *see* character semantics
- special 67, 821–856
- supplementary *see* supplementary characters
- transcoding 198–199
- unsupported 203
- characters, not glyphs
 - in spoofing 249
 - Unicode principle 15
- CharMapML
 - see* UTS #22, Character Mapping Markup Language (CharMapML)
- charsets
 - IANA registered names 41
- charts, character code *see* code charts
- Cherokee 738
 - reference materials 945
- Chinese 676–677
 - Cantonese 677
 - Hakka 694
 - Mandarin 677
 - Minnan (Hokkien/Fujian, incl. Taiwanese) 694
 - simplified and traditional 676
- Chu hán 675
- Chu Nôm 916
- citations for
 - properties 77
 - Unicode algorithms 78
 - Unicode Standard 76
- CJK ideographs 262, 671–686
 - accounting numbers 178
 - CJK Compatibility Ideographs 685–686
 - CJK Compatibility Supplement 686
 - CJK Strokes 688, 923
 - CJK Unified Ideographs 671–685
 - CJK Unified Ideographs Extension A 673
 - CJK Unified Ideographs Extension B 685
 - CJK Unified Ideographs Extension C 685
 - CJK Unified Ideographs Extension D 685
 - CJK Unified Ideographs Extension E 685
 - code charts 870
 - compatibility ideographs in Plane 2 52
 - component structure 680
 - encoding blocks 672
 - ideographic description sequences 689–692
 - ideographic variation mark (U+303E) 691
 - KangXi radicals 684, 687–688
 - names 870
 - numbers 777
 - numeric values 178, 207
 - order of encoding 682
 - radicals 687–688
 - source standards 918–922
 - unknown or unavailable 287
 - Vietnamese 670
- CJK Miscellaneous Area 50
- CJK punctuation and symbols 286
 - compatibility forms 288
 - overscores and underscores 288
 - quotation marks 274
 - sesame dots 287
 - vertical forms 288
- CJK-JRG (Chinese/Japanese/Korean Joint Research Group) 914
- CJKV Ideographs Area 50
- CLDR (Unicode Common Locale Data Repository) 890
 - cluster boundaries 219
- code charts 857–873, 889
 - representative glyphs 858
- code point sequences
 - notation 876
- code points 7, 29
 - assigned 11, 30
 - assignment 46
 - categories 30
 - default ignorable 203, 255
 - definition 90
 - designated 30
 - notation 875
 - number in Unicode Standard 1
 - private-use *see* private-use code points
 - reserved *see* reserved code points
 - semantics 32
 - surrogate *see* surrogates
 - unassigned *see* unassigned code points
 - undesigned 30
- code positions *see* code points
- code set independence 18
- code unit sequences
 - definition 119
 - ill-formed (definition) 121
 - notation 876

- well-formed (definition) 121
- code units
 - definition 119
 - isolated 118
- code values *see* code units
- coded character representations
 - see* coded character sequences
- coded character sequences
 - definition 91
- coded characters *see* encoded characters
- codespace *see* Unicode codespace
- coeng 621, 624
- Collation Algorithm, Unicode (UCA) 12
- collation *see* sorting
- collation tables 198
- combining character sequences 56, 105
 - defective 225
 - definition 107
 - Latin 293
 - line breaking 221
 - matching 221
 - order of base character and marks 222, 329
 - rendering 221
 - selection 219
 - truncation 222–223
- combining characters 55–60, 109–114, 221–229
 - blocking reordering 832
 - canonical ordering 62, 137, 168
 - class zero 169
 - combining marks 329–330
 - definition 105
 - dependence 329
 - display order 58
 - keyboard input 222
 - ligatures 59
 - multiple 57
 - multiple base characters 59
 - normalization of 208
 - ordering conventions 56
 - rendering of marks 224–229
 - reordrant 169
 - script-specific 56
 - split 170
 - strikethrough 172
 - subjoined 172
 - typographical interaction 58, 168
 - vertical stacking 58
 - see also* diacritics
- Combining Class (normative property) 168
- combining classes 135, 168, 227–228
 - class zero characters 168
 - definition 135
- combining grapheme joiner (U+034F) 831
- combining half marks 192, 337
- combining marks *see* combining characters
- comma below 294
- Compatibility and Specials Area 26, 50
- compatibility characters 22
- compatibility composite characters 27
 - see* compatibility decomposable characters
- compatibility decomposable characters 26
 - definition 115
- compatibility decomposition 63
 - definition 115
- compatibility decomposition mappings 115
- Compatibility Encoding Scheme for UTF-16
 - see* UTR #26, Compatibility Encoding Scheme for UTF-16: 8-Bit (CESU-8)
- compatibility equivalence
 - definition 116
- compatibility equivalent character sequences
 - conformance 81
- compatibility mappings
 - see* compatibility decomposition mappings
- compatibility precomposed characters
 - see* compatibility decomposable characters
- compatibility variants 26
 - mapping 246
- composite characters
 - see* decomposable characters
- Composition Exclusion (normative property) 99
- compression 210
 - see also* UTS #6, A Standard Compression Scheme for Unicode (SCSU)
- conferences 889
- conformance 73–158
 - definitions 87–92
 - examples 69
 - ISO/IEC 10646 implementations 905
 - requirements 79–84
- confusables 248
- conjunct consonants
 - Indic 219, 449
 - Myanmar 615
 - selection of clusters 219
- contextual shaping
 - apostrophe 276
 - Arabic 369
 - not used for Hebrew final forms 364
 - quotation marks 272
 - Syriac 396
- contour tones 327
- control codes 31, 68, 822
 - graphics for 797
 - names 189
 - properties 823
 - semantics 32, 823
 - specified in Unicode 823

