

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

# Tables

|             |  |        |
|-------------|--|--------|
| Table 0-1.  | Extended BNF                                   | xxxv   |
| Table 0-2.  | Character Class Examples                       | xxxvii |
| Table 0-3.  | Operators                                      | xxxvii |
| Table 2-1.  | The 10 Unicode Design Principles               | 14     |
| Table 2-2.  | Types of Code Points                           | 25     |
| Table 2-3.  | The Seven Unicode Encoding Schemes             | 33     |
| Table 3-1.  | Normative Character Properties                 | 67     |
| Table 3-2.  | Informative Character Properties               | 68     |
| Table 3-3.  | Examples of Unicode Encoding Forms             | 76     |
| Table 3-4.  | UTF-16 Bit Distribution                        | 77     |
| Table 3-5.  | UTF-8 Bit Distribution                         | 77     |
| Table 3-6.  | Well-Formed UTF-8 Byte Sequences               | 78     |
| Table 3-7.  | Summary of UTF-16BE, UTF-16LE, and UTF-16      | 80     |
| Table 3-8.  | Summary of UTF-32BE, UTF-32LE, and UTF-32      | 81     |
| Table 3-9.  | Sample Combining Classes                       | 84     |
| Table 3-10. | Canonical Ordering Results                     | 85     |
| Table 3-11. | Hangul Syllable No-Break Rules                 | 86     |
| Table 3-12. | Syllable Break Examples                        | 87     |
| Table 3-13. | Context Specification for Casing               | 89     |
| Table 3-14. | Case Detection Examples                        | 91     |
| Table 4-1.  | Sources for Case Mapping Information           | 97     |
| Table 4-2.  | General Category                               | 99     |
| Table 4-3.  | Primary Numeric Ideographs                     | 100    |
| Table 4-4.  | Ideographs Used as Accounting Numbers          | 101    |
| Table 4-5.  | Unusual Properties                             | 103    |
| Table 5-1.  | Surrogate Support Levels                       | 112    |
| Table 5-2.  | Surrogate Level Examples                       | 113    |
| Table 5-3.  | Hex Values for Acronyms                        | 116    |
| Table 5-4.  | NLF Platform Correlations                      | 117    |
| Table 5-5.  | Typing Order Differing from Canonical Order    | 127    |
| Table 5-6.  | Permuting Combining Class Weights              | 128    |
| Table 5-7.  | Syntactic Classes for Identifiers              | 131    |
| Table 5-8.  | Casing and Normalization in Strings            | 139    |
| Table 5-9.  | Paired Stateful Controls                       | 143    |
| Table 6-1.  | Typology of Scripts in the Unicode Standard    | 151    |
| Table 6-2.  | Unicode Space Characters                       | 155    |
| Table 6-3.  | Unicode Dash Characters                        | 156    |
| Table 6-4.  | East Asian Quotation Marks                     | 158    |
| Table 6-5.  | Opening and Closing Forms                      | 158    |
| Table 7-1.  | Nonspacing Marks Used with Greek               | 174    |
| Table 7-2.  | Greek Spacing and Nonspacing Pairs             | 178    |
| Table 7-3.  | Font Styles and Georgian Forms                 | 182    |
| Table 8-1.  | Digit Names                                    | 197    |
| Table 8-2.  | Glyph Variation in Eastern Arabic-Indic Digits | 197    |
| Table 8-3.  | Primary Arabic Joining Classes                 | 199    |
| Table 8-4.  | Derived Arabic Joining Classes                 | 200    |

|              |  |     |
|--------------|--|-----|
| Table 8-5.   | Arabic Glyph Types . . . . .                           | 200 |
| Table 8-6.   | Ligature Notation . . . . .                            | 202 |
| Table 8-7.   | Dual-Joining Arabic Characters . . . . .               | 203 |
| Table 8-8.   | Right-Joining Arabic Characters . . . . .              | 204 |
| Table 8-9.   | Miscellaneous Syriac Diacritic Use . . . . .           | 209 |
| Table 8-10.  | Additional Syriac Joining Classes . . . . .            | 210 |
| Table 8-11.  | Dual-Joining Syriac Characters . . . . .               | 211 |
| Table 8-12.  | Right-Joining Syriac Characters . . . . .              | 211 |
| Table 8-13.  | Alaph-Joining Syriac Characters . . . . .              | 212 |
| Table 8-14.  | Syriac Ligatures . . . . .                             | 212 |
| Table 8-15.  | Thaana Glyph Placement . . . . .                       | 213 |
| Table 9-1.   | Sample Half-Forms . . . . .                            | 229 |
| Table 9-2.   | Sample Ligatures . . . . .                             | 229 |
| Table 9-3.   | Sample Half-Ligature Forms . . . . .                   | 230 |
| Table 9-4.   | Gurmukhi Conjuncts . . . . .                           | 235 |
| Table 9-5.   | Gujarati Conjuncts . . . . .                           | 236 |
| Table 9-6.   | Oriya Conjuncts . . . . .                              | 237 |
| Table 9-7.   | Oriya Vowel Placement . . . . .                        | 238 |
| Table 9-8.   | Vowel Reordering . . . . .                             | 240 |
| Table 9-9.   | Vowel Splitting and Reordering . . . . .               | 240 |
| Table 9-10.  | Ligating Vowel Signs . . . . .                         | 242 |
| Table 9-11.  | Malayalam Orthographic Reform . . . . .                | 248 |
| Table 9-12.  | Malayalam Conjuncts . . . . .                          | 249 |
| Table 9-13.  | Positions of Limbu Combining Marks . . . . .           | 262 |
| Table 10-1.  | Glyph Positions in Thai Syllables . . . . .            | 267 |
| Table 10-2.  | Glyph Positions in Lao Syllables . . . . .             | 269 |
| Table 10-3.  | Myanmar Syllabic Structure . . . . .                   | 273 |
| Table 10-4.  | Independent Vowel Characters . . . . .                 | 274 |
| Table 10-5.  | Two Registers of Khmer Consonants . . . . .            | 276 |
| Table 10-6.  | Khmer Subscript Consonant Signs . . . . .              | 277 |
| Table 10-7.  | Composite Dependent Vowel Signs with Nikahit . . . . . | 279 |
| Table 10-8.  | Subscript Independent Vowel Signs . . . . .            | 279 |
| Table 10-9.  | Tai Le Tone Marks . . . . .                            | 284 |
| Table 10-10. | Myanmar Digits . . . . .                               | 284 |
| Table 10-11. | Hanunóo and Buhid Vowel Sign Combinations . . . . .    | 287 |
| Table 11-1.  | Initial Sources for Unified Han . . . . .              | 294 |
| Table 11-2.  | Common Han Characters . . . . .                        | 296 |
| Table 11-3.  | Source Encoding for Sword Variants . . . . .           | 300 |
| Table 11-4.  | Ideographs Not Unified . . . . .                       | 302 |
| Table 11-5.  | Ideographs Unified . . . . .                           | 302 |
| Table 11-6.  | Han Ideograph Arrangement . . . . .                    | 303 |
| Table 11-7.  | Sources Added for Extension B . . . . .                | 304 |
| Table 11-8.  | Mandarin Tone Marks . . . . .                          | 310 |
| Table 11-9.  | Minnan and Hakka Tone Marks . . . . .                  | 311 |
| Table 11-10. | Line-Based Placement of Jungseong . . . . .            | 316 |
| Table 12-1.  | Labialized Forms in -WAA . . . . .                     | 322 |
| Table 12-2.  | Labialized Forms in -WE . . . . .                      | 323 |
| Table 13-1.  | Similar Characters in Linear B and Cypriot . . . . .   | 346 |
| Table 14-1.  | Other Currency Symbols . . . . .                       | 352 |
| Table 14-2.  | Mathematical Alphanumeric Symbols . . . . .            | 355 |
| Table 14-3.  | Use of Symbol Pieces . . . . .                         | 366 |
| Table 14-4.  | Japanese Era Names . . . . .                           | 373 |
| Table 14-5.  | Precomposed Note Characters . . . . .                  | 379 |
| Table 14-6.  | Alternative Noteheads . . . . .                        | 379 |

*Tables*

|             |   |      |
|-------------|---|------|
| Table 14-7. | Augmentation Dots and Articulation Symbols . . . . .      | 380  |
| Table 14-8. | Examples of Ornamentation . . . . .                       | 380  |
| Table 15-1. | Control Codes Specified in the Unicode Standard . . . . . | 386  |
| Table 15-2. | Bidirectional Ordering Controls . . . . .                 | 392  |
| Table 15-3. | Unicode Encoding Form Signatures . . . . .                | 402  |
| Table C-1.  | Timeline . . . . .  | 1348 |
| Table C-2.  | Zero Extending . . . . .                                  | 1351 |
| Table D-1.  | Versions of Unicode and ISO/IEC 10646-1 . . . . .         | 1355 |
| Table D-2.  | Allocation of BMP Code Points . . . . .                   | 1356 |
| Table D-3.  | Allocation of Supplementary Code Points . . . . .         | 1356 |
| Table D-4.  | Overall Allocation of Unicode Code Points . . . . .       | 1356 |