

This PDF file is an excerpt from *The Unicode Standard, Version 4.0*, issued by the Unicode Consortium and published by Addison-Wesley. The material has been modified slightly for this online edition, however the PDF files have not been modified to reflect the corrections found on the Updates and Errata page (<http://www.unicode.org/errata/>). For information on more recent versions of the standard, see <http://www.unicode.org/standard/versions/enumeratedversions.html>.

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 Addison-Wesley was aware of a trademark claim, the designations have been printed in initial capital letters. However, not all words in initial capital letters are trademark designations.

The Unicode® Consortium is a registered trademark, and Unicode™ is a trademark of Unicode, Inc. The Unicode logo is a trademark of Unicode, Inc., and may be registered in some jurisdictions.

The authors and publisher have taken care in preparation of this book, 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.

Dai Kan-Wa Jiten used as the source of reference Kanji codes was written by Tetsuji Morohashi and published by Taishukan Shoten.

Cover and CD-ROM label design: Steve Mehallo, <http://www.mehallo.com>

The publisher offers discounts on this book when ordered in quantity for bulk purchases and special sales. For more information, customers in the U.S. please contact U.S. Corporate and Government Sales, (800) 382-3419, corpsales@pearsontechgroup.com. For sales outside of the U.S., please contact International Sales, +1 317 581 3793, international@pearsontechgroup.com

Visit Addison-Wesley on the Web: <http://www.awprofessional.com>

Library of Congress Cataloging-in-Publication Data

The Unicode Standard, Version 4.0 : the Unicode Consortium /Joan Aliprand... [et al.].

p. cm.

Includes bibliographical references and index.

ISBN 0-321-18578-1 (alk. paper)

1. Unicode (Computer character set). I. Aliprand, Joan.

QA268.U545 2004

005.7'2—dc21

2003052158

Copyright © 1991–2003 by Unicode, Inc.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior written permission of the publisher or Unicode, Inc. Printed in the United States of America. Published simultaneously in Canada.

For information on obtaining permission for use of material from this work, please submit a written request to the Unicode Consortium, Post Office Box 39146, Mountain View, CA 94039-1476, USA, Fax +1 650 693 3010 or to Pearson Education, Inc., Rights and Contracts Department, 75 Arlington Street, Suite 300 Boston, MA 02116, USA, Fax: +1 617 848 7047.

ISBN 0-321-18578-1

Text printed on recycled paper

1 2 3 4 5 6 7 8 9 10—CRW—0706050403

First printing, August 2003

I.2 General Index

The General Index covers the contents of this book. To find topics in the Unicode Standard Annexes, Unicode Technical Standards, and Unicode Technical Reports, refer to their 4.0.0 versions on the CD-ROM accompanying this book, or use the search feature on the Unicode Web site for the latest versions.

For definitions of terms used in this book, see the *Glossary*. To find the code points for specific characters or the code ranges for particular scripts, see *Section I.1, Unicode Names Index*.

- A**
- abjads 148, 191
 - abstract character sequences
 - definition 64
 - abstract characters 12
 - definition 64
 - abugidas 149, 217, 265
 - accent marks *see* diacritics
 - accented characters
 - encoding 13
 - Latin 166
 - normalization 114
 - accounting numbers, ideographic 100
 - Aegean numbers 345
 - Afrikaans 168
 - Ainu 313
 - Algonquian 331
 - Ali Gali 325
 - aliases in code charts 415
 - allocation areas 36
 - allocation of encoded characters 35–42, 1356
 - Alphabetic (informative property) 102
 - alphabets 148
 - European 165–188
 - mathematical 354–357
 - Alpine 339
 - alternate formatting characters (deprecated) 103, 394–395
 - Amharic 322
 - angle brackets (U+2329 and U+232A)
 - deprecated for technical publication 366
 - Annexes, Unicode Standard (UAX) xxxiii, 51
 - abstracts 1343
 - as components of Unicode Standard 58
 - conformance 62
 - list of 62, 1343
 - annotation characters 402–404
 - use in plain text discouraged 403
 - ANSI/ISO C
 - wchar_t and Unicode 109
 - apostrophe (U+0027) 153, 159
 - Arabic 195–205
 - Arabic-Indic digits 196–197
 - signs used with 198
 - ArabicShaping.txt 199, 202, 210
 - archaic scripts 337–346
 - areas of the Unicode Standard 36
 - Armenian 180–181
 - arrows 363
- ASCII
- characters with multiple semantics 153
 - punctuation 152
 - transparency of UTF-8 30
 - Unicode modeled on 1
 - zero extension 109, 1350
- Asian Scripts Area 40
- Assamese 232
- assigned code points 12, 25
- Athapascan 331
- atomic character boundaries 121
- B**
- Bangla 232–233
 - base characters
 - definition 70
 - multiple 46
 - ordered before combining marks 123, 186
 - Basic Multilingual Plane (BMP) 1, 35
 - allocation areas 39
 - allocation of code points 1356
 - representation in UTF-16 29
 - Basque 168
 - benefits of Unicode 1
 - Bengali 232–233
 - Bidi Class (normative property) 98
 - Bidi Mirrored (normative property) 101
 - Bidi Mirroring Glyph (informative property) 101
 - BidiMirroring.txt 101
 - bidirectional algorithm, Unicode 19, 42, 61
 - see also* UAX #9, Bidirectional Algorithm
 - bidirectional ordering 19
 - controls 103, 392
 - bidirectional text 42, 61
 - Middle Eastern scripts 191
 - nonspacing marks in 126
 - punctuation in 152
 - big-endian 32
 - definition 61
 - Bihari 219
 - binary comparison and sort order
 - caution for UTF-16 30
 - UTF differences 133, 135
 - UTF-8 32
 - blocks of the Unicode Standard 36, 147
 - Blocks.txt 36, 1358
 - BMP *see* Basic Multilingual Plane
 - BNF (Backus-Naur Form) xxxv
 - BOM (U+FEFF) (byte order mark) 32, 47, 79–81, 400–402