- control sequences 822
 - conversion of characters 127, 198–199, 257
 - convertibility
 - as Unicode design principle 23
 - Coptic 309, 313–315
 - reference materials 946
 - Coptic Epact numbers 781
 - corporate use subarea 842
 - corrigenda 76
 - CR (U+000D carriage return) 211, 823
 - CRLF (carriage return and line feed) 211
 - Croatian 298
 - digraphs 298
 - culturally expected sorting 12, 232
 - Cuneiform
 - Old Persian 433
 - Sumero-Akkadian 428–431
 - Ugaritic 432
 - Cuneiform and Hieroglyphic Area 51
 - Cuneiform and Hieroglyphs 427–442
 - currency symbols block 765–767
 - currency symbols encoded in other blocks .. 766
 - currency symbols, other 767
 - dollar sign, form and usage 766
 - euro sign 767
 - lari sign 767
 - lira sign, compatibility usage 766
 - lira sign, Turkish 767
 - peso signs, usage 766
 - ruble sign 767
 - rupee signs, Indian, usage 767
 - yen and yuan signs, usage 766
 - cursive joining 827–831
 - Arabic 377–384
 - control characters for 193, 371–372, 530, 826
 - Mandaic 404
 - Mongolian 529–531
 - N’Ko 725
 - Phags-pa 569
 - Syriac 396–399
 - transparency 830
 - cursive scripts 361
 - Cypriot 344
 - reference materials 953
 - see also* Linear B
 - Cyrillic 316–319
 - Czech 298
- D**
- danda, in Devanagari block 461
 - Danish 297
 - dashes 269
 - Database, Unicode Character
 - see* Unicode Character Database (UCD)
 - dead consonants, Indic 448
 - dead keys 222
 - decomposable characters 63
 - definition 115
 - normalization of 208
 - decomposition 63, 115–117
 - canonical *see* canonical decomposition
 - compatibility *see* compatibility decomposition
 - definition 115
 - in normalization 208
 - mapping, definition 115
 - mappings noted in code charts 866
 - default case
 - algorithms 85, 152–158
 - conversion 154
 - detection 156
 - folding 155
 - default caseless matching 157
 - default grapheme clusters 219
 - see also* UAX #29, Unicode Text Segmentation
 - Default Ignorable Code Point (property) 255
 - default ignorable code points 203, 255
 - default property values
 - definition 96
 - defective combining character sequences 225
 - definition 107
 - dependent vowel signs
 - Indic 447
 - Khmer 626
 - Philippine scripts 650
 - deprecated characters 74, 861
 - alternate format 194, 836–837
 - definition 91
 - Derived Age (property) 203
 - derived properties
 - definition 103
 - DerivedCoreProperties.txt 153, 164, 255
 - DerivedNormalizationProps.txt 245
 - Deseret 743–745
 - reference materials 946
 - design goals of Unicode 4
 - design principles of Unicode 14–24
 - designated code points 30
 - Devanagari 445–466
 - Dhivehi 509
 - diacritics 55, 329
 - alternative glyphs 293, 329
 - Czech 294
 - display in isolation 60, 269, 330
 - double 113, 192, 331
 - German dialectology 335
 - Greek 306–307, 310
 - Latin 293–296

Latvian 294
 mathematical 773
 on i and j 295
 rendering 224–229
 Slovak 294
 spacing clones of 327, 331
 symbol 55, 336
see also combining characters
 dictionary symbols 807
 digit form names 373
 digits 207
 Arabic 777
 Arabic-Indic 373–374
 compatibility 777
 decimal 177
 glyph variants 779
 hexadecimal 777
 Myanmar 777
 national shapes 837
 Shan 777
 superscript and subscript 778
 Tai Laing 777
 Tai Tham 777
 digraphs 298, 301, 303
 dingbats 809–811
 directionality 20, 53
 East Asian scripts 670
 Middle Eastern scripts 361
 Mongolian 528
 musical symbols 751
 normative property 173
 Ogham 358
 Old Italic 348
 Philippine scripts 651
 Runic 351
 discussion list for Unicode 889
 Dogri 462
 Domino Tiles 812
 dotless i 240, 295
 dotted circle
 in code charts 106, 330
 in fallback rendering 224
 to indicate diacritic 55
 to indicate vowel sign placement 56
 double diacritics 113, 192, 331
 Duployan 759–760
 reference materials 946
 Dutch 297, 298
 dynamic composition
 as Unicode design principle 22
 Dzongkha 515

E

East Asian scripts 669–712
 writing direction 53
 see also CJK ideographs
 Eastern Arabic-Indic digits 373
 EBCDIC
 newline function 212
 see UTR #16, UTF-EBCDIC
 editing, text boundaries for 219–220
 efficiency
 as Unicode design principle 15
 Egyptian hieroglyphs 434–438
 reference materials 947
 Elbasan 355
 reference materials 947
 ellipsis 277–278
 e-mail discussion list for Unicode 889
 emoji 805, 889
 animal symbols 808
 charts 889
 cultural symbols 808
 zodiacal symbols 808
 emoji modifiers 809
 emoticons 809
 Enclosed Alphanumerics 818
 enclosing marks 336
 definition 106
 encoded characters 7, 29
 allocation 44–52
 definition 90
 encoding form conversion
 definition 126
 encoding forms 33–39
 ISO/IEC 10646 definitions 900
 encoding forms, Unicode
 see Unicode encoding forms
 encoding model for Unicode characters 33, 42
 see also UTR #17, Unicode Character Encoding
 Model
 encoding schemes 40–43
 encoding schemes, Unicode
 see Unicode encoding schemes
 endian ordering
 see byte order mark (BOM) (U+FEFF)
 end-user subarea 843
 English 297
 equivalent sequences 208
 as Unicode design principle 23
 case-insensitivity 233, 242
 combining characters in matching 221
 conformance 82
 Hangul syllables 701
 in sorting and searching 232

