

Electronic Edition

This file is part of the electronic edition of *The Unicode Standard, Version 5.0*, provided for online access, content searching, and accessibility. It may not be printed. Bookmarks linking to specific chapters or sections of the whole Unicode Standard are available at

<http://www.unicode.org/versions/Unicode5.0.0/bookmarks.html>

Purchasing the Book

For convenient access to the full text of the standard as a useful reference book, we recommend purchasing the printed version. The book is available from the Unicode Consortium, the publisher, and booksellers. Purchase of the standard in book format contributes to the ongoing work of the Unicode Consortium. Details about the book publication and ordering information may be found at

<http://www.unicode.org/book/aboutbook.html>

Joining Unicode

You or your organization may benefit by joining the Unicode Consortium: for more information, see [Joining the Unicode Consortium](http://www.unicode.org/consortium/join.html) at

<http://www.unicode.org/consortium/join.html>

This PDF file is an excerpt from *The Unicode Standard, Version 5.0*, issued by the Unicode Consortium and published by Addison-Wesley. The material has been modified slightly for this electronic 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/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 the publisher was aware of a trademark claim, the designations have been printed with initial capital letters or in all capitals.

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 the 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, www.mehallo.com

The publisher offers excellent discounts on this book when ordered in quantity for bulk purchases or special sales, which may include electronic versions and/or custom covers and content particular to your business, training goals, marketing focus, and branding interests. For more information, please contact U.S. Corporate and Government Sales, (800) 382-3419, corpsales@pearsoned.com. For sales outside the United States please contact International Sales, international@pearsoned.com

Visit us on the Web: www.awprofessional.com

Library of Congress Cataloging-in-Publication Data

The Unicode Standard / the Unicode Consortium ; edited by Julie D. Allen ... [et al.]. — Version 5.0.
p. cm.

Includes bibliographical references and index.

ISBN 0-321-48091-0 (hardcover : alk. paper)

1. Unicode (Computer character set) I. Allen, Julie D.

II. Unicode Consortium.

QA268.U545 2007

005.7'22—dc22

2006023526

Copyright © 1991–2007 Unicode, Inc.

All rights reserved. Printed in the United States of America. This publication is protected by copyright, and permission must be obtained from the publisher prior to any prohibited reproduction, storage in a retrieval system, or transmission in any form or by any means, electronic, mechanical, photocopying, recording, or likewise. For information regarding permissions, write to Pearson Education, Inc., Rights and Contracts Department, 75 Arlington Street, Suite 300, Boston, MA 02116. Fax: (617) 848-7047

ISBN 0-321-48091-0

Text printed in the United States on recycled paper at Courier in Westford, Massachusetts.

First printing, October 2006

Tables

Table 2-1.	The 10 Unicode Design Principles	13
Table 2-2.	User-Perceived Characters with Multiple Code Points	16
Table 2-3.	Types of Code Points	27
Table 2-4.	The Seven Unicode Encoding Schemes	36
Table 2-5.	Interaction of Combining Characters	51
Table 2-6.	Nondefault Stacking	52
Table 3-1.	Named Unicode Algorithms	81
Table 3-2.	Normative Character Properties	86
Table 3-3.	Informative Character Properties	87
Table 3-4.	Examples of Unicode Encoding Forms	101
Table 3-5.	UTF-16 Bit Distribution	103
Table 3-6.	UTF-8 Bit Distribution	103
Table 3-7.	Well-Formed UTF-8 Byte Sequences	104
Table 3-8.	Summary of UTF-16BE, UTF-16LE, and UTF-16	106
Table 3-9.	Summary of UTF-32BE, UTF-32LE, and UTF-32	108
Table 3-10.	Sample Combining Classes	116
Table 3-11.	Canonical Ordering Results	117
Table 3-12.	Hangul Syllable No-Break Rules	119
Table 3-13.	Korean Syllable Break Examples	121
Table 3-14.	Context Specification for Casing	124
Table 3-15.	Case Detection Examples	126
Table 4-1.	Sources for Case Mapping Information	133
Table 4-2.	Class Zero Combining Marks—Reordrant	135
Table 4-3.	Thai and Lao Logical Order Exceptions	136
Table 4-4.	Class Zero Combining Marks—Split	136
Table 4-5.	Class Zero Combining Marks—Subjoined	137
Table 4-6.	Class Zero Combining Marks—Strikethrough	137
Table 4-7.	General Category	139
Table 4-8.	Primary Numeric Ideographs	140
Table 4-9.	Ideographs Used as Accounting Numbers	141
Table 4-10.	Unusual Properties	146
Table 5-1.	Hex Values for Acronyms	162
Table 5-2.	NLF Platform Correlations	163
Table 5-3.	Typing Order Differing from Canonical Order	175
Table 5-4.	Permuting Combining Class Weights	176
Table 5-5.	Casing and Normalization in Strings	189
Table 5-6.	Paired Stateful Controls	194
Table 6-1.	Typology of Scripts in the Unicode Standard	201

