Preface

This book, *The Unicode Standard, Version 3.0*, is the authoritative source of information on the Unicode character encoding standard, the international character code for information processing that includes all major scripts of the world and is the foundation for development of software for worldwide use. As well as encoding characters used for written communication in a simple and consistent manner, the Unicode Standard defines character properties and algorithms for use in implementations.

Version 3.0 expands on material from Versions 2.0 and 2.1 and supersedes all other previous versions. The previous versions of the Unicode Standard are:

- *The Unicode Standard, Version 1.0*, Volume 1 (1991)
- *The Unicode Standard, Version 1.0*, Volume 2 (1992)
- *The Unicode Standard, Version 1.1*, Unicode Technical Report #4 (1993)
- *The Unicode Standard, Version 2.0* (1996)
- *The Unicode Standard, Version 2.1*, Unicode Technical Report #8 (1998)

Major additions to Version 3.0 include:

- conformance rules for transformation formats
- new scripts including Ethiopic, Khmer, Mongolian, Myanmar, and Sinhala
- restructured and enhanced character block descriptions
- clarified bidirectional algorithm
- updated implementation guidelines
- a Shift-JIS index

The Unicode Standard maintains consistency with the international standard ISO/IEC 10646. Version 3.0 of the Unicode Standard corresponds to ISO/IEC 10646-1:2000.

0.1 About the Unicode Standard

This book defines Version 3.0 of the Unicode Standard. The general principles and architecture of the Unicode Standard, requirements for conformance, and guidelines for implementers precede the actual coding information. Useful ancillary information is given in the appendices. The accompanying CD-ROM contains tables of use to implementers and all technical reports published to date.

Concepts, Architecture, Conformance, and Guidelines

The first five chapters of Version 3.0 introduce the Unicode Standard and provide the information an engineer needs to produce a conforming implementation. Basic text processing, working with combining marks, encoding forms, and doing bidirectional text layout are all described. A special chapter on implementation guidelines answers many common questions that arise when implementing Unicode.

> *Chapter 1* introduces the standard's basic concepts, design basis, and coverage, and discusses basic text handling requirements.
>
> *Chapter 2* sets forth the fundamental principles underlying the Unicode Standard and covers specific topics such as text processes, overall character properties, and the use of combining marks.
>
> *Chapter 3* constitutes the formal statement of conformance. This chapter also presents the normative algorithms for three processes: the canonical ordering of combining marks, the encoding of Korean Hangul syllables by conjoining *jamo*, and the formatting of bidirectional text.
>
> *Chapter 4* describes character properties in detail, both normative (required) and informative. Tables giving additional character property information appear on the CD-ROM.
>
> *Chapter 5* discusses implementation issues, including compression, strategies for dealing with unknown and unsupported characters, and transcoding to other standards.

Character Block Descriptions

Chapters 6 through 13 contain the character block descriptions that give basic information about each script or collection and may discuss specific characters or pertinent layout information.

> *Chapter 6* describes the general punctuation characters.
>
> *Chapter 7* presents the European Alphabetic scripts, including Latin, Greek, Cyrillic, Armenian, Georgian, Runic, Ogham, and associated combining marks.
>
> *Chapter 8* presents the Middle Eastern, right-to-left scripts: Hebrew, Arabic, Syriac, and Thaana.
>
> *Chapter 9* covers the South and Southeast Asian scripts, including Devanagari, Bengali, Gurmukhi, Gujarati, Oriya, Tamil, Telugu, Kannada, Malayalam, Sinhala, Tibetan, Thai, Lao, Khmer, and Myanmar.
>
> *Chapter 10* presents the East Asian scripts, including Han, Hiragana, Katakana, Hangul, Bopomofo, and Yi.
>
> *Chapter 11* presents other scripts, including Ethiopic, Cherokee, Canadian Aboriginal Syllabics, and Mongolian.
>
> *Chapter 12* presents symbols, including currency, letterlike and technical symbols, and mathematical operators.
>
> *Chapter 13* describes special characters such as the Private Use Area, surrogates, and specials.

Charts and Index

The next two chapters document the Unicode Standard's character code assignments, their names and important descriptive information, and Han indices that aid in locating specific ideographs encoded in Unicode.

Chapter 14 gives the code charts and the Character Names List. The code charts contain the normative character encoding assignments, and the names list contains normative information as well as useful cross references and informational notes.

Chapter 15 provides a radical-stroke index to East Asian ideographs, as well as a Shift-JIS index.

Appendices and Tables

The appendices contain detailed background information on important topics: character encoding systems, submission of proposals, and the history of Unicode and its relationship to ISO/IEC 10646.

Appendix A describes the history of Han Unification in the Unicode Standard.

Appendix B gives instructions on how to submit characters for consideration as additions to the Unicode Standard.

Appendix C details the relationship between the Unicode Standard and ISO/IEC 10646.

Appendix D lists the changes to the Unicode Standard since Version 2.0.

The appendices are followed by a glossary of terms, a bibliography, and two indices: an index to Unicode characters and an index to the text of Chapters 1 through 15.

The Unicode Character Database and Technical Reports

The Unicode Character Database is the name for a collection of files that contain character code values, character names, and character property data. It is described more fully in the file UnicodeCharacterDatabase.html. Version 3.0.0 of the database is provided on the accompanying CD-ROM. Updates and revisions will be made available online. For information on the latest available version see:

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

The following Unicode Technical Reports are formally part of this standard:

- UTR #11: East Asian Width, Version 5.0
- UTR #13: Unicode Newline Guidelines, Version 5.0
- UTR #14: Line Breaking Properties, Version 6.0
- UTR #15: Unicode Normalization Forms, Version 18.0

The latest available version of these reports is provided on the CD-ROM. Updates and revisions will be made available online. For information on the latest available version, see http://www.unicode.org/unicode/standard/versions/.

On the CD-ROM

The CD-ROM contains the *Unicode Character Database*, which gives character codes, character names, character properties, and decompositions for decomposable or compatibility characters. In addition to the Unicode Character Database and Unicode Technical Reports that are part of this standard, the CD-ROM also contains additional technical reports (covering topics such as compression, collation, and transformation formats), as well as property-based mapping tables (for example, tables for case) and transcoding tables for international, national, and industry character sets (including the Han cross-reference table). For the complete contents of the CD-ROM, see its READ ME file. Please consult the Unicode Consortium's online resources (see *Section 0.3, Resources*) to obtain the most up-to-date versions of the materials on the CD-ROM.

0.2 Notational Conventions

Throughout this book, certain typographic conventions are used. In running text, an individual Unicode value is expressed as *U+nnnn*, where *nnnn* is a four-digit number in hexadecimal notation, using the digits 0–9 and the letters A–F (for 10 through 15, respectively).

- U+0416 is the Unicode value for the character named CYRILLIC CAPITAL LETTER ZHE.

In tables, the *U+* may be omitted for brevity.

A range of Unicode values is expressed as *U+xxxx→U+yyyy*, or *U+xxxx—U+yyyy*, or *xxxx..yyyy*, where *xxxx* and *yyyy* are the first and last Unicode values in the range, and the arrow, long dash, or two dots indicate a contiguous range inclusive of the endpoints.

- The range U+0900→U+097F contains 128 character values.

All Unicode characters have unique names, which are identical to those of the English-language edition of International Standard ISO/IEC 10646. Unicode character names contain only uppercase Latin letters A through Z, digits, space, and hyphen-minus; this convention makes it easy to generate computer-language identifiers automatically from the names. Unified East Asian ideographs are named CJK UNIFIED IDEOGRAPH-X, where X is replaced with the hexadecimal Unicode value—for example, CJK UNIFIED IDEOGRAPH-4E00. The names of Hangul syllables are generated algorithmically; for details, see Hangul Syllable Names in *Section 3.11, Conjoining Jamo Behavior*.

In running text, a formal Unicode name is shown in small capitals (for example, GREEK SMALL LETTER MU), and alternative names (aliases) appear in italics (for example, *umlaut*). Italics are also used to refer to a text element that is not explicitly encoded (for example, *pasekh alef*) or to set off a foreign word (for example, the Welsh word *ynghyd*). Phonemic transcriptions are shown between slashes, as in Khmer /khnyom/.

The symbols used in the character names list are described at the beginning of *Chapter 14, Code Charts*.

In the text of this book, the word “Unicode” when used alone as a noun refers to the Unicode Standard.

In this book, unambiguous dates of the current common era, such as 1999, are unlabeled. In cases of ambiguity, CE is used. Dates before the common era are labeled with BCE.

Extended BNF

The Unicode Standard and technical reports use an extended BNF format for describing syntax. As different conventions are used for BNF, *Table 0-1, Extended BNF*, lists the notation used here.

A sequence of characters is sometimes listed in text with angle brackets, such as <a, grave> or <U+0061, U+0300>.

Character Classes. A character class is constructed from one or two base sets. It is either a single base set, the negation of a base set, or the (set) difference between two base sets. The

Table 0-1. Extended BNF

Symbols	Meaning
x := ...	production rule
x y	the sequence consisting of x then y
x*	zero or more occurrences of x
x?	zero or one occurrence of x
x+	one or more occurrences of x
x | y	either x or y
(x)	for grouping
x || y	equivalent to (x | y | (x y))
{ x }	equivalent to (x)?
"abc"	string literals (“_” is sometimes used to denote space for clarity)
’abc’	string literals (alternative form)
\u1234	Unicode characters within string literals or character classes
\v00101234	Unicode scalar values within string literals or character classes
U+HHHH	Unicode character literal: equivalent to ‘\uHHHH’
U-HHHHHHHH	Unicode character literal: equivalent to ‘\vHHHHHHHH’
charClass	character class (syntax below)

base sets themselves are bounded by brackets, and contain lists of characters, ranges of characters, general categories, or negations of general categories. The syntax follows:

```
charClass := baseSet | '¬' baseSet | baseSet '-' baseSet
baseSet := '[' item (','? item)* ']'
item    := char | char '-' char | '{' '¬'? category '}'
```

General categories are defined in *Chapter 4, Character Properties*, such as [{Uppercase Letter}] for *uppercase letter*. Main categories such as [{Mark}] are the equivalent of a list of multiple subcategories: [{Non-Spacing Mark}{Spacing Combining Mark}{Enclosing Mark}]. Examples are found in *Table 0-2, Character Class Examples*.

Table 0-2. Character Class Examples

Syntax	Matches
[a-z]	English lowercase letters
[a-z]-[c]	English lowercase letters except for c
¬[c]	all characters but c
[0-9]	European decimal digits
[\u0030-\u0039]	(same as above, using Unicode escapes)
[0-9,A-F,a-f]	hexadecimal digits
[{Letter},{Non-Spacing Mark}]	all letters and non-spacing marks
[{L},{Mn}]	(same as above, using abbreviated notation)
[{¬Cn}]	all assigned Unicode characters
[\u0600-\u06FF]-[{Cn}]	all assigned Arabic characters

Operators

Operators used in this standard are listed in *Table 0-3, Operators*.

Table 0-3. Operators

÷	allow break here (see *Section 5.15, Locating Text Element Boundaries*)
×	do not allow a break here
→	is transformed to, or behaves like
/	integer division (rounded down)
%	modulo operation; equivalent to the integer remainder for positive numbers.

0.3 Resources

The Unicode Consortium provides a number of online resources for obtaining information and data about the Unicode Standard, as well as updates and corrigenda. They are listed below.

Unicode Web Site

- http://www.unicode.org

Unicode Anonymous FTP Site

- ftp://ftp.unicode.org

Unicode Public Mailing List

- unicode@unicode.org

Subscription instructions for the public mailing list are posted on the Unicode Web site.

How to Contact the Unicode Consortium

Contact the Consortium for membership information and to order publications (including additional copies of this book).

- Electronic mail address: info@unicode.org
- Postal address:

  P.O. Box 391476
  Mountain View, CA 94039-1476
  USA

Please check the Web site for up-to-date contact information, including telephone, fax, and courier delivery address.

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.

ISBN 0-201-61633-5

This book is set in Minion, designed by Rob Slimbach at Adobe Systems, Inc. It was typeset using FrameMaker 5.5 running under Windows NT. ASMUS, Inc. created custom software for chart layout. The Han radical-stroke index was typeset by Apple Computer, Inc. The following companies and organizations supplied fonts:

Apple Computer, Inc.
Atelier Fluxus Virus
Beijing Zhong Yi (Zheng Code) Electronics Company
DecoType, Inc.
IBM Corporation
Monotype Typography, Inc.
Microsoft Corporation
Peking University Founder Group Corporation
Production First Software

Additional fonts were supplied by individuals as listed in the *Acknowledgments*.

The publisher offers discounts on this book when ordered in quantity for special sales. For more information please contact:

Corporate, Government, and Special Sales
Addison Wesley Longman, Inc.
One Jacob Way
Reading, Massachusetts 01867

Visit A-W on the Web: http://www.awl.com/cseng/

First printing, January 2000.