- Boop, Betty1411
 Bopomofo 310–311
 boundaries, text 12, 47, 102, 121–122, 130
 see also UAX #14, Line Breaking Properties
 see also UAX #29, Text Boundaries
 boustrophedon 43, 341
 Brahmi 149, 217, 266
 Braille 374–375
 Breton 168
 Buhid 286
 Bulgarian 179
 bullets 160
 Burmese *see* Myanmar
 Byelorussian 179
 byte order mark (BOM) (U+FEFF) ... 32, 47, 79–81,
 400–402
 byte ordering
 changing 60
 conformance 61
 byte serialization 32, 47
- C**
- C language
 wchar_t and Unicode 109
 C0 and C1 control codes 25, 39, 102, 385
 Cambodian *see* Khmer
 Canadian Aboriginal Syllabics 331
 canonical composite characters
 see canonical decomposable characters
 canonical composition
 defined in UAX #15, Unicode Normalization
 Forms
 canonical decomposable characters
 definition 72
 canonical decomposition 21
 definition 72
 canonical decomposition mappings 71
 canonical equivalence
 definition 72
 nonspacing marks 127
 canonical mappings
 see canonical decomposition mappings
 canonical ordering of combining marks 84–85
 canonical precomposed characters
 see canonical decomposable characters
 canonical-equivalent character sequences
 conformance 59, 60
 Cantonese 298
 capital letters 96, 136, 165
 carriage return (U+000D) (CR) 116, 386
 carriage return and line feed (CRLF) 116
 case
 and text processes 13
 beyond ASCII 137
 case folding 138
 case operations (conformance) 62, 89–91
 case operations and normalization 139
 case operations, reversibility 138
 cased (definition) 89
 case-insensitive comparison ... 91, 133, 134, 138
 casing context (definition) 89
 conversion 90
 detection 90
 European alphabets 165
 exceptional Latin pairs 169, 170
 Georgian 182
 mapping tables 108
 mappings 89, 97, 136–138
 mappings noted in code charts 415
 Turkish I 137, 169
 Case (normative property) 96, 136
 CaseFolding.txt 97, 138, 139
 Catalan 168
 CD-ROM xxxiii–xxxiv
 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)
 character encoding forms (CEF) 26–32, 1350
 see also Unicode encoding forms
 character encoding model 26, 33
 see also UTR #17, Character Encoding Model
 character encoding schemes (CES) 32–35
 see also Unicode encoding schemes
 character encoding standards
 coverage by Unicode 3
 character literals, Unicode
 code point notation U+ xxxvi
 character mapping
 interchange format *see* UTR #22, Character Map-
 ping Markup Language (CharMapML)
 character names 1353
 aliases in code charts 415
 conventions xxxiv
 for control codes 102
 in code charts 413–415
 character properties *see* properties
 see also individual properties, e.g. combining
 classes
 character semantics 1, 63–64, 1353
 as Unicode design principle 17
 ASCII 153
 definition 64
 character sequences
 abstract *see* abstract character sequences
 canonical-equivalent *see* canonical-equivalent
 character sequences
 compatibility-equivalent *see* compatibility-equiv-
 alent character sequences
 conformance 59
 character shaping selectors (deprecated) 394
 character tabulation (U+0009) (HT) 386
 characters
 abstract *see* abstract characters
 adding to Unicode Standard 7
 arrangement in Unicode 42
 assigned 12, 25
 blocks 36, 147
 boundaries 121
 canonical decomposable *see* canonical decompos-
 able characters
 classes xxxvi
 code charts 413–418
 coded *see* encoded characters
 combining *see* combining characters
 compatibility decomposable *see* compatibility
 decomposable characters
 composite *see* decomposable characters

- concept of 15, 46
- conformance definitions 64–66
- confusable 141
- conversion 107–109
- decomposable *see* decomposable characters
- deprecated *see* deprecated characters
- encoded *see* encoded characters
- encoding forms *see* encoding forms
- encoding schemes *see* encoding schemes
- format control 25, 48, 154, 383–409
- glyphs, relationship to 15
- graphic 25
- identity (definition) 63
- interpretation 5, 59
- layout control 48, 387–392
- modification 60
- names list 413–415
- names *see* character names
- not encoded in Unicode 2
- number encoded in Version 4.0 2, 1356
- obsolete 65
- online charts 1346
- precomposed *see* decomposable characters
- properties *see* properties
- semantics *see* character semantics
- special 47, 383–409
- supplementary *see* supplementary characters
- transcoding 107–109
- unsupported 110–111
- characters, not glyphs
 - in spoofing 141
 - Unicode principle 15
- CharMapML
 - see* UTR #22, Character Mapping Markup Language (CharMapML)
- charsets
 - IANA registered names 33
- charts, character code *see* code charts
- Cherokee 330
- Chinese 297–299
 - Cantonese 298
 - Hakka 311
 - Mandarin 298
 - Minnan (Hokkien/Fujian, incl. Taiwanese) .. 311
 - simplified and traditional 297
- Chữ hán 297
- Chữ Nôm 1342
- citations for
 - properties 57
 - Unicode algorithms 57
 - Unicode Standard 57
- CJK ideographs 150, 293–305
 - accounting numbers 100
 - CJK Compatibility Ideographs 305
 - CJK Compatibility Supplement 305
 - CJK Unified Ideographs 293–304
 - CJK Unified Ideographs Extension A 295
 - CJK Unified Ideographs Extension B 304
- code charts 417
 - compatibility ideographs in Plane 2 37
 - component structure 301
 - encoding blocks 295
 - ideographic description sequences 307–309
 - ideographic variation mark (U+303E) 309
 - KangXi radicals 306, 1189
 - numeric values 100, 114
 - order of encoding 302
 - radicals 306
 - radical-stroke index 1189
 - source standards 293–295, 304
 - unknown or unavailable 161
 - Vietnamese 291
- CJK Miscellaneous Area 40
- CJK punctuation and symbols 160
 - compatibility forms 161
 - quotation marks 157
- CJK Radical (property) 307
- CJK-JRG (Chinese/Japanese/Korean Joint Research Group) 1341
- CJKV Ideographs Area 40
- cluster boundaries 121
- code charts 413–418
 - online 1346
 - representative glyphs 414
- code point sequences
 - notation xxxv
- code points 5, 24
 - assigned 12, 25
 - assigned on Basic Multilingual Plane (BMP) .. 1356
 - assigned on supplementary planes 1356
 - assignment 42, 1356
 - categories 25
 - default ignorable 111, 142
 - definition 64
 - designated 25
 - notation xxxiv
 - number in Unicode Standard 1
 - private-use *see* private-use code points
 - reserved *see* reserved code points
 - semantics 26
 - surrogate *see* surrogates
 - unassigned *see* unassigned code points
 - undesignated 25
- code positions *see* code points
- code set independence 15
- code unit sequences
 - definition 73
 - ill-formed (definition) 74
 - notation xxxv
 - well-formed (definition) 75
- code units
 - definition 73
 - isolated 73
- code values *see* code units
- coded character representations
 - definition 65
- coded character sequences
 - see* coded character representations
- coded characters *see* encoded characters
- codespace *see* Unicode codespace
- coeng 275, 277
- collation algorithm, Unicode (UCA) 14
- collation *see* sorting
- collation tables 108
- combining character sequences 44
 - canonical ordering 84–85
 - defective 126
 - definition 70
 - Latin 166
 - line-breaking 123