Table 6-2.	Unicode Space Characters	206
Table 6-3.	Unicode Dash Characters	207
Table 6-4.	East Asian Quotation Marks	210
Table 6-5.	Opening and Closing Forms	211
Table 6-6.	Names for the @	215
Table 7-1.	Nonspacing Marks Used with Greek	237
Table 7-2.	Greek Spacing and Nonspacing Pairs	241
Table 8-1.	Arabic Digit Names	272
Table 8-2.	Glyph Variation in Eastern Arabic-Indic Digits	273
Table 8-3.	Primary Arabic Joining Classes	275
Table 8-4.	Derived Arabic Joining Classes	276
Table 8-5.	Arabic Glyph Types	276
Table 8-6.	Arabic Ligature Notation	278
Table 8-7.	Dual-Joining Arabic Characters	279
Table 8-8.	Right-Joining Arabic Characters	281
Table 8-9.	Miscellaneous Syriac Diacritic Use	287
Table 8-10.	Additional Syriac Joining Classes	288
Table 8-11.	Dual-Joining Syriac Characters	289
Table 8-12.	Right-Joining Syriac Characters	290
Table 8-13.	Alaph-Joining Syriac Characters	290
Table 8-14.	Syriac Ligatures	291
Table 8-15.	Thaana Glyph Placement	292
Table 9-1.	Devanagari Vowel Letters	299
Table 9-2.	Sample Devanagari Half-Forms	309
Table 9-3.	Sample Devanagari Ligatures	310
Table 9-4.	Sample Devanagari Half-Ligature Forms	311
Table 9-5.	Bengali Vowel Letters	313
Table 9-6.	Bengali Consonant-Vowel Combinations	314
Table 9-7.	Gurmukhi Vowel Letters	318
Table 9-8.	Gurmukhi Conjuncts	319
Table 9-9.	Additional Pairin and Addha Forms in Gurmukhi	319
Table 9-10.	Use of Joiners in Gurmukhi	320
Table 9-11.	Gujarati Vowel Letters	321
Table 9-12.	Gujarati Conjuncts	321
Table 9-13.	Oriya Vowel Letters	322
Table 9-14.	Oriya Conjuncts	323
Table 9-15.	Oriya Vowel Placement	323
Table 9-16.	Tamil Vowel Reordering	326
Table 9-17.	Tamil Vowel Splitting and Reordering	326
Table 9-18.	Tamil Ligatures with u	327
Table 9-19.	Telugu Vowel Letters	331
Table 9-20.	Kannada Vowel Letters	332
Table 9-21.	Malayalam Vowel Letters	335
Table 9-22.	Malayalam Orthographic Reform	335
Table 9-23.	Malayalam Conjuncts	336

