The Unicode Standard
Version 8.0 – Core Specification

To learn about the latest version of the Unicode Standard, see http://www.unicode.org/versions/latest/.

The Unicode Standard / the Unicode Consortium ; edited by Julie D. Allen ... [et al.]. — Version 8.0
Includes bibliographical references and index.
ISBN 978-1-936213-10-8 (http://www.unicode.org/versions/Unicode8.0.0/)
1. Unicode (Computer character set) I. Allen, Julie D. II. Unicode Consortium.
QA268.U545 2015

ISBN 978-1-936213-10-8
Published in Mountain View, CA
August 2015

Preface

This is *The Unicode Standard, Version 8.0*. It supersedes all earlier versions of the Unicode Standard.

Why Unicode?

The Unicode Standard and its associated specifications provide programmers with a single universal character encoding, extensive descriptions, and a vast amount of data about how characters function. The specifications and data describe how to form words and break lines; how to sort text in different languages; how to format numbers, dates, times, and other elements appropriate to different languages; how to display languages whose written form flows from right to left, such as Arabic and Hebrew, or whose written form splits, combines, and reorders, such as languages of South Asia. These specifications include descriptions of how to deal with security concerns regarding the many "look-alike" characters from alphabets around the world. Without the properties and algorithms in the Unicode Standard and its associated specifications, interoperability between different implementations would be impossible, and much of the vast breadth of the world's languages would lie outside the reach of modern software.

What's New?

Key new features that have been defined and documented since the publication of *The Unicode Standard, Version 7.0* include:

- a rewritten description of casing to account for the addition of a set of lowercase Cherokee syllables
- substantially revised documentation for emoji symbols, including the new symbol modifiers for implementing skin tone diversity
- an updated description of New Tai Lue specifying the change of model from logical to visual
- descriptions for five new scripts and Sutton SignWriting
- improvements to existing descriptions, including Malayalam, Bengali, Devanagari, CJK punctuation, and tag characters

Support for Languages and Symbol Sets. The new scripts added in Version 8.0 are:

Ahom	Hatran	Old Hungarian
Anatolian Hieroglyphs	Multani	Sutton SignWriting

A total of 7,716 characters were added in Version 8.0 of the Unicode Standard, including 5,771 CJK unified ideographs.

With Version 8.0, support for lesser-used languages was extended worldwide, including:

- letters to support the Ik language in Uganda, Kulango in the Côte d'Ivoire, and other languages of Africa
- the Ahom script for support of the Tai Ahom language in India
- Arabic letters to support Arwi—the Tamil language written in the Arabic script

Conformance Updates. Only minor conformance updates were made in Version 8.0.

Property and Behavioral Updates. The core data files of the Unicode Character Database were updated for the new additions in Version 8.0.

Detailed Change Information. See *Appendix D, Changes from Previous Versions* and http://www.unicode.org/versions/Unicode8.0.0/ for detailed information about the changes from the previous versions of the standard, including character counts, conformance clause and definition updates, and significant changes to the Unicode Character Database and Unicode Standard Annexes.

Organization of This Standard

This core specification, together with the Unicode code charts, the Unicode Character Database, and the Unicode Standard Annexes, defines Version 8.0 of the Unicode Standard. The core specification contains the general principles, requirements for conformance, and guidelines for implementers. The character code charts and names are also available online.

Concepts, Architecture, Conformance, and Guidelines. The first five chapters of Version 8.0 introduce the Unicode Standard and provide the fundamental information needed to produce a conforming implementation. Basic text processing, working with combining marks, encoding forms, and normalization 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 several processes, including normalization, Korean syllable boundary determination, and default casing.

Chapter 4 describes character properties in detail, both normative (required) and informative. Additional character property information appears in Unicode Standard Annex #44, "Unicode Character Database."

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 23* contain the character block descriptions that provide basic information about each script or group of symbols and may discuss specific characters or pertinent layout information. Some of this information is required to produce conformant implementations of these scripts and other collections of characters.

Code Charts. *Chapter 24* describes the conventions used in the code charts and the list of character names. 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.

Appendices. The appendices contain additional information.

Appendix A documents the notational conventions used by the standard.

Appendix B provides abstracts of Unicode Technical Reports and lists other important Unicode resources.

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

Appendix D lists the changes to clauses and definitions in the Unicode Standard since Version 7.0.

Appendix E describes the history of Han unification in the Unicode Standard.

Appendix F provides additional documentation for characters encoded in the CJK Strokes block (U+31C0..U+31EF).

References and Index. The appendices are followed by a bibliography and an index to the text of this core specification.

Glossary and Character Index. A glossary of Unicode terms and the Unicode Character Name Index may be found at:

http://www.unicode.org/glossary/

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

Unicode Standard Annexes

The Unicode Standard Annexes form an integral part of the Unicode Standard. Conformance to a version of the Unicode Standard includes conformance to its Unicode Standard Annexes. All versions, including the most up-to-date versions of all Unicode Standard Annexes, are available at:

http://www.unicode.org/reports/index.html#annexes

The following is a list of Unicode Standard Annexes:

Unicode Standard Annex #9, "Unicode Bidirectional Algorithm," describes specifications for the positioning of characters in text containing characters flowing from right to left, such as Arabic or Hebrew.

Unicode Standard Annex #11, "East Asian Width," presents the specification of an informative property for Unicode characters that is useful when interoperating with East Asian legacy character sets.

Unicode Standard Annex #14, "Unicode Line Breaking Algorithm," presents the specification of line breaking properties for Unicode characters.

Unicode Standard Annex #15, "Unicode Normalization Forms," describes Unicode normalization and provides examples and implementation strategies for it.

Unicode Standard Annex #24, "Unicode Script Property," describes two related Unicode code point properties. Both properties share the use of Script property values. The Script property itself assigns single script values to all Unicode code points, identifying a primary script association, where possible. The Script_Extensions property assigns sets of Script property values, providing more detail for cases where characters are commonly used with multiple scripts.

Unicode Standard Annex #29, "Unicode Text Segmentation," describes algorithms for determining default boundaries between certain significant text elements: grapheme clusters ("user-perceived characters"), words, and sentences.

Unicode Standard Annex #31, "Unicode Identifier and Pattern Syntax," describes specifications for recommended defaults for the use of Unicode in the definitions of identifiers and in pattern-based syntax.

Unicode Standard Annex #34, "Unicode Named Character Sequences," defines the concept of a Unicode named character sequence.

Unicode Standard Annex #38, "Unicode Han Database (Unihan)," describes the organization and content of the Unihan database.

Unicode Standard Annex #41, "Common References for Unicode Standard Annexes," contains the listing of references shared by other Unicode Standard Annexes.

Unicode Standard Annex #42, "Unicode Character Database in XML," describes an XML representation of the Unicode Character Database.

Unicode Standard Annex #44, "Unicode Character Database," provides the core documentation for the Unicode Character Database (UCD). It describes the layout and organization of the Unicode Character Data-

base and how the UCD specifies the formal definition of Unicode character properties.

Unicode Standard Annex #45, "U-Source Ideographs," describes U-source ideographs as used by the Ideographic Rapporteur Group (IRG) in its CJK ideograph unification work.

The Unicode Character Database

The Unicode Character Database (UCD) is a collection of data files containing character code points, character names, and character property data. It is described more fully in *Section 4.1, Unicode Character Database* and in Unicode Standard Annex #44, "Unicode Character Database." All versions, including the most up-to-date version of the Unicode Character Database, are found at:

http://www.unicode.org/ucd/

Information on versioning and on all versions of the Unicode Standard can be found at:

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

Unicode Code Charts

The Unicode code charts contain the character encoding assignments and the names list. The archival, reference set of versioned 8.0 code charts may be found at:

http://www.unicode.org/charts/PDF/Unicode-8.0/

For easy lookup of characters, see the current code charts:

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

An interactive radical-stroke index to CJK ideographs is located at:

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

Unicode Technical Standards and Unicode Technical Reports

Unicode Technical Reports and Unicode Technical Standards are separate publications and do not form part of the Unicode Standard.

All versions of all Unicode Technical Reports and Unicode Technical Standards are available at:

http://www.unicode.org/reports/

See *Appendix B, Unicode Publications and Resources*, for a summary overview of important Unicode Technical Standards and Unicode Technical Reports.

Updates and Errata

Reports of errors in the Unicode Standard, including the Unicode Character Database and the Unicode Standard Annexes, may be reported using the reporting form:

http://www.unicode.org/reporting.html

A list of known errata is maintained at:

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

Any currently listed errata will be fixed in subsequent versions of the standard.

Acknowledgements

The Unicode Standard, Version 8.0 is the result of the dedication and contributions of many people over several years. We would like to acknowledge the individuals whose contributions were central to the design, authorship, and review of this standard. A complete listing of acknowledgements can be found at:

http://www.unicode.org/acknowledgements/