- matching123
 - order of base character and marks123, 186
 - rendering123
 - selection121
 - truncation 123–124
 - combining characters 43–46, 82–83, 122–124, 186–188
 - and sorting85
 - combining diacritical marks 186–188
 - definition70
 - display order44
 - in identifiers130
 - keyboard input123
 - ligatures46
 - multiple44
 - multiple base characters46
 - normalization of115
 - ordering conventions44
 - properties70
 - rendering of marks 125–129
 - typographical interaction44, 97
 - vertical stacking44
 - see also* diacritics
 - Combining Class (normative property)97
 - combining classes 82–90, 97, 127–128
 - class zero characters84, 97
 - definition83
 - combining grapheme joiner (U+034F)392
 - combining half marks103, 188
 - combining marks *see* combining characters
 - Compatibility and Specials Area23, 40
 - compatibility characters20, 23
 - mapping23
 - compatibility composite characters23
 - see* compatibility decomposable characters
 - compatibility composition
 - defined in* UAX #15, Unicode Normalization Forms
 - compatibility decomposable characters23
 - definition71
 - compatibility decomposition21
 - definition71
 - compatibility decomposition mappings71
 - Compatibility Encoding Scheme for UTF-16
 - see* UTR #26, Compatibility Encoding Scheme for UTF-16: 8-Bit (CESU-8)
 - compatibility equivalence
 - definition72
 - compatibility mappings
 - see* compatibility decomposition mappings
 - compatibility precomposed characters
 - see* compatibility decomposable characters
 - compatibility variant23
 - compatibility-equivalent character sequences
 - conformance60
 - composite character sequences
 - see* combining character sequences
 - composite characters
 - see* decomposable characters
 - compatibility *see* compatibility decomposable characters
 - Composition Exclusion (normative property)67
 - CompositionExclusions.txt1357
 - compression115
 - see also* UTS #6, A Standard Compression Scheme for Unicode (SCSU)
 - conferences1346
 - conformance55–91
 - definitions63–66
 - examples48
 - ISO/IEC 10646 implementations1353
 - requirements58–61
 - confusables141
 - conjunct consonants
 - Indic121, 222
 - Myanmar272
 - selection of clusters121
 - contextual shaping
 - apostrophe159
 - Arabic195
 - Mongolian326
 - not used for Hebrew final forms193
 - quotation marks156
 - Syriac210
 - contour tones185
 - control codes25, 48, 385
 - graphics for365
 - names102
 - properties386
 - semantics26, 386
 - specified in Unicode386
 - control sequences385
 - conversion of characters107–109
 - convertibility
 - as Unicode design principle22
 - Coptic177
 - corporate use subarea399
 - corrigenda56
 - CR (U+000D carriage return)116, 386
 - CRLF (carriage return and line feed)116
 - Croatian168
 - digraphs169
 - culturally expected sorting14, 133
 - currency symbols351–352
 - encoded in script blocks352
 - cursive joining389–391
 - Arabic199–204
 - control characters for103, 195–196, 388
 - Mongolian326
 - Syriac210–212
 - cursive scripts191
 - Cypriot346
 - see also* Linear B
 - Cyrillic179
 - Czech168
- ## D
- danda, in Devanagari block231
 - Danish167
 - dashes155
 - Database, Unicode Character
 - see* Unicode Character Database (UCD)
 - dead consonants, Indic221
 - dead keys123
 - decomposable characters21
 - definition71
 - normalization of115
 - decomposition21, 71–72
 - canonical *see* canonical decomposition
 - compatibility *see* compatibility decomposition

- definition 71
 - in normalization 115
 - mappings noted in code charts 416
 - default grapheme clusters 121
 - see also* UAX #29, Text Boundaries
 - default ignorable code points 111, 142
 - Default Ignorable Code Points (property) 142
 - default property values 110
 - definition 69
 - defective combining character sequences 126
 - definition 71
 - dependent vowel signs
 - Indic 220
 - Khmer 278
 - Philippine scripts 286
 - deprecated characters 55
 - alternate formatting 103, 394–395
 - definition 65
 - Derived Age (property) 111
 - derived properties
 - definition 68
 - DerivedAge.txt 1357, 1358, 1359
 - DerivedCoreProperties.txt 89, 97, 142
 - DerivedNormalizationProps.txt 140
 - Deseret 332–333
 - design goals of Unicode 3
 - design principles of Unicode 14–22
 - designated code points 25
 - Devanagari 219–231
 - Dhivehi 213
 - diacritics 43, 186–188
 - alternative glyphs 187
 - Czech 168
 - double 82, 103, 186
 - Greek 174–175, 177
 - Latin 166, 168–169
 - Latvian 168
 - mathematical 357
 - on i and j 169
 - rendering 125–129
 - rendering in isolation 46, 187
 - Romanian 168
 - Slovak 168
 - spacing clones of 46, 167
 - Turkish 168
 - see also* combining characters
 - digit form names 197
 - digits 100, 114
 - Arabic-Indic 196–197
 - national shapes 395
 - dingbats 371–372
 - directionality 19, 42
 - East Asian scripts 292
 - Middle Eastern scripts 191
 - Mongolian 325
 - normative property 98
 - Ogham 338
 - Old Italic 339
 - Philippine scripts 287
 - Runic 341
 - discussion list for Unicode xxxvii
 - dotless i 137, 169
 - dotted circle
 - in code charts 70, 187
 - in fallback rendering 125
 - to indicate diacritic 43
 - to indicate vowel sign placement 44
 - double diacritics 82, 103, 186
 - Dutch 167
 - dynamic composition
 - as Unicode design principle 20
 - Dzongkha 251
- ## E
- East Asian scripts 291–317
 - writing direction 42
 - see also* CJK ideographs
 - Eastern Arabic-Indic digits 197
 - EBCDIC
 - newline function 117
 - see* UTR #16, UTF-EBCDIC
 - editing, text boundaries for 121–122
 - efficiency
 - as Unicode design principle 15
 - e-mail discussion list for Unicode xxxvii
 - Enclosed Alphanumerics 373
 - enclosing marks 82, 188
 - not allowed in identifiers 130
 - encoded characters 5, 24
 - allocation 35–42, 1356
 - definition 64
 - encoding form conversion
 - definition 78
 - surrogates in 112
 - encoding forms 26–32
 - ISO/IEC 10646 definitions 1350
 - encoding forms, Unicode
 - see* Unicode encoding forms
 - encoding model for Unicode characters 26, 33
 - see also* UTR #17, Character Encoding Model
 - encoding schemes 32–35
 - see* Unicode encoding schemes
 - encoding schemes, Unicode
 - see* Unicode encoding schemes
 - end user subarea 399
 - endian ordering
 - see* byte order mark (BOM) (U+FEFF)
 - English 166
 - equivalent sequences 114
 - as Unicode design principle 20
 - case-insensitivity 134, 138
 - combining characters in matching 123
 - conformance 60
 - Hangul syllables 315
 - in sorting and searching 132
 - language-specific 72
 - security implications 140
 - see also* canonical equivalence
 - see also* compatibility equivalence
 - see also* encoding forms, encoding schemes
 - errata 56, 1346, 1360
 - escape sequences 385
 - not used in Unicode 1, 4
 - Esperanto 168
 - Estonian 168
 - Ethiopic 322–324
 - Etruscan 339
 - euro sign (U+20AC) 351
 - European alphabetic scripts 165–188
 - eyelash-RA 226

F

fallback rendering of nonspacing marks125
 FAQ (Frequently Asked Questions)1346
 Faroese167
 Farsi195, 196
 featural syllabaries149
 FF (U+000C form feed)116, 386
 file separator (FS) (U+001C information separator
 four)386
 Finnish167
 Finno-Ugric Transcription (FUT)
see Uralic Phonetic Alphabet (UPA)
 fixed-width Unicode encoding form (UTF-32) ..29, 76
 flat tables108
 Flemish167
 fonts
 and Unicode characters16
 for mathematical alphabets 356–357
 style variation for symbols349
 form feed (U+000C) (FF)116, 386
 format control characters 25, 48, 154, 383–409
 deprecated 394–395
 ignored in identifiers131
 reserved ranges111
 stateful143
 fraction characters358
 fraction slash (U+2044)159, 358
 French168
 Frisian168
 FTP site, Unicode Consortiumxxvii
 fullwidth forms in East Asian encodings313
 see also UAX #11, East Asian Width
 futhark341

G

Garshuni206
 Ge'ez322
 General Category (normative property)98
 list of values99
 general punctuation152–162
 General Scripts Area39
 geometrical symbols368–369
 Georgian182–183
 German167
 geta mark (U+3013)161
 Glagolitic179
 glyph selection tables108
 glyphs5, 15
 characters, relationship to15
 diacritics alternative187
 Greek alternative175–176
 Latin alternative167, 168
 mathematical alternative360
 representative in code charts414
 standardized variants397
 symbols alternative349
 golden numbers342
 Gothic343
 grapheme clusters12, 46
 see also UAX #29, Text Boundaries
 default121
 grapheme joiner, combining (U+034F)392
 graphic characters25

Greek174–178
 alternative glyphs175–176
 letters as symbols175–176, 361
 see also Cypriot, Linear B
 Greenlandic168
 group separator (GS) (U+001D information separa-
 tor three)386
 guillemets157
 Gujarati236
 Gurmukhi234–235

H

Hakka311
 halant217
 see also virama
 half marks, combining103, 188
 half-consonants, Indic223
 halfwidth forms in East Asian encodings313
 see also UAX #11, East Asian Width
 Han ideographs *see* CJK ideographs
 Han unification299–304
 and language tags120
 history1341–1342
 language usage297
 source separation rule295, 300
 source standards293–295, 304
 Hangul syllables291, 314–316
 as grapheme clusters47
 boundaries86
 canonical decomposition88
 collation314
 composition87
 conjoining jamo85–89
 equivalent sequences315
 Hangul Compatibility Jamo314–315
 Hangul Jamo314
 Hangul Syllables block315–316
 Johab set315
 names88
 normalization315
 precomposed86
 standard86
 Hangzhou numerals358
 Hanja *see* CJK ideographs
 Hanunóo286
 Hanzi *see* CJK ideographs
 harakat, Arabic pronunciation marks195
 hasant232
 hash tables108
 Hebrew192–194
 higher-level protocols
 definition66
 high-surrogate code points396
 definition72
 high-surrogate code units
 definition72
 Hindi219
 Hiragana312
 historic scripts337–346
 horizontal tab (HT) (U+0009 character tabulation) ...386
 HTML newline function117
 Hungarian168
 hyphenation388
 as a text process12

hyphens 155, 388

I

I Ching symbols 372
 IANA charset names 33
 Icelandic 167
 identifiers 130–132
 see also UAX #15 (Annex 7, Programming Language Identifiers, normalization)
 Ideographic (informative property) 102
 Ideographic Rapporteur Group (IRG) 294, 1342
 ideographs *see* CJK ideographs
 ill-formed
 definition 74
 implementation guidelines 107–143
 in a Unicode encoding form
 definition 75
 in-band mechanisms 408
 Indic scripts 217–250
 principles, in terms of Devanagari 220–225
 relation to ISCII standard 219
 Indonesian 166
 industry character sets
 covered in Unicode 3
 information separators (U+001C..U+001F) 386
 informative properties
 definition 67
 inside-out rule 125
 interchange restrictions 26
 International Phonetic Alphabet (IPA) . 148, 170–171
 spacing modifier letters 184
 see also phonetic alphabets
 internationalization 15
 Internationalization & Unicode Conferences (IUC) .. 1346
 Internet protocols
 UTF-8 as preferred encoding 30
 Inuktitut 331
 invisible operators 393
 iota subscript 175
 IPA *see* International Phonetic Alphabet
 IRG (Ideographic Rapporteur Group) 294, 1342
 Irish 167, 338
 ISCII standard and Unicode 219
 ISO/IEC 10646 5, 1347–1353
 codespace 1350
 conformance of Unicode implementations . 1353
 encoding forms 1350
 synchrony with Unicode Standard 6, 1352
 timeline compared to Unicode versions 1348
 UCS transformation formats (UTF) 1351
 Italian 167
 ITC Zapf Dingbats 371
 IUC (Internationalization & Unicode Conference) ... 1346

J

Jamo Short Name (normative property) 67
 Jamo.txt 88
 jamos *see* Hangul syllables
 Japanese 291
 Jawi 204
 Johab 315
 joiners 196
 combining grapheme joiner (U+034F) 392
 word joiner (U+2060) 387

 zero width joiner (U+200D) 195–196, 390
 justification 126

K

Kana (Hiragana and Katakana) 312–313
 Kanbun 305
 KangXi radicals 306, 1189
 Kanji *see* CJK ideographs
 Kannada 245–247
 Katakana 312–313
 KC (normalization form)
 see Normalization Form KC
 KD (normalization form)
 see Normalization Form KD
 keytop labels 365
 Khmer 274–283
 characters not recommended 280
 syllable components, order of 281
 killer
 Myanmar 272
 see also virama
 Korean Hangul *see* Hangul
 Kurdish 195

L

Ladino 192
 language tags 119, 405–408
 and Han unification 120
 use strongly discouraged 408
 Lao 269–270
 last-resort glyphs 142
 Latin 166–173
 alternative glyphs 167, 168
 Basic Latin 166
 encoding blocks 36
 IPA Extensions 170–171
 Latin Extended Additional 172–173
 Latin Extended-A 167–169
 Latin Extended-B 169–170
 Latin Ligatures 173
 Latin-1 Supplement 167
 Latvian 168
 layout control characters 48, 387–392
 ignored in identifiers 131
 leading surrogates *see* high-surrogate code units
 legibility criterion for plain text 18
 letter spacing 388
 letterlike symbols 353–357
 LF (U+000A line feed) 116, 386
 liaison members, Unicode Consortium 6
 ligatures 389–391
 Arabic 201–202
 combining characters on 46
 control characters for 103
 for nonspacing marks 128
 Latin 173
 selection 122
 Syriac 212
 Limbu 260–262
 line breaking 116–119, 387–389
 control characters 104
 in South Asian scripts 268, 273, 283
 recommendations 118
 see also UAX #14, Line Breaking Properties

- line feed (U+000A) (LF)116, 386
line separator (U+2028) (LS)116, 388
line tabulation (U+000B) (VT)386
Linear B345
see also Cypriot
linear boundaries122
Lithuanian168
little-endian32
definition61
logical order
as Unicode design principle18
logosyllabaries150
lowercase96, 136, 165
low-surrogate code points396
definition72
low-surrogate code units
definition72
LS (U+2028 line separator)116, 388
- M**
- MacOS newline function117
mail discussion list for Unicodexxvii
major version56
Malay166
Malayalam248–249
Maltese168
Manchu325
Mandarin298
mapping tables *see* tables of character data
Marathi219, 226, 230
markup languages
and Unicode conformance408
line breaking116
see also UTR #20, Unicode in XML and Other
Markup Languages
Mathematical (informative property)361
mathematical expression formatting characters ..103
see also UTR #25, Unicode Support for Mathematics
mathematical symbols360–364
alphabets354–357
alphanumeric354–357
fonts356–357
formatting characters393
fragments for typesetting366
invisible operators393
operators360–361
standardized variants363
MathML362
matras220
Middle Eastern scripts191–214
Min298
Minnan (Hokkien/Fujian, incl. Taiwanese)311
minor version56
minus sign361
commercial (U+2052)160
mirrored property
see Bidi Mirrored (normative property)
Miscellaneous Symbols370
missing glyphs142
Modifier Letters, Spacing184–185
Mongolian325–328
writing direction42, 325
- multibyte encodings
compared to UTF-830
multistage tables108
musical symbols376–380
Myanmar271–273
- N**
- NEL (U+0085 next line)116, 386
Nepali219
neutral directional characters98
new characters or scripts
how to propose7
newline function (NLF)117–119, 386
newline guidelines116–119
next line (U+0085) (NEL)116, 386
NFC (Normalization Form C)21
NFD (Normalization Form D)21
NFKC (Normalization Form KC)21
NFKD (Normalization Form KD)21
NLF (newline function)117–119, 386
no-break space (U+00A0)387
base for diacritic in isolation46, 187
noncharacter code points *see* noncharacters
noncharacters26, 47, 400
conformance59
definition65
deletion60
handling60
in code charts417
interchange restrictions26
semantics26
U+10FFFF (not a character code)400
U+FDD0..U+FDEF26
U+FFFE (not a character code)47, 400
U+FFFF (not a character code)26, 400
nondecomposable characters21
non-joiner, zero width (U+200C)195–196, 390
nonlinear boundaries122
non-overlap principle in Unicode encoding forms ..27
nonspacing marks
definition70
positioning128
rendering125–129
see also combining characters
see also diacritics
normalization21, 114–115
and case operations139
conformance62
of private use characters398
see also UAX #15, Unicode Normalization Forms
Normalization Form C (NFC)21
Normalization Form D (NFD)21
Normalization Form KC (NFKC)21
Normalization Form KD (NFKD)21
normative behaviors
definition63
normative properties
definition66
list67
may change66
Norwegian167
notational conventionsxxxiv–xxxvii
notational systems150
nukta227

- null (U+0000)
 as Unicode string terminator 386
 number forms 358–359
 CJK ideographs 114
 number handling 114
 numerals, old-style 153
 numeric separators 152
 numeric shape selectors (deprecated) 395
 Numeric Value (normative property) 100
 numero sign (U+2116) 353
- O**
- object replacement character (U+FFFC) 404
 obsolete characters 65
 octet xxxv
 Ogham 338
 Old Italic 339–340
 old-style numerals 153
 Oriya 237–238
 Oromo 322
 Osmanya 329
 Other_ID_Start (property) 131
 out-of-band mechanisms 408
 overlapping encodings 27
- P**
- Panjabi 234
 paragraph or section marks 160
 paragraph separator (U+2029) (PS) 116, 388
 Pashto 195
 Persian 195, 196
 Philippine scripts 286–287
 phonemes 150
 phonetic alphabets 148
 IPA Extensions 170–171
 Phonetic Extensions 171–172
 Spacing Modifier Letters 184–185
 Uralic Phonetic Alphabet (UPA) 160, 171
 see also International Phonetic Alphabet (IPA)
 Pinyin 169
 pivot code, Unicode as 107
 plain text
 as Unicode design principle 18
 legibility criterion 18
 planes of Unicode codespace 35
 Plane 0 (BMP) 35
 Plane 1 (SMP) 35, 41
 Plane 2 (SIP) 35, 37
 Plane 14 (SSP) 36
 Planes 15–16 (Private Use) 37, 399
 points, Hebrew pronunciation marks 192
 policies of the Unicode Consortium 1346
 Polish 168
 Portuguese 167
 precomposed characters
 see decomposable characters
 compatibility *see* compatibility decomposable
 characters
 prefixed format control characters 103
 Private Use Area (PUA) 40, 398
 private use characters
 semantics 26
 Private Use planes 37, 399
 private-use code points 26, 110
 conformance 59
 definition 69
 high-surrogates 396
 processing code, choice of Unicode encoding form 31
 properties 17, 66–70, 95–104
 aliases (definition) 68
 and Unicode algorithms 66
 data tables 108
 default values (definition) 69
 derived *see* derived properties
 in Unicode Character Database (UCD) 36
 informative *see* informative properties
 normative references to 57, 62
 normative *see* normative properties
 of control codes 386
 provisional *see* provisional properties
 simple *see* simple properties
 see also individual properties, e.g. combining
 classes
 property values
 aliases (definition) 69
 default 110, 398
 default (definition) 69
 normative references to 62
 PropertyAliases.txt xxxvi, 69, 1358
 PropertyValueAliases.txt xxxvi, 69, 1358
 Provençal 168
 provisional properties
 definition 67
 PS (U+2029 paragraph separator) 116, 388
pulli 239
 PUA (Private Use Area) 40, 398
 punctuation 152–162
 ASCII 152
 blocks containing 147
 CJK 160
 doubled 159
 in bidirectional text 152
 paired 153
 small form variants 162
 typographic forms 152
 vertical forms 161
 Punjabi 234
- Q**
- quotation marks 156–158
 East Asian 158
 European 157
- R**
- radicals, KangXi and other CJK 306
 radical-stroke index 1189
 record separator (RS) (U+001E information separator two) 386
 recycling symbols 370
 referencing 62
 properties 57
 Unicode algorithms 57
 Unicode Standard 57
 regular expressions 119
 and line breaking 116
 see also UTR #18, Unicode Regular Expression
 Guidelines

- rendering of text5, 12, 16
 - unsupported characters110
 - repertoire of abstract characters24
 - replacement character (U+FFFD)35, 48, 61, 404
 - reserved code points25, 110
 - definition65
 - in code charts417
 - preservation in interchange26
 - see also* unassigned code points
 - Rhaeto-Romanic168
 - rich text18
 - right single quotation mark (U+2019)
 - preferred for apostrophe153
 - right-to-left text42
 - East Asian scripts292
 - Middle Eastern scripts191
 - roadmap for script additions40
 - Roman numerals114, 358
 - Romanian168
 - Romany168
 - Runic341–342
 - Russian179
- S**
- Sami168
 - sample code, on CD-ROMxxxiv
 - Sanskrit219
 - scalar values, Unicode
 - see* Unicode scalar values
 - scripts
 - added in Version 4.02
 - adding to Unicode Standard7
 - in Unicode Standard2
 - roadmap for future additions40
 - types of151
 - see also* UAX #24, Script Names
 - SCSU
 - see* UTS #6, A Standard Compression Scheme for Unicode
 - searching132–134
 - as a text process12
 - case-insensitive134, 138
 - section or paragraph marks160
 - security issues140
 - self-synchronization of encoding forms28
 - semantics *see* character semantics
 - sequences
 - notationxxxv
 - Serbian
 - corresponding digraphs in Croatian169
 - Shan284
 - Shavian334
 - Show Hidden60, 125, 142, 397
 - SHY (U+00AD soft hyphen)388
 - signature for Unicode data48, 401–402
 - simple properties
 - definition68
 - simplified Chinese297
 - Sindhi195, 231
 - Sinhala250
 - SIP (Supplementary Ideographic Plane)35, 37
 - slash, fraction (U+2044)159
 - Slovak168
 - Slovenian168
 - small letters96, 136, 165
 - SMP (Supplementary Multilingual Plane)35, 41
 - soft hyphen (U+00AD) (SHY)388
 - Somali329
 - Sorbian168
 - sorting14, 132
 - and combining characters85
 - as a text process12
 - case-insensitive133
 - culturally expected14, 133
 - language-insensitive133
 - see also* Unicode collation algorithm (UCA)
 - source separation rule295, 300
 - South Asian scripts217–262
 - Southeast Asian scripts265–287
 - space (U+0020)
 - base for diacritic in isolation46, 187
 - space characters154, 387–389
 - graphics for365
 - spacing clones of diacritics46, 167
 - spacing marks
 - definition70
 - Spacing Modifier Letters184–185
 - Spanish167
 - special characters47, 383–409
 - SpecialCasing.txt89, 97
 - Specials401–404
 - spell-checking
 - as a text process12
 - spellings, alternative
 - see* equivalent sequences
 - spoofing141
 - SSP (Supplementary Special-purpose Plane)36
 - stacked boundaries121
 - stacking sequences44
 - nondefault45
 - Standard Compression Scheme for Unicode (SCSU)
 - see* UTS #6, A Standard Compression Scheme for Unicode
 - standard Korean syllables86
 - standardized variants327, 397
 - mathematical symbols363
 - StandardizedVariants.txt327, 363, 397
 - standards coverage3
 - stateful encoding
 - not used in Unicode4
 - paired format controls143
 - string comparison14
 - string literals, Unicode
 - code point notation `\u1234`xxxvi
 - strings, Unicode34, 74
 - null termination386
 - strong directional characters98
 - styled text18
 - sublinear searching134
 - subsets, supported50
 - conformance59
 - ISO/IEC 10646 specification for1352
 - substitution character48
 - superscripts and subscripts359
 - supplementary characters
 - in UTF-16 strings34
 - tables for108
 - Supplementary Ideographic Plane (SIP)35, 37
 - Supplementary Multilingual Plane (SMP)35, 41

- supplementary planes
 - allocation of code points 1356
 - representation in UTF-16 29
 - representation in UTF-8 30
 - Supplementary Private Use Areas 37, 399
 - Supplementary Special-purpose Plane (SSP) 36
 - supported subsets 50
 - conformance 59
 - surrogate code points *see* surrogates
 - surrogate pairs 29
 - definition 73
 - processing 31, 111–113
 - surrogates 26, 72–73, 396
 - interchange restrictions 26
 - isolated surrogates ill-formed 76
 - isolated surrogates uninterpreted 73
 - isolated surrogates, handling 34
 - support levels 112
 - Surrogates Area 40, 396
 - Suzhou-style numerals 358
 - Swahili 166
 - Swedish 167
 - syllabaries 148
 - alphabetic property 102
 - featural 149
 - symbols 349–380
 - appearance variation 349
 - arrows 363
 - currency 351–352
 - dingbats 371–372
 - Enclosed Alphanumerics 373
 - fragments for mathematical typesetting 366
 - geometrical 368–369
 - Khmer lunar calendar 283
 - letterlike 353–357
 - mathematical 360–364
 - mathematical alphanumeric 354–357
 - miscellaneous 370
 - musical 376–380
 - number forms 358–359
 - recycling 370
 - technical 365–367
 - Symbols Area 40
 - symmetric swapping format characters (deprecated) 394
 - Syriac 206–212
- T**
- tab (U+0009 character tabulation) 386
 - tables of character data 108–109
 - optimization 108
 - supplementary characters 108
 - tag characters 405–409
 - use strongly discouraged 405
 - Tagalog 286
 - Tagbanwa 286
 - tags, language 119, 405–408
 - use strongly discouraged 408
 - Tai Le 284–285
 - Tamil 239–243
 - TCHAR in Win32 API 109
 - Technical Notes (UTN) 1345
 - Technical Reports (UTR) xxxiii, 50
 - abstracts 1344
 - Technical Standards (UTS) xxxiii, 50
 - abstracts 1344
 - technical symbols 365–367
 - Telugu 244
 - terminal emulation 350
 - text boundaries 12, 47, 102, 121–122, 130
 - see also* UAX #14, Line Breaking Properties
 - see also* UAX #29, Text Boundaries
 - text elements 5, 12, 121
 - boundaries 130
 - for sorting 133
 - variable-width nature 31
 - text processes 4, 12–14
 - text rendering 5, 12, 16
 - text selection, boundaries for 121–122
 - Thaana 213–214
 - Thai 266–268
 - Tibetan 251–259
 - Tigre 322
 - tilde (U+007E) 154
 - titlecase 96, 136
 - Todo 325
 - tone letters 185
 - tone marks
 - Bopomofo spacing 310, 311
 - Tai Le 284
 - Thai 266
 - Vietnamese 172
 - traditional Chinese 297
 - trailing surrogates *see* low-surrogate code units
 - transcoding 107–109
 - tables 108
 - triangulation in transcoding 107
 - truncation
 - combining character sequences 123–124
 - surrogates and 112, 113
 - Turkish 168
 - case mapping of I 137, 169
 - two-stage tables 108
- U**
- U+ notation xxxvi
 - U+10FFFF (not a character code) 400
 - U+FFFE (not a character code) 400
 - U+FFFF (not a character code) 400
 - UAX (Unicode Standard Annex) xxxiii, 51
 - abstracts 1343
 - as component of Unicode Standard 58
 - conformance 62
 - list of 62, 1343
 - UCA *see* Unicode collation algorithm
 - UCD *see* Unicode Character Database
 - UCS (Universal Character Set) *see* ISO/IEC 10646
 - UCS-2 1350
 - UCS-4 1350
 - Ugaritic 344
 - Ukrainian 179
 - unassigned code points 25, 58, 110
 - defined as reserved code points 65
 - handling 56
 - properties of 110
 - semantics 59
 - see also* reserved code points
 - undesignated code points 25

- Unicode 1.0 Name (informative property)101
- Unicode algorithms
 - and properties66
 - conformance62
 - definition66
 - normative references to57, 62
- Unicode bidirectional algorithm19, 42
 - see also* UAX #9, Bidirectional Algorithm
- Unicode Character Database (UCD)xxxiii, 96
 - as component of Unicode Standard58
 - changes56
 - properties in36
- Unicode character encoding model26, 33
 - see also* UTR #17, Character Encoding Model
- Unicode character literals
 - code point notation U+ xxxvi
- Unicode codespace
 - allocation numbers1356
 - definition64
 - planes35
 - same as ISO/IEC 106461350
 - size1, 24
- Unicode collation algorithm (UCA)14
 - see also* UTS #10, Unicode Collation Algorithm
- Unicode conferences1346
- Unicode Consortium6
 - addressesxxxviii
 - Consortium membership in standards bodies . .6
 - e-mail discussion listxxxvii
 - FTP sitexxxvii
 - liaison members6
 - membership6
 - policies1346
 - Web sitexxxiii, xxxvii, 1345
- Unicode data signature48, 401–402
- Unicode encoding forms73–78
 - advantages of each31
 - conformance28, 60
 - definition74
 - fixed-width (UTF-32)29, 76
 - signatures402
 - variable-width (UTF-16)29, 76
 - variable-width (UTF-8)30, 77
 - see also* encoding forms
- Unicode encoding schemes
 - conformance78–81
 - definition78
 - endian ordering32
 - see also* encoding schemes
- Unicode escape sequence notation \u1234 xxxvi
- Unicode scalar values
 - definition73
- Unicode Standard
 - adding new characters or scripts7
 - allocation of encoded characters35–42
 - application areas3
 - architecture11–14
 - areas36
 - benefits1
 - blocks36, 147
 - code charts413–418
 - components58
 - conformance55–91
 - conformance of ISO/IEC 10646 implementations . .1353
 - corrections56
 - definitions for conformance63–66
 - design goals3
 - design principles14–22
 - errata56, 1346, 1360
 - normative references to57, 62
 - number of characters2, 1356
 - number of code points1, 24
 - online code charts1346
 - script coverage2
 - security issues140
 - synchrony with ISO/IEC 106461352
 - updates1346
 - user community3
 - versions *see* versions of the Unicode Standard
 - see also* Version 4.0
- Unicode Standard Annexes (UAX) xxxiii, 51
 - abstracts1343
 - as components of Unicode Standard58
 - conformance62
 - list of62, 1343
- Unicode string literals
 - code point notation \u1234 xxxvi
- Unicode strings34
 - definition74
- Unicode Technical Committee (UTC)6
- Unicode Technical Notes (UTN)1345
- Unicode Technical Reports (UTR)xxxiii, 50
 - abstracts1344
- Unicode Technical Standards (UTS)xxxiii, 50
 - abstracts1344
- UnicodeData.txt56, 89, 97
- unification
 - as Unicode design principle19
 - see also* Han unification
- Unified CJK Ideograph (property)307
- Unified Repertoire and Ordering (URO) . .300, 1342
 - see also* Han unification
- Unihan.txt68, 303, 417
- unit separator (US) (U+001F information separator one)386
- Universal Character Set (UCS) *see* ISO/IEC 10646
- universality
 - as Unicode design principle14
- Unix
 - and UTFs31
 - newline function117
 - UTF-32 in29
 - UTF-8 in15
- unsupported characters110–111
- update version56, 95
- uppercase96, 136, 165
- Uralic Phonetic Alphabet (UPA)160, 171
- Urdu195
- URO (Unified Repertoire and Ordering) . .300, 1342
 - see also* Han unification
- UTF, Unicode Transformation Formats27, 74
 - advantages of each31
 - as encoding form or scheme81
 - binary comparison and sort order differences . .133, 135
 - in APIs109
 - in ISO/IEC 106461351
- UTF-1629, 76
 - binary comparison and sort order caution . . .30
 - bit distribution (table)77
 - BOM in79, 401

- encoding form (definition) 76
 - encoding scheme (definition) 79
 - encoding schemes 32
 - in ISO/IEC 10646 1351
 - in UTF-8 order 136
 - surrogates and string handling 34, 111
 - UTF-16BE (Big-endian) 402
 - encoding scheme 33
 - encoding scheme (definition) 79
 - UTF-16LE (Little-endian) 402
 - encoding scheme 33
 - encoding scheme (definition) 79
 - UTF-32 29, 76
 - BOM in 80
 - encoding form (definition) 76
 - encoding scheme (definition) 80
 - encoding schemes 32
 - in Unix 29
 - UTF-32BE (Big-endian)
 - encoding scheme 33
 - encoding scheme (definition) 80
 - UTF-32LE (Little-endian)
 - encoding scheme 33
 - encoding scheme (definition) 80
 - UTF-8 30, 77
 - ASCII transparency 30
 - binary comparison and sort order 32
 - bit distribution (table) 77
 - BOM in 79, 81, 401
 - byte ranges 77
 - compared to multibyte encodings 30
 - encoding form (definition) 77
 - encoding scheme 32
 - encoding scheme (definition) 79
 - in ISO/IEC 10646 1351
 - in Unix 15
 - in UTF-16 order 135
 - non-shortest form is invalid 77, 140
 - preferred encoding for Internet protocols 30
 - security and 140
 - signature 79, 81, 401
 - UTF-EBCDIC
 - see* UTR #16, UTF-EBCDIC
 - UTN (Unicode Technical Note) 1345
 - UTR (Unicode Technical Report) xxxiii, 50
 - abstracts 1344
 - UTS (Unicode Technical Standard) xxxiii, 50
 - abstracts 1344
- V**
- valid (synonym for well-formed) 75
 - variable-width Unicode encoding form (UTF-16) . . 29, 76
 - variable-width Unicode encoding form (UTF-8) . . 30, 77
 - variation selectors 104, 397
 - ideographic variation mark (U+303E) 309
 - Mongolian free variation selectors 327
 - variation sequences 397
 - Version 4.0 58
 - additions xxxi, 2
 - changes since Version 3.0 1357–1360
 - correlation with ISO/IEC 10646 1350
 - number of characters 2, 1356
- versions of the Unicode Standard xxxiii, 55, 1346, 1355–1357
- backward compatibility 55
 - compared to ISO/IEC 10646 editions 1355
 - content 56
 - interaction in implementations 111
 - numbering 56
 - property changes 56
 - stability 56
 - updates 1346
- vertical tab (VT) (U+000B line tabulation) . 116, 386
- vertical text 42, 152, 161
- East Asian scripts 292
 - Mongolian 325
- Vietnamese 172
- ideographs 291
- virama 149, 217
- definition 221
 - Khmer 277
 - Myanmar 272
 - Philippine scripts 286
 - virama-like characters 104
- visual order used for Thai and Lao 19
- vowel marks, Middle Eastern scripts 191
- vowel signs
- Indic 44, 220
 - Khmer 278
 - Philippine scripts 286
- VT (U+000B line tabulation) 116, 386
- W**
- wchar_t
- and Unicode encoding forms 31
 - in C language 109
- weak directional characters 98
- Web site, Unicode Consortium xxxiii, xxxvii, 1345
- Weierstrass elliptic function symbol 354
- well-formed
- definition 75
- Welsh 168
- Where is my Character? 1346
- wide characters
- datatype in C 109
- wiggly fence (U+29DB) 363
- Windows newline function 117
- word breaks 123, 387–389
- in South Asian scripts 268, 273, 283
- word joiner (U+2060) 387
- writing direction *see* directionality
- writing systems 148–151
- Wu (Shanghainese) 298
- X**
- XML
- see* UTR #20, Unicode in XML and Other Markup Languages
- Y**
- yen currency sign 351
 - Yi 317
 - Yiddish 192
 - ypogegrammeni 175
 - yuan currency sign 351

Z

- Zapf Dingbats371
- zero extension relation among encodings1350
- zero width joiner (U+200D) 195–196, 390
- zero width no-break space (U+FEFF)47, 61, 387
 - initial81, 402
- zero width non-joiner (U+200C) 195–196, 390
- zero width space (U+200B)387
 - for word breaks in South Asian scripts . 268,273, 283
- zero width space characters388
- ZWJ *see* zero width joiner (U+200D)
- ZWNBSP *see* zero width no-break space (U+FEFF)
- ZWNJ *see* zero width non-joiner (U+200C)
- ZWSP *see* zero width space (U+200B)