Table 10-1.	Sinhala Vowel Letters	342
Table 10-2.	Phags-pa Positional Forms of I, U, E, and O	357
Table 10-3.	Contextual Glyph Mirroring in Phags-pa	358
Table 10-4.	Phags-pa Standardized Variants	359
Table 10-5.	Positions of Limbu Combining Marks	362
Table 10-6.	Kharoshthi Vowel Signs	367
Table 10-7.	Kharoshthi Vowel Modifiers	368
Table 10-8.	Kharoshthi Consonant Modifiers	368
Table 10-9.	Examples of Kharoshthi Virama	369
Table 11-1.	Glyph Positions in Thai Syllables	374
Table 11-2.	Glyph Positions in Lao Syllables	377
Table 11-3.	Myanmar Syllabic Structure	381
Table 11-4.	Independent Khmer Vowel Characters	383
Table 11-5.	Two Registers of Khmer Consonants	385
Table 11-6.	Khmer Subscript Consonant Signs	385
Table 11-7.	Khmer Composite Dependent Vowel Signs with Nikahit	387
Table 11-8.	Khmer Subscript Independent Vowel Signs	388
Table 11-9.	Tai Le Tone Marks	393
Table 11-10.	Myanmar Digits	394
Table 11-11.	New Tai Lue Vowel Placement	395
Table 11-12.	New Tai Lue Registers and Tones	395
Table 11-13.	Hanunóo and Buhid Vowel Sign Combinations	397
Table 11-14.	Balinese Base Consonants and Conjunct Forms	399
Table 11-15.	Sasak Extensions for Balinese	401
Table 11-16.	Balinese Consonant Clusters with u and u:	402
Table 12-1.	Initial Sources for Unified Han	410
Table 12-2.	Blocks Containing Han Ideographs	411
Table 12-3.	Common Han Characters	413
Table 12-4.	Source Encoding for Sword Variants	418
Table 12-5.	Ideographs Not Unified	421
Table 12-6.	Ideographs Unified	421
Table 12-7.	Han Ideograph Arrangement	422
Table 12-8.	Sources Added for Extension B	423
Table 12-9.	Mandarin Tone Marks	431
Table 12-10.	Minnan and Hakka Tone Marks	432
Table 12-11.	Separating Jamo Characters	437
Table 12-12.	Line-Based Placement of Jungseong	438
Table 13-1.	Labialized Forms in Ethiopic -WAA	446
Table 13-2.	Labialized Forms in Ethiopic -WE	447
Table 13-3.	N’Ko Tone Diacritics on Vowels	460
Table 13-4.	Other N’Ko Diacritic Usage	460
Table 13-5.	N’Ko Letter Shaping	462
Table 13-6.	IPA Transcription of Deseret	467
Table 14-1.	Similar Characters in Linear B and Cypriot	480
Table 14-2.	Cuneiform Script Usage	485

Table 15-1.	Currency Symbols Encoded in Other Blocks	491
Table 15-2.	Mathematical Alphanumeric Symbols	496
Table 15-3.	Use of Mathematical Symbol Pieces	510
Table 15-4.	Japanese Era Names	518
Table 15-5.	Examples of Ornamentation	524
Table 15-6.	Representation of Ancient Greek Vocal and Instrumental Notation . .	526
Table 16-1.	Control Codes Specified in the Unicode Standard	533
Table 16-2.	Letter Spacing	535
Table 16-3.	Bidirectional Ordering Controls	543
Table 16-4.	Unicode Encoding Scheme Signatures	551
Table 16-5.	U+FEFF Signature in Other Charsets	552
Table A-1.	Extended BNF	1079
Table A-2.	Character Class Examples	1081
Table A-3.	Operators	1081
Table C-1.	Timeline	1092
Table C-2.	Zero Extending	1096
Table D-1.	Versions of Unicode and ISO/IEC 10646-1	1100
Table D-2.	Allocation of Code Points by Type	1101
Table D-3.	Clause and Definition Numbering	1102
Table F-1.	Constraints on Property Values	1123