

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](mailto:corpsales@pearsontechgroup.com). For sales outside of the U.S., please contact International Sales, +1 317 581 3793, [international@pearsontechgroup.com](mailto: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

# Figures

Figure 1-1.	Wide ASCII. . . . .	2
Figure 1-2.	Universal, Efficient, and Unambiguous. . . . .	4
Figure 2-1.	Text Elements and Characters . . . . .	13
Figure 2-2.	Characters Versus Glyphs . . . . .	16
Figure 2-3.	Unicode Character Code to Rendered Glyphs . . . . .	17
Figure 2-4.	User Characters as Multiple Code Points . . . . .	17
Figure 2-5.	Bidirectional Ordering . . . . .	19
Figure 2-6.	Equivalent Sequences . . . . .	21
Figure 2-7.	Types of Decomposables . . . . .	22
Figure 2-8.	Codespace and Encoded Characters . . . . .	24
Figure 2-9.	Overlap in Legacy Mixed-Width Encodings . . . . .	27
Figure 2-10.	Boundaries and Interpretation . . . . .	27
Figure 2-11.	Unicode Encoding Forms . . . . .	28
Figure 2-12.	Unicode Encoding Schemes . . . . .	34
Figure 2-13.	Unicode Allocation . . . . .	38
Figure 2-14.	Allocation on the BMP . . . . .	39
Figure 2-15.	Allocation on Plane 1. . . . .	41
Figure 2-16.	Indic Vowel Signs. . . . .	44
Figure 2-17.	Stacking Sequences . . . . .	44
Figure 2-18.	Interaction of Combining Characters . . . . .	45
Figure 2-19.	Nondefault Stacking . . . . .	46
Figure 2-20.	Ligated Multiple Base Characters. . . . .	46
Figure 3-1.	Enclosing Marks. . . . .	82
Figure 3-2.	Positioning of Double Diacritics . . . . .	82
Figure 4-1.	Positions of Common Combining Marks . . . . .	98
Figure 5-1.	Two-Stage Tables. . . . .	109
Figure 5-2.	Ideographic Numbers . . . . .	114
Figure 5-3.	Normalization . . . . .	115
Figure 5-4.	Consistent Character Boundaries. . . . .	121
Figure 5-5.	Dead Keys Versus Handwriting Sequence. . . . .	123
Figure 5-6.	Truncating Composed Character Sequences . . . . .	124
Figure 5-7.	Inside-Out Rule . . . . .	125
Figure 5-8.	Fallback Rendering . . . . .	125
Figure 5-9.	Bidirectional Placement . . . . .	126
Figure 5-10.	Justification. . . . .	127
Figure 5-11.	Positioning with Ligatures . . . . .	128
Figure 5-12.	Positioning with Contextual Forms . . . . .	129
Figure 5-13.	Positioning with Enhanced Kerning . . . . .	129
Figure 5-14.	Sublinear Searching . . . . .	134
Figure 5-15.	Case Mapping for Turkish I . . . . .	137
Figure 6-1.	Overriding Inherent Vowels. . . . .	149
Figure 6-2.	European Quotation Marks . . . . .	157
Figure 6-3.	Asian Quotation Marks. . . . .	158
Figure 7-1.	Alternative Glyphs . . . . .	168
Figure 7-2.	Diacritics on <i>i</i> and <i>j</i> . . . . .	169
Figure 7-3.	Vietnamese Letters and Tone Marks . . . . .	172

Figure 7-4.	Georgian Displayed with Ecclesiastical Font . . . . .	183
Figure 7-5.	Tone Letters . . . . .	185
Figure 7-6.	Double Diacritics . . . . .	186
Figure 7-7.	Positioning of Double Diacritics . . . . .	187
Figure 7-8.	Combining Half Marks . . . . .	188
Figure 8-1.	Directionality and Cursive Connection . . . . .	195
Figure 8-2.	Using a Joiner . . . . .	196
Figure 8-3.	Using a Non-joiner . . . . .	196
Figure 8-4.	Combinations of Joiners and Non-joiners . . . . .	196
Figure 8-5.	Syriac Abbreviation . . . . .	207
Figure 8-6.	Use of SAM . . . . .	208
Figure 9-1.	Dependent Versus Independent Vowels . . . . .	221
Figure 9-2.	Dead Consonants . . . . .	222
Figure 9-3.	Conjunct Formations . . . . .	222
Figure 9-4.	Preventing Conjunct Forms . . . . .	223
Figure 9-5.	Half-Consonants . . . . .	223
Figure 9-6.	Independent Half-Forms . . . . .	223
Figure 9-7.	Consonant Forms . . . . .	224
Figure 9-8.	Rendering Order . . . . .	228
Figure 9-9.	Marathi Allographs . . . . .	230
Figure 9-10.	Bengali Khanda Ta . . . . .	233
Figure 9-11.	Spacing Forms of Vowels . . . . .	242
Figure 9-12.	Tibetan Syllable Structure . . . . .	252
Figure 9-13.	Justifying Tseks . . . . .	259
Figure 10-1.	Common Ligatures . . . . .	280
Figure 10-2.	Common Multiple Forms . . . . .	280
Figure 10-3.	Examples of Syllabic Order . . . . .	282
Figure 10-4.	Ligation in <i>Muul</i> Style . . . . .	283
Figure 11-1.	Han Spelling . . . . .	297
Figure 11-2.	Context for Characters . . . . .	297
Figure 11-3.	Three-Dimensional Conceptual Model . . . . .	299
Figure 11-4.	Source Separation . . . . .	300
Figure 11-5.	Not Cognates, Not Unified . . . . .	301
Figure 11-6.	Component Structure . . . . .	301
Figure 11-7.	The Most Superior Node of a Component . . . . .	301
Figure 11-8.	Using the Ideographic Description Characters . . . . .	308
Figure 11-9.	Separating Jamo Characters . . . . .	315
Figure 12-1.	IPA Transcription of Deseret . . . . .	333
Figure 13-1.	Distribution of Old Italic . . . . .	340
Figure 14-1.	Easily Confused Shapes for Mathematical Glyphs . . . . .	356
Figure 15-1.	Letter Spacing . . . . .	388
Figure 15-2.	Sample Display Actions . . . . .	391
Figure 15-3.	Annotation Characters . . . . .	403
Figure 15-4.	Tag Characters . . . . .	406