Electronic Edition

This file is part of the electronic edition of *The Unicode Standard, Version 5.0*, provided for online access, content searching, and accessibility. It may not be printed. Bookmarks linking to specific chapters or sections of the whole Unicode Standard are available at

http://www.unicode.org/versions/Unicode5.0.0/bookmarks.html

Purchasing the Book

For convenient access to the full text of the standard as a useful reference book, we recommend purchasing the printed version. The book is available from the Unicode Consortium, the publisher, and booksellers. Purchase of the standard in book format contributes to the ongoing work of the Unicode Consortium. Details about the book publication and ordering information may be found at

http://www.unicode.org/book/aboutbook.html

Joining Unicode

You or your organization may benefit by joining the Unicode Consortium: for more information, see Joining the Unicode Consortium at

http://www.unicode.org/consortium/join.html

Library of Congress Cataloging-in-Publication Data

The Unicode Standard / the Unicode Consortium ; edited by Julie D. Allen ... [et al.]. — Version 5.0.
p. cm.
Includes bibliographical references and index.
ISBN 0-321-48091-0 (hardcover : alk. paper)
1. Unicode (Computer character set) I. Allen, Julie D.
II. Unicode Consortium.
QA268.U545 2007
005.7'22—dc22

2006023526

ISBN 0-321-48091-0
Text printed in the United States on recycled paper at Courier in Westford, Massachusetts.
First printing, October 2006

Glossary

Abjad. A writing system in which only consonants are indicated. The term "abjad" is derived from the first four letters of the traditional order of the Arabic script: *alef, beh, jeem, dal*. (See *Section 6.1, Writing Systems*.)

Abstract Character. A unit of information used for the organization, control, or representation of textual data. (See definition D7 in *Section 3.4, Characters and Encoding*.)

Abstract Character Sequence. An ordered sequence of one or more abstract characters. (See definition D8 in *Section 3.4, Characters and Encoding*.)

Abugida. A writing system in which consonants are indicated by the base letters that have an inherent vowel, and in which other vowels are indicated by additional distinguishing marks of some kind modifying the base letter. The term "abugida" is derived from the first four letters of the Ethiopic script in the Semitic order: *alf, bet, gaml, dant*. (See *Section 6.1, Writing Systems*.)

Accent Mark. A mark placed above, below, or to the side of a character to alter its phonetic value. (See also *diacritic*.)

Acrophonic. Denoting letters or numbers by the first letter of their name. For example, the Greek acrophonic numerals are variant forms of such initial letters.

Aksara. (1) In Sanskrit grammar, the term for "letter" in general, as opposed to consonant (*vyanjana*) or vowel (*svara*). Derived from the first and last letters of the traditional ordering of Sanskrit letters—"a" and "ksha". (2) More generally, in Indic writing systems, *aksara* refers to a "syllable," consisting of a consonant plus vowel sequence, where the vowel may or may not be the inherent vowel of the consonant letter. When multiple consonants are involved, the *aksara* represents the entire orthographic syllable, which can include two or more leading consonants that may be visually presented in conjunct forms; in such cases, the *aksara* may not be identical to the phonological syllable.

Algorithm. A term used in a broad sense in the Unicode Standard, to mean the logical description of a process used to achieve a specified result. This does not require the actual procedure described in the algorithm to be followed; any implementation is conformant as long as the results are the same.

Alphabet. A writing system in which both consonants and vowels are indicated. The term "alphabet" is derived from the first two letters of the Greek script: *alpha, beta*. (See *Section 6.1, Writing Systems*.)

Alphabetic Property. Informative property of the primary units of alphabets and/or syllabaries. (See *Section 4.10, Letters, Alphabetic, and Ideographic*.)

Alphabetic Sorting. (See *collation*.)

Annotation. The association of secondary textual content with a point or range of the primary text. (The value of a particular annotation *is* considered to be a part of the "content" of the text. Typical examples include glossing, citations, exemplification, Japanese yomi, and so on.)

ANSI. (1) The American National Standards Institute. (2) The Microsoft collective name for all Windows code pages. Sometimes used specifically for code page 1252, which is a superset of ISO/IEC 8859-1.

Apparatus Criticus. Collection of conventions used by editors to annotate and comment on text.

Arabic Digits. Forms of decimal digits used in most parts of the Arabic world (for instance, U+0660 ٠, U+0661 ١, U+0662 ٢, U+0663 ٣). Although *European digits* (1, 2, 3,…) derive historically from these forms, they are visually distinct and are coded separately. (Arabic digits are sometimes called Indic numerals; however, this nomenclature leads to confusion with the digits currently used with the scripts of India.) Arabic digits are referred to as *Arabic-Indic digits* in the Unicode Standard. Variant forms of Arabic digits used chiefly in Iran and Pakistan are referred to as *Eastern Arabic-Indic digits*. (See *Section 8.2, Arabic*.)

ASCII. (1) The American Standard Code for Information Interchange, a 7-bit coded character set for information interchange. It is the U.S. national variant of ISO/IEC 646 and is formally the U.S. standard ANSI X3.4. It was proposed by ANSI in 1963 and finalized in 1968. (2) The set of 128 Unicode characters from U+0000 to U+007F, including control codes as well as graphic characters. (3) ASCII has been incorrectly used to refer to various 8-bit character encodings that include ASCII characters in the first 128 positions.

Assigned Character. Synonym for assigned to an abstract character. This refers to graphic, format, control, and private-use characters that have been encoded in the Unicode Standard. (See *Section 2.4, Code Points and Characters*.)

Assigned Code Point. (See *designated code point*.)

Atomic Character. A character that is not decomposable. (See *decomposable character*.)

Base Character. Any graphic character except for those with the General Category of Combining Mark (M). (See definition D51 in *Section 3.6, Combination*.) In a combining character sequence, the base character is the initial character, which the combining marks are applied to.

Basic Multilingual Plane. Plane 0, abbreviated as BMP.

Bicameral. A script that distinguishes between two cases. (See *case*.) Most often used in the context of European alphabets.

BIDI. Abbreviation of bidirectional, in reference to mixed left-to-right and right-to-left text.

Bidirectional Display. The process or result of mixing left-to-right text and right-to-left text in a single line. (See Unicode Standard Annex #9, "The Bidirectional Algorithm.")

Big-endian. A computer architecture that stores multiple-byte numerical values with the most significant byte (MSB) values first.

Binary Files. Files containing nontextual information.

Block. A grouping of related characters within the Unicode encoding space. A block may contain unassigned positions, which are reserved.

BMP. Acronym for *Basic Multilingual Plane*.

BMP Character. A Unicode encoded character having a BMP code point. (See *supplementary character*.)

BMP Code Point. A Unicode code point between U+0000 and U+FFFF. (See *supplementary code point*.)

BNF. Acronym for *Backus-Naur Form*, a formal meta-syntax for describing context-free syntaxes. (For details, see *Appendix A, Notational Conventions*.)

BOCU-1. Acronym for Binary Ordered Compression for Unicode. A Unicode compression scheme that is MIME-compatible (directly usable for e-mail) and preserves binary order, which is useful for databases and sorted lists.

BOM. Acronym for *byte order mark*.

Bopomofo. An alphabetic script used primarily in the Republic of China (Taiwan) to write the sounds of Mandarin Chinese and some other dialects. Each symbol corresponds to either the syllable-initial or syllable-final sounds; it is therefore a subsyllabic script in its primary usage. The name is derived from the names of its first four elements. More properly known as *zhuyin zimu* or *zhