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.

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

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

Appendix B

Abstracts of Unicode Technical Reports

The following abstracts are divided into three categories: Unicode Standard Annexes, Unicode Technical Standards, and Unicode Technical Reports. They are listed numerically within each category. There are gaps because some of them have been superseded or incorporated into the standard. The Unicode Web site has a complete list of where all of them are available.

A Unicode Technical Report (UTR) may contain informative material, normative specifications, or both. Each UTR may specify a base version of the Unicode Standard. In that case, conformance to the UTR requires conformance to that version of the standard or higher.

There are two specially distinguished types of approved Unicode Technical Reports that are given more authoritative status by the Unicode Consortium.

A Unicode Standard Annex (UAX) forms an integral part of the Unicode Standard, carrying the same version number as the standard, but is published as a separate document. Note that conformance to a version of the Unicode Standard includes conformance to its Unicode Standard Annexes.

A Unicode Technical Standard (UTS) is an independent specification. Conformance to the Unicode Standard does not imply conformance to any UTS. Each UTS specifies a base version of the Unicode Standard.

B.1 Unicode Standard Annexes

UAX #9: The Bidirectional Algorithm

This document describes specifications for the positioning of characters flowing from right to left, such as Arabic or Hebrew.

UAX #11: East Asian Width

This report presents the specifications of an informative property for Unicode characters that is useful when interoperating with East Asian legacy character sets.

UAX #14: Line Breaking Properties

This report presents the specification of line breaking properties for Unicode characters.

UAX #15: Unicode Normalization Forms

This document describes specifications for four normalized forms of Unicode text. With these forms, equivalent text (canonical or compatibility) will have identical binary representations. When implementations keep strings in a normalized form, they can be assured that equivalent strings have a unique binary representation.

UAX #24: Script Names

This document provides an assignment of script names to all Unicode code points. This information is useful in mechanisms such as regular expressions, where it produces much better results than simple matches on block names.

UAX #29: Text Boundaries

This document describes guidelines for determining default boundaries between certain significant text elements: grapheme clusters ("user characters"), words, and sentences.

B.2 Unicode Technical Standards

UTS #6: A Standard Compression Scheme for Unicode

This report presents the specifications of a compression scheme for Unicode and sample implementation.

UTS #10: Unicode Collation Algorithm

This report provides the specification of the Unicode Collation Algorithm, which provides a specification for how to compare two Unicode strings while remaining conformant to the requirements of the Unicode Standard.

B.3 Unicode Technical Reports

UTR #16: UTF-EBCDIC

This document presents the specifications of UTF-EBCDIC: EBCDIC Friendly Unicode (or UCS) Transformation Format.

UTR #17: Character Encoding Model

This document clarifies a number of the terms used to describe character encodings and indicates where the different encoding forms of the Unicode Standard fit in. It elaborates the Internet Architecture Board's (IAB) three-layer "text stream" definitions into a five-layer structure.

UTR #18: Unicode Regular Expression Guidelines

This document describes guidelines for how to adapt regular expression engines for use with the Unicode Standard.

UTR #20: Unicode in XML and Other Markup Languages

This document contains guidelines on the use of the Unicode Standard in conjunction with markup languages such as XML.

UTR #22: Character Mapping Markup Language (CharMapML)

This document specifies an XML format for the interchange of mapping data for character encodings. It provides a complete description for such mappings in terms of a defined mapping to and from Unicode code points, and a description of alias tables for the interchange of mapping table names.

UTR #26: Compatibility Encoding Scheme for UTF-16: 8-Bit (CESU-8)

This document specifies an 8-bit Compatibility Encoding Scheme for UTF-16 (CESU) that is intended for internal use within systems processing Unicode to provide an ASCII-compatible 8-bit encoding that is similar to UTF-8 but preserves UTF-16 binary collation. *It is not intended or recommended as an encoding used for open information exchange.* The Unicode Consortium does not encourage the use of CESU-8, but does recognize the existence of data in this encoding and supplies this technical report to clearly define the format and to distinguish it from UTF-8. This encoding does not replace or amend the definition of UTF-8.

B.4 Other Unicode References

There is a wealth of other information available on the Unicode Web site. Some of the most important of these references are listed here.

Unicode Technical Notes

http://www.unicode.org/notes/

Unicode Technical Notes (UTN) publish information that may be of interest to implementers or readers of the Unicode Standard, or to users of programs implementing the standard. However, the Technical Notes are not formally reviewed by the Unicode Technical Committee and are not part of the Unicode Standard. Their publication does not imply endorsement by the Unicode Consortium in any way. Current topics at the time of publication include the following:

- UTN #1, “Issues in Indic Language Collation”
- UTN #2, “Rendering Combining Marks”
- UTN #3, “Encoding Scripts from the Past”
- UTN #4, “Leaks in the Unicode Pipeline”
- UTN #5, “Canonical Equivalences in Applications”
- UTN #6, “BOCU-1: MIME-Compatible Compression”
- UTN #7, “Migrating Software to Supplementary Characters”
- UTN #8, “Toward a Model for Language Identification”

FAQ (Frequently Asked Questions)

http://www.unicode.org/faq/

The FAQ pages provide an invaluable resource for understanding the Unicode Standard, and its implications for users and implementers.

Charts

http://www.unicode.org/charts/

The charts section of the Web site provides online charts for all of the Unicode characters, plus specialized charts for normalization, collation, case mapping, script names, and Unified CJK Ideographs.

Conferences

http://www.unicode.org/conference/

The Internationalization & Unicode Conferences are of particular value to anyone implementing the Unicode Standard or working on internationalization. A variety of tutorials and conference sessions cover current topics related to the Unicode Standard, the World Wide Web, software, internationalization, and localization.

Policies

http://www.unicode.org/policies/

These pages describe Unicode Consortium policies on stability, patents, and Unicode Web site privacy. The stability policies are particularly important for implementers, documenting invariants for the Unicode Standard that allow implementations to be compatible with future and past versions.

Updates and Errata

http://www.unicode.org/errata/

This page lists periodic updates with corrections of typographic errors and new clarifications of the text.

Versions

http://www.unicode.org/versions/

This page describes the version numbering used in the Unicode Standard, the nature of the Unicode character repertoire, and ways to cite and reference the Unicode Standard, the Unicode Character Database, and Unicode Technical Reports. It also specifies the exact contents of each and every version of the Unicode Standard, back to Unicode 1.0.0.

Where Is My Character?

http://www.unicode.org/standard/where/

This page provides basic guidance to finding Unicode characters, especially those whose glyphs do not appear in the charts, or which are represented by sequences of Unicode characters.