- language-specific 117
 security implications 248
see also canonical equivalence
see also compatibility equivalence
see also encoding forms, encoding schemes
- errata xxxvi, 76, 890
 escape sequences 823
 not used in Unicode 1, 4
 Esperanto 298
 Estonian 298
 Ethiopic 714–717
 reference materials 947
 Etruscan 347
 European scripts 291–338
 ancient 339–359
 eyelash-RA 454
- F**
- fallback rendering 255
 of nonspacing marks 224
 FAQ (Frequently Asked Questions) 889
 Faroese 297
 Farsi 369, 372
 featural syllabaries 261
 FF (U+000C form feed) 211, 823
 file separator (U+001C) 823
 Finnish 297
 Finno-Ugric Transcription (FUT)
 see Uralic Phonetic Alphabet (UPA)
 fixed-width Unicode encoding form (UTF-32) ... 35,
 123
 flat tables 198
 Flemish 297
 fleurons 811
 fonts
 and Unicode characters 16
 for mathematical alphabets 772–774
 style variation for symbols 763
 form feed (U+000C) (FF) 211, 823
 format control characters 30, 68, 267, 821–837
 deprecated 836–837
 prefixed 194, 333
 stateful 834
 fraction characters 788
 fraction slash (U+2044) 277, 784
 French 298
 Frisian 298
 FTP site, Unicode Consortium 889
 fullwidth forms in East Asian encodings 698
 futhark 350
- Ge'ez 714
 General Category (normative property) 174
 list of values 174
 general punctuation 265–289
 General Scripts Area 50
 geometrical symbols 802–804
 Georgian 323–324
 German 297
 geta mark (U+3013) 287
 Glagolitic 320
 reference materials 947
 Glossary 890
 glyph selection tables 198
 glyphs 6, 15
 characters, relationship to 15
 diacritics alternative 293, 329
 Greek alternative 307–309
 Latin alternative 293
 mathematical alternative 790
 missing 255
 representative in code charts 858
 standardized variants 838
 symbols alternative 763
 golden numbers 352
 Gothic 354
 reference materials 948
 Grantha 599–601
 reference materials 948
 grapheme base 329
 definition 107
 grapheme clusters 11, 60–61
 see also UAX #29, Unicode Text Segmentation
 default 219
 definition 108
 grapheme extender
 definition 108
 grapheme joiner, combining (U+034F) 831
 graphic characters 30
 Greek 306–311
 acrophonic numerals 207, 311
 alternative glyphs 307–309
 ancient musical notation 756–758
 editorial marks 282
 letters as symbols 307–309, 791
 see also Cypriot, Linear B
 Greek editorial marks
 reference materials 948
 Greenlandic 298
 group separator (U+001D) 823
 guillemets 272
 Gujarati 478
 Gurmukhi 473–477
- G**
- Garshuni 392

H

Hakka 694
 halant 443
 see also virama
 half marks, combining 192, 337
 half-consonants, Indic 450
 halfwidth forms in East Asian encodings 698
 Han ideographs *see* CJK ideographs
 Han unification 678–685
 and language tags 217
 history 913–922
 language usage 675
 source separation rule 673, 679
 source standards 918–922
 hand symbols 808
 Hangul Area 50
 Hangul syllables 669, 699–702
 and combining marks 113
 as grapheme clusters 61
 canonical decomposition 144
 collation 701
 composition 146
 conjoining jamo 142–151
 equivalent sequences 701
 Hangul Compatibility Jamo 700
 Hangul Jamo 699–702
 Hangul Syllables block 701–702
 Johab set 701
 name generation 147
 normalization 700
 standard 143
 Hangzhou numerals 784
 Hanja *see* CJK ideographs
 Hanunóo 650
 Hanzi *see* CJK ideographs
 harakat 370
 hasant 467
 hash tables 199
 Hatran 426
 reference materials 948
 Hebrew 363–368
 hentaigana 697
 hieroglyphs
 Anatolian 441–442
 Egyptian 434–438
 Meroitic 439–440
 high surrogate
 definition 118
 high-surrogate code points 79, 844
 high-surrogate code units 118
 higher-level protocols
 definition 92
 Hindi 445

Hiragana 696
 horizontal tab (U+0009) 823
 HTML newline function 212
 Hungarian 298
 hyphenation 826
 as a text process 10
 hyphens 269, 826

I

I Ching symbols 813
 IANA charset names 41
 Icelandic 297
 identifiers 231
 see also UAX #31, Unicode Identifier and Pattern
 Syntax
 Ideographic (informative property) 190
 ideographic description sequences 690
 Ideographic Rapporteur Group (IRG) 916
 Ideographic Variation Database *see* UTS #37, Unicode
 Ideographic Variation Database
 ideographs *see also* CJK ideographs
 IDNA *see* UTS #46, Unicode IDNA Compatibility
 Processing
 IICore 673, 916
 ill-formed
 definition 121
 Imperial Aramaic 414–415
 reference materials 948
 implementation guidelines 197–257
 in a Unicode encoding form
 definition 122
 in-band mechanisms 856
 India
 Official scripts 443–505
 Indian rupee signs, usage 767
 Indic scripts 443–505
 principles, in terms of Devanagari 446–453
 relation to ISCII standard 445
 Indonesia and Oceania
 scripts of 649–668
 Indonesian 297
 industry character sets
 covered in Unicode 3
 information separators (U+001C..U+001F) 823
 informative properties
 definition 99
 Inscriptional Pahlavi 420
 Inscriptional Parthian 420
 inside-out rule 224
 interchange restrictions 31
 International Phonetic Alphabet (IPA) 260, 300–301
 reference materials 949
 Spacing Modifier Letters 326

see also phonetic alphabets
 internationalization 18
 Internationalization & Unicode Conference 889
 Internet protocols
 UTF-8 as preferred encoding 37
 Inuktitut 740
 invisible operators 796
 iota subscript 307
 IPA *see* International Phonetic Alphabet
 IRG (Ideographic Rapporteur Group) 916
 Irish 297, 358
 ISCII standard and Unicode 445
 ISO/IEC 10646 893–905
 conformance of Unicode implementations .. 905
 encoding forms 900
 synchrony with Unicode Standard 902
 timeline compared to Unicode versions 895
 Italian 297
 ITC Zapf Dingbats 809
 IUC *see* Internationalization & Unicode Conference

J

jamos *see* Hangul syllables
 Japanese 669
 Javanese 661–664
 reference materials 950
 Jawi 388
 jihvamuliya 466, 580
 Johab 701
 joiners 371
 combining grapheme joiner (U+034F) 831
 word joiner (U+2060) 825
 zero width joiner (U+200D) 371–372, 828
 justification 226

K

Kaithi 577–579
 reference materials 950
 Kana (Hiragana and Katakana) 696–697
 Kanbun 686
 KangXi radicals 684, 687–688
 Kanji *see* CJK ideographs
 Kannada 495–498
 Kashmiri 463
 Katakana 696–697
 Kawi 655, 657
 Kayah Li 643
 reference materials 950
 KC (normalization form)
 see Normalization Form KC
 KD (normalization form)
 see Normalization Form KD
 keypad labels 797

Khamti Shan 618
 Kharoshthi 556–557
 reference materials 950
 Khmer 621–631
 characters not recommended 628
 syllable components, order of 629
 Khojki 588–589
 reference materials 951
 Khudawadi 590–591
 reference materials 951
 killer 262
 Batak 666
 Brahmi 552
 Meetei Mayek 540
 Myanmar (asat) 616
 see also virama
 Konkani 462
 Korean Hangul *see* Hangul
 Kurdish 388

L

Ladino 363
 language tags 217, 852–856
 and Han unification 217
 use strongly discouraged 852, 855
 Lanna 634
 Lao 611–613
 last-resort glyphs 255
 Latin 293–305
 alternative glyphs 293
 Basic Latin 297
 encoding blocks 45
 IPA Extensions 300–301
 Latin Extended Additional 303–305
 Latin Extended-A 297
 Latin Extended-B 298–300
 Latin Extended-C 303
 Latin Extended-D 304
 Latin Extended-E 305
 Latin Ligatures 303
 Latin-1 Supplement 297
 Phonetic Extensions 302–305
 Latvian 298, 305
 cedilla 294
 layout control characters 68, 825–835
 leading surrogates
 see high-surrogate code units
 legibility criterion for plain text 19
 Lepcha 547–549
 reference materials 951
 letter spacing 826
 letterlike symbols 768–774
 LF (U+000A line feed) 211, 823

- ligatures 827–831
 - Arabic 380–381
 - combining characters on 59
 - control characters for 193
 - for nonspacing marks 228
 - Latin 303
 - selection 220
 - Syriac 399
 - Limbu 536–539
 - reference materials 952
 - line breaking 211–215, 825–827
 - control characters 192
 - in South Asian scripts 609, 617, 631
 - recommendations 213
 - see also* UAX #14, Unicode Line Breaking Algorithm
 - line feed (U+000A) (LF) 211, 823
 - line separator (U+2028) (LS) 211, 827
 - line tabulation (U+000B) (VT) 823
 - Linear A 341
 - reference materials 952
 - Linear B 342–343
 - reference materials 953
 - see also* Cypriot
 - linear boundaries 220
 - Lisu 706–708
 - reference materials 953
 - Lithuanian 298
 - little-endian 40
 - definition 83
 - Locale Data Markup Language
 - see* UTS #35, Unicode Locale Data Markup Language (LDML)
 - logical order
 - as Unicode design principle 19
 - exceptions to 170
 - logograph 262
 - logosyllabaries 262
 - low surrogate
 - definition 118
 - low-surrogate code points 79, 844
 - low-surrogate code units 118
 - lowercase 164, 238, 291
 - LS (U+2028 line separator) 211, 827
 - Lycian 345
 - reference materials 954
 - Lydian 345
 - reference materials 954
- M**
- MacOS newline function 212
 - Mahajani 586–587
 - reference materials 954
 - Mahjong Tiles 812
 - mail discussion list for Unicode 889
 - Maithili 462
 - major version 75
 - Malay 297
 - Malay, Patani 610
 - Malayalam 499–505
 - Maltese 298
 - Manchu 528
 - Mandaic 403–404
 - reference materials 954
 - Mandarin 677
 - Manden 722
 - Manichaean 416–419
 - reference materials 955
 - map symbols 807
 - mapping tables *see* tables of character data
 - Marathi 445, 454, 460
 - Marchen 571
 - reference materials 955
 - markup languages
 - and Unicode conformance 856
 - line breaking 211
 - Mathematical (informative property) 788
 - mathematical expression format characters 194
 - see also* UTR #25, Unicode Support for Mathematics
 - mathematical symbols 788–795
 - alphabets 770–774
 - alphanumeric 769–774
 - fonts 772–774
 - format characters 796
 - fragments for typesetting 798
 - invisible operators 796
 - operators 789–792
 - reference materials 955
 - standardized variants 795
 - MathML 792
 - matras 168, 447
 - Meetei Mayek 540–541
 - reference materials 955
 - Mende Kikakui 732–733
 - reference materials 955
 - Meroitic
 - cursive 439–440
 - hieroglyphs 439–440
 - reference materials 956
 - Miao 709–710
 - reference materials 956
 - Middle Eastern scripts 361–510
 - ancient 407–426
 - Min 677
 - Minnan (Hokkien/Fujian, incl. Taiwanese) 694
 - minor version 75

- minus sign 791
 - commercial (U+2052) 280
 - mirrored property
 - see* Bidi Mirrored (normative property)
 - mirroring of paired punctuation 271
 - Miscellaneous Symbols 806
 - missing glyphs 255
 - Modi 596–598
 - reference materials 956
 - modifier letters 325–328
 - Modifier Letters, Spacing 302
 - Mongolian 527–535, 564
 - writing direction 528
 - Mro 542
 - reference materials 956, 957
 - Multani 592
 - reference materials 957
 - multibyte encodings
 - compared to UTF-8 37
 - multistage tables 198
 - musical symbols 750–758
 - ancient Greek 756–758
 - Balinese 659
 - Byzantine 755
 - directionality 751
 - Gregorian 753
 - Kievan 754
 - reference materials 957
 - Western 750–754
 - Myanmar 614–620
 - digits 777
 - Myanmar Extended-A 618
 - Myanmar Extended-B 618
 - reference materials 958
- ## N
- N’Ko 722–726
 - reference materials 959
 - Nabataean 424
 - reference materials 958
 - named character sequences 183
 - names, character *see* character names
 - namespace 89
 - NEL (U+0085 next line) 211, 823
 - Nepali 445
 - neutral directional characters 173
 - New Tai Lue 634–636
 - Newa 513–514
 - reference materials 958
 - newline function (NLF) 212, 824
 - newline guidelines 211–215
 - next line (U+0085) (NEL) 211, 823
 - NFC (Normalization Form C) 62
 - NFD (Normalization Form D) 62
 - NFKC (Normalization Form KC) 62
 - NFKD (Normalization Form KD) 62
 - NLF (newline function) 212, 824
 - no-break space (U+00A0) 825
 - base for diacritic in isolation 60, 269, 330
 - no-break space, narrow (U+202F) 533
 - noncharacter code points *see* noncharacters
 - noncharacters 31, 845
 - conformance 79
 - definition 92
 - handling 82
 - in code charts 861
 - interchange restrictions 31
 - semantics 32
 - U+10FFFF (not a character code) 845
 - U+FDD0..U+FDEF 31, 845
 - U+FFFE (not a character code) 67, 846
 - U+FFFF (not a character code) 31, 845
 - nondecomposable characters 64
 - non-joiner, zero width (U+200C) 371–372, 829
 - nonlinear boundaries 220
 - non-overlap principle in Unicode encoding forms 33
 - nonspacing marks 329
 - definition 106
 - display in isolation (U+202F) 60, 269, 330
 - positioning 228
 - rendering 224–229
 - see also* combining characters
 - see also* diacritics
 - normalization 62, 208–209
 - and case operations 244
 - canonical ordering algorithm 62, 137, 168
 - conformance 84
 - of private-use characters 842
 - see also* UAX #15, Unicode Normalization Forms
 - stability 134
 - Normalization Form C (NFC) 62
 - Normalization Form D (NFD) 62
 - Normalization Form KC (NFKC) 62
 - Normalization Form KD (NFKD) 62
 - normalization forms 134–141
 - definition 140
 - specification 136
 - normative behaviors
 - definition 87
 - normative properties
 - definition 98
 - list 99
 - may change 98
 - Norwegian 297
 - notational conventions 875–879
 - notational systems 263, 747–762
 - nukta 370, 390, 455

null (U+0000)
 as Unicode string terminator 824
 number forms
 CJK ideographs 207
 numbers
 Coptic Epact 781
 handling 207
 ideographic accounting 178
 numerals 775–785
 acrophonic 311
 Chinese counting rods 786
 Coptic 315
 Cuneiform 431
 Ethiopic 716
 Greek acrophonic 207
 Hangzhou 784
 Meroitic cursive 440
 old-style 277
 Roman 207, 788
 Rumi 782
 Suzhou-style 784
 numeric separators 280
 numeric shape selectors (deprecated) 837
 Numeric Type (normative property) 177
 Numeric Value (normative property) 177
 numero sign (U+2116) 768

O

object replacement character (U+FFFC) 851
 octet 877
 Ogham 358
 reference materials 959
 Ol Chiki 544–545
 reference materials 959
 Old Church Slavonic 316
 Old Hungarian 353
 reference materials 959
 Old Italic 347–349
 reference materials 959
 Old North Arabian 409
 reference materials 960
 Old Permic 357
 reference materials 960
 Old Persian 433
 reference materials 960
 Old South Arabian 410–411
 reference materials 960
 Old Turkic 572
 reference materials 961
 old-style numerals 277
 Oriya 480–482
 ornamental dingbats 811
 Oromo 714

Osage 742
 reference materials 959
 Osmanya 718
 reference materials 961
 out-of-band mechanisms 856
 overlapping encodings 33
 overscores 277

P

Pahawh Hmong 646–647
 reference materials 961
 Pahlavi, Inscriptional 420
 reference materials 949
 Pahlavi, Psalter 421
 Palmyrene 425
 reference materials 962
 Panjabi 473
 paragraph or section marks 280
 paragraph separator (U+2029) (PS) 211, 827
 Parthian, Inscriptional 420
 reference materials 949
 Pashto 369
 Patani Malay 610
 Pau Cin Hau 648
 reference materials 962
 Persian 369, 372
 Phags-pa 564–570
 reference materials 962
 Phaistos Disc symbols 815
 Phake 620
 Philippine scripts 650–652
 reference materials 963
 Phoenician 412
 reference materials 963
 phonemes 263
 phonetic alphabets 260
 IPA Extensions 300–301
 Phonetic Extensions 302–305
 Spacing Modifier Letters 326–328
 Uralic Phonetic Alphabet (UPA) 280, 302
 see also International Phonetic Alphabet (IPA)
 Pinyin 297
 pivot code, Unicode as 198
 plain text
 as Unicode design principle 18
 legibility criterion 19
 planes of Unicode codespace 44
 Plane 0 (BMP) 44
 Plane 1 (SMP) 44, 51
 Plane 14 (SSP) 45
 Plane 2 (SIP) 44, 52
 Planes 15–16 (Private Use) 52, 843
 Playing Cards 812

- points, Hebrew pronunciation marks 363
 - policies of the Unicode Consortium 890
 - Polish 298
 - Portuguese 297
 - precomposed characters
 - see decomposable characters
 - compatibility *see* compatibility decomposable characters
 - prefixed format control characters 194
 - prepended concatenation marks 256, 333
 - Private Use Area (PUA) 50, 842
 - Private Use planes 45, 52, 843
 - private-use characters
 - properties 841
 - semantics 32
 - private-use code points 31, 203
 - conformance 80
 - definition 104
 - high surrogates 844
 - processing code, Unicode as 38
 - properties 18, 94–104, 159–195
 - aliases 162
 - aliases (definition) 103
 - and Unicode algorithms 98
 - data tables 198
 - derived *see* derived properties
 - in Unicode Character Database (UCD) 46
 - informative *see* informative properties
 - normative references to 77, 84
 - normative *see* normative properties
 - of control codes 823
 - provisional *see* provisional properties
 - simple *see* simple properties
 - see also individual properties, e.g. combining classes*
 - property values
 - aliases 162
 - aliases (definition) 104
 - default 96
 - default (definition) 96
 - normative references to 84
 - PropertyAliases.txt 103, 878
 - PropertyValueAliases.txt 104, 878
 - PropList.txt 166
 - Provençal 298
 - provisional properties
 - definition 100
 - PS (U+2029 paragraph separator) 211, 827
 - Psalter Pahlavi 421
 - reference materials 964
 - PUA (Private Use Area) 50, 842
 - pulli* 483
 - punctuation 265–289
 - blocks containing 259
 - CJK 286
 - doubled 277
 - in bidirectional text 265
 - paired 271
 - small form variants 288
 - typographic forms 265
 - vertical forms 288
 - Punctuation and Symbols Area 50
 - Punjabi 473
- ## Q
- quotation marks 272–275
 - East Asian 274
 - European 272
- ## R
- radicals, KangXi and other CJK 687–688
 - radical-stroke index 684
 - record separator (U+001E) 823
 - recycling symbols 807–808
 - referencing 84
 - properties 77
 - Unicode algorithms 78
 - Unicode Standard 76
 - regional indicator symbols 819
 - regular expressions 216
 - and line breaking 211
 - see also* UTS #18, Unicode Regular Expressions
 - Rejang 665
 - reference materials 964
 - rendering of text 6, 10, 17
 - fallback 255
 - unsupported characters 203
 - repertoire of abstract characters 29
 - replacement character (U+FFFD) . . . 43, 68, 83, 127, 257, 851
 - reserved code points 30, 203
 - definition 92
 - in code charts 861
 - preservation in interchange 31
 - see also* unassigned code points
 - Rhaeto-Romanic 298
 - rich text 18
 - right single quotation mark (U+2019)
 - preferred for apostrophe 276
 - right-to-left text 53
 - East Asian scripts 670
 - Middle Eastern scripts 361
 - roadmap for script additions 46
 - Roman numerals 207, 788
 - Romanian 298
 - comma below 295
 - Romany 298

- Rong 547–549
- Rumi numeral symbols 782
- Runic 350–352
reference materials 964
- Russian 316
- S**
- Samaritan 401–402
reference materials 964
- Sami 298
- Sanskrit 445
- Saurashtra 550
reference materials 965
- scalar values, Unicode
see Unicode scalar values
- scripts
in Unicode Standard 3
roadmap for future additions 46
types of 264
see also UAX #24, Unicode Script Property
- SCSU
see UTS #6, A Standard Compression Scheme for Unicode
- searching 232–234
as a text process 10
case-insensitive 233, 242
- section or paragraph marks 280
- security issues 248
- self-synchronization of encoding forms 34
- semantics
see character semantics
- sequences
notation 876
- Serbian
corresponding digraphs in Croatian 298
- Shan 632
digits 777
- Sharada 580–581
reference materials 965
- Shavian 359, 706
reference materials 965
- Show Hidden 81, 224, 255, 839
- SHY (U+00AD soft hyphen) 826
- Sibe 528
- Siddham 584–585
reference materials 965
- signature for Unicode data 67, 847–849
- simple properties
definition 103
- simplified Chinese 676
- Sindhi 369, 462
- Sinhala 511–512
reference materials 966
- Sinological dot 304
- SIP (Supplementary Ideographic Plane) 44, 52
- slash, fraction (U+2044) 277
- Slovak 298
- Slovenian 298
- small letters 164, 238, 291
- SMP (Supplementary Multilingual Plane) 44, 51
- soft hyphen (U+00AD) (SHY) 826
- Somali 718
- Sora Sompeng 604
reference materials 966
- Sorbian 298
- sorting 12, 232
and combining grapheme joiner 832
as a text process 10
case-insensitive 233
culturally expected 12, 232
language-insensitive 232
see also Unicode Collation Algorithm (UCA)
- source separation rule 673, 679
- South and Central Asian scripts
Ancient 551–572
Other historic 573–604
Other modern 507–550
- South Asian scripts 443–539
- Southeast Asian scripts 605–648
- space (U+0020)
base for diacritic in isolation 60, 269, 330
- space characters 268, 825–827
graphics for 797
- space, zero width (U+200B) 268
- spacing clones of diacritics 327, 331
- spacing marks 329
definition 107
- Spacing Modifier Letters 326–328
- Spanish 297
- special characters 67, 821–856
- SpecialCasing.txt 152, 166
- Specials 847–851
- spell-checking
as a text process 11
- spellings, alternative
see equivalent sequences
- spoofing 248
- SSP (Supplementary Special-purpose Plane) 45
- stability 101, 161
as Unicode design principle 23
- stacked boundaries 219
- stacking sequences 57
nondefault 58
- Standard Compression Scheme for Unicode (SCSU)
see UTS #6, A Standard Compression Scheme for Unicode
- standardized variants 531, 838

- in the code charts 868
 - mathematical symbols 795
 - StandardizedVariants.txt 531, 795
 - standards coverage 3
 - starters 136
 - stateful encoding
 - not used in Unicode 4
 - paired format controls 834
 - string comparison 12
 - string literals, Unicode
 - code point notation `\u1234` 878
 - strings, Unicode 43, 120
 - null termination 824
 - strong directional characters 173
 - styled text 18
 - sublinear searching 233
 - subsets, supported 71
 - conformance 80
 - ISO/IEC 10646 specification for 903
 - substitution character
 - see replacement character
 - Sumero-Akkadian 428–431
 - Sumero-Akkadian Cuneiform
 - reference materials 966
 - Sundanese 667–668
 - reference materials 967
 - superscripts 327
 - and subscripts 786
 - supplementary characters
 - in UTF-16 strings 43
 - tables for 199
 - Supplementary General Scripts Area 50
 - Supplementary Ideographic Plane (SIP) 44, 52
 - Supplementary Multilingual Plane (SMP) 44, 51
 - supplementary planes
 - representation in UTF-16 36
 - representation in UTF-8 37
 - Supplementary Private Use Areas 52, 843
 - Supplementary Special-purpose Plane (SSP) 45
 - supported subsets 71
 - conformance 80
 - supralineation 314
 - surrogate code points
 - see surrogates
 - surrogate pairs 36, 124
 - definition 118
 - processing 38, 205–206
 - surrogates 31, 118, 844
 - interchange restrictions 31
 - isolated surrogates, handling 43
 - isolated surrogates, ill-formed 124
 - isolated surrogates, uninterpreted 118
 - support levels 205
 - Surrogates Area 50, 844
 - Sutton SignWriting 761–762
 - reference materials 967
 - Suzhou-style numerals 784
 - svasti signs 522
 - Swahili 297
 - Swedish 297
 - syllabaries 261
 - alphabetic property 190
 - featural 261
 - Syloti Nagri 575–576
 - symbols 763–820
 - animal 808
 - appearance variation 763
 - arrows 794–795
 - box drawing 802
 - cultural 808
 - currency symbols block 765–767
 - dictionary 807
 - dingbats 809–811
 - emoji 805, 819
 - Enclosed Alphanumerics 818
 - fragments for mathematical typesetting 798
 - game 808
 - gender 808
 - genealogical 808
 - geometrical 802–804
 - hand 808
 - Khmer lunar calendar 631
 - letterlike 768–774
 - map 807
 - mathematical 788–795
 - mathematical alphanumeric 769–774
 - miscellaneous 806
 - musical 750–758
 - numerals 775–785
 - recycling 807–808
 - regional indicator 819
 - technical 797–801
 - weather 807
 - zodiacal 808
 - symmetric swapping format characters 836
 - Syriac 392–399
 - reference materials 967
- ## T
- tab (U+0009 character tabulation) 823
 - tab, vertical (U+000B) 211, 823
 - tables of character data 198–199
 - optimization 199
 - supplementary characters 199
 - tag characters 852–856
 - Tagalog 650
 - Tagbanwa 650

- tags, language 217, 852–856
 - use strongly discouraged 855
 - Tai Laing
 - digits 777
 - Tai Le 632–633
 - reference materials 967
 - Tai Tham 637–639
 - digits 777
 - reference materials 968
 - Tai Viet 640–642
 - Tai Xuan Jing symbols 814
 - Takri 582–583
 - reference materials 968
 - Tamil 483–491
 - Tangut 711–712
 - components 712
 - radicals 712
 - reference materials 968
 - tashkil 370
 - tashkil, harakat, points 372
 - TCHAR in Win32 API 202
 - Technical Notes (UTN) 888
 - Technical Reports (UTR) 883
 - abstracts 886
 - Technical Standards (UTS) xxxv, 883
 - abstracts 884
 - technical symbols 797–801
 - Telugu 492–494
 - terminal emulation 764
 - text boundaries 61, 191, 219–220, 230
 - see also* UAX #14, Unicode Line Breaking Algorithm
 - see also* UAX #29, Unicode Text Boundaries
 - text elements 6, 10, 219
 - boundaries 230
 - for sorting 232
 - variable-width nature 38
 - text processes 6, 10–13
 - text rendering 6, 10, 17
 - text selection, boundaries for 219–220
 - Thaana 509–510
 - reference materials 968
 - Thai 607–610
 - Tibetan 515–526
 - Tifinagh 719
 - Tigre 714
 - tilde (U+007E) 280
 - Tirhuta 593–595
 - reference materials 968
 - titlecase 164, 238
 - Todo 528
 - tone letters 327–328
 - tone marks
 - Bopomofo spacing 693, 694
 - Chinantec 328
 - Chinese 328
 - Tai Le 632
 - Thai 607
 - Vietnamese 296
 - traditional Chinese 676
 - traffic signs 807
 - trailing surrogates
 - see* low-surrogate code units
 - transcoding 198–199
 - tables 198
 - Transport and Map Symbols 809
 - triangulation in transcoding 198
 - tries 198
 - truncation
 - combining character sequences 222–223
 - surrogates and 206
 - Turkish 298
 - case mapping of I 240, 295
 - cedilla 295
 - lira sign 767
 - two-stage tables 199
- ## U
- U+ notation 878
 - U+10FFFF (not a character code) 845
 - U+FEFF (BOM) 847–849
 - U+FFFE (not a character code) 846
 - U+FFFF (not a character code) 845
 - UAX (Unicode Standard Annex) xxxiii, 883
 - as component of Unicode Standard 79
 - conformance 85
 - list of 85
 - UCA *see* Unicode Collation Algorithm
 - UCD *see* Unicode Character Database
 - UCS (Universal Character Set)
 - see* ISO/IEC 10646
 - UCS-2 900
 - UCS-4 900
 - Ugaritic 432
 - reference materials 969
 - Ukrainian 316
 - unassigned code points 30, 79, 203
 - defined as reserved code points 92
 - handling 74
 - properties of 96
 - semantics 79
 - see also* reserved code points
 - underscores 277
 - undesignated code points 30
 - Unicode 1.0 Name (informative property) 189
 - Unicode algorithms
 - and properties 98

- conformance 84
- definition 92
- normative references to 78, 84
- Unicode Bidirectional Algorithm 20, 53
 - see also* UAX #9, Unicode Bidirectional Algorithm
- Unicode Character Database (UCD) . xxxv, 161, 890
 - as component of Unicode Standard 79
 - changes 74
 - properties in 46
- Unicode character encoding model 33, 42
 - see also* UTR #17, Unicode Character Encoding Model
- Unicode character literals
 - code point notation U+ 878
- Unicode codespace
 - allocation numbers 908
 - definition 90
 - planes 44
 - size 1, 29
- Unicode Collation Algorithm (UCA) 12
 - see also* UTS #10, Unicode Collation Algorithm
- Unicode Common Locale Data Repository (CLDR) . 890
- Unicode conferences 889
- Unicode Consortium 882
 - addresses 891
 - Consortium membership in standards bodies 882
 - e-mail discussion list 889
 - FTP site 889
 - membership 882
 - policies 890
 - website 889
- Unicode data signature 67, 847–849
- Unicode data types 201–202
 - for C 201–202
- Unicode encoding forms 119–126
 - advantages of each 38
 - conformance 34, 82
 - definition 120
 - fixed-width (UTF-32) 35, 123
 - signatures 848, 849
 - variable-width 36, 124
 - see also* encoding forms
- Unicode encoding schemes
 - conformance 130–133
 - definition 130
 - endian ordering 40
 - see also* encoding schemes
- Unicode escape sequence notation \u1234 878
- Unicode Regular Expressions *see* UTS #18, Unicode Regular Expressions
- Unicode scalar values
 - definition 119
- Unicode security 248
 - see also* UTS #39, Unicode Security Mechanisms
- Unicode Standard
 - allocation of encoded characters 44–52
 - architecture 10–13
 - areas 45
 - benefits 1
 - blocks 45, 259
 - code charts 857–873, 889
 - components 79
 - conformance 73–158
 - conformance of ISO/IEC 10646 implementations 905
 - corrections 76
 - definitions for conformance 87–92
 - design goals 4
 - design principles 14–24
 - errata 76, 890
 - normative references to 76, 84
 - number of characters 3
 - number of code points 1, 29
 - script coverage 3
 - security issues 248
 - synchrony with ISO/IEC 10646 902
 - updates 890
 - versions *see* versions of the Unicode Standard
 - see also* Version 9.0
- Unicode Standard Annexes (UAX) xxxiii, 883
 - as components of Unicode Standard 79
 - conformance 85
 - list of 85
- Unicode string literals
 - code point notation \u1234 878
- Unicode strings 43
 - definition 120
- Unicode Technical Committee (UTC) 882
- Unicode Technical Notes (UTN) 888
- Unicode Technical Reports (UTR) 883
 - abstracts 886
- Unicode Technical Standards (UTS) xxxv, 883
 - abstracts 884
- UnicodeData.txt 152, 166
- unification
 - as Unicode design principle 21
 - see also* Han unification
- Unified Repertoire and Ordering (URO) .. 679, 915
 - see also* Han unification
- UniHan Database 161, 683, 684, 870, 890, 916
- UniHan.zip 101, 161
- unit separator (U+001F) 823
- Universal Character Set (UCS)
 - see* ISO/IEC 10646
- universality
 - as Unicode design principle 14
- Unix

and UTFs 38
 newline function 212
 UTF-32 in 35
 UTF-8 in 18
 unsupported characters 203
 upadhamaniya 466, 580
 update version 75
 uppercase 164, 238, 291
 Uralic Phonetic Alphabet (UPA) 280, 302
 Urdu 369
 URO (Unified Repertoire and Ordering) ... 679, 915
 see also Han unification
 UTF, Unicode Transformation Formats 33, 120
 advantages of each 38
 as encoding form or scheme 133
 binary comparison and sort order differences ..
 233, 235
 in APIs 202
 UTF-16 36, 124, 901
 binary comparison and sort order caution ... 36
 bit distribution (table) 124
 BOM in 131, 847
 encoding form (definition) 124
 encoding scheme (definition) 131
 encoding schemes 40
 in ISO/IEC 10646 901
 in UTF-8 order 236
 surrogates and string handling 43, 205
 UTF-16BE (Big-endian) 848
 encoding scheme 41
 encoding scheme (definition) 130
 UTF-16LE (Little-endian) 848
 encoding scheme 41
 encoding scheme (definition) 130
 UTF-32 35, 123
 as processing code 38
 BOM in 132
 encoding form (definition) 123
 encoding scheme (definition) 132
 encoding schemes 40
 in Unix 35
 UTF-32BE (Big-endian)
 encoding scheme 41
 encoding scheme (definition) 131
 UTF-32LE (Little-endian)
 encoding scheme 41
 encoding scheme (definition) 132
 UTF-8 36, 124, 901
 ASCII transparency 36
 binary comparison and sort order 39
 bit distribution (table) 125
 BOM in 130, 133, 848
 byte ranges 125
 compared to multibyte encodings 37

encoding form (definition) 124
 encoding scheme 40
 encoding scheme (definition) 130
 in Unix 18
 in UTF-16 order 235
 non-shortest form is invalid 124, 248
 preferred encoding for Internet protocols 37
 security and 248
 signature 130, 133, 848
 UTF-EBCDIC
 see UTR #16, UTF-EBCDIC
 UTN (Unicode Technical Note) 888
 UTR (Unicode Technical Report) 883
 abstracts 886
 UTS (Unicode Technical Standard) xxxv, 883
 abstracts 884
 Uyghur 369, 527, 564

V
 Vai 727–728
 reference materials 969
 valid (synonym for well-formed) 122
 variable-width Unicode encoding form 36, 124
 variants
 compatibility 26
 fullwidth and halfwidth 289
 mathematical symbols 795
 small form 288
 standardized 838
 variation selectors 195, 838
 ideographic variation mark (U+303E) 691
 Mongolian free variation selectors 531
 variation sequences 838
 for Phags-pa 568–570
 Version 9.0 79
 number of characters 3
 versions of the Unicode Standard xxxv, 74, 890, 907–
 908
 backward compatibility 74
 compared to ISO/IEC 10646 editions 907
 content 75
 interaction in implementations 203
 numbering 75
 property changes 74
 stability 74
 updates 890
 vertical tab (U+000B) 211, 823
 vertical text 53, 266, 288
 East Asian scripts 670
 Mongolian 528
 Vietnamese 296, 303
 ideographs 670
 virama 262, 443

definition	448
Kharoshthi	560
Khmer	624
Myanmar	615
Philippine scripts	650
virama-like characters	193
visual order used for Thai and Lao	21
vowel harmony	
Mongolian	532
vowel marks, Middle Eastern scripts	361
vowel separator	
Mongolian	533
vowel signs	
Indic	56, 447
Khmer	626
Philippine scripts	650

W

Warang Citi	543
reference materials	969
wchar_t	
and Unicode encoding forms	38
in C language	202
weak directional characters	173
weather symbols	807
website, Unicode Consortium	889
Weierstrass elliptic function symbol	769
well-formed	
definition	121
Welsh	298
Where Is My Character?	890
wide characters	
data type in C	202
wiggly fence (U+29DB)	793
Windows newline function	212
word breaks	221, 825–827
in South Asian scripts	609, 617, 631
word joiner (U+2060)	825
writing direction <i>see</i> directionality	
writing systems	260–264
Wu (Shanghainese)	677

X

Xibe	528
Xishuangbanna Dai	634

Y

Yi	703–705
reference materials	969
Yiddish	363
Yijing Hexagram Symbols	813
ypogegrammeni	307

Z

Zapf Dingbats	809
zero extension relation among encodings	900
zero width joiner (U+200D)	371–372, 828
zero width no-break space (U+FEFF)	67, 83, 825
initial	133, 848
zero width non-joiner (U+200C)	371–372, 829
zero width space (U+200B)	826
for word breaks in South Asian scripts	609, 617, 631
zero-width space characters	826
ZWJ <i>see</i> zero width joiner (U+200D)	
ZWNBSP <i>see</i> zero width no-break space (U+FEFF)	
ZWNJ <i>see</i> zero width non-joiner (U+200C)	
ZWSP <i>see</i> zero width space (U+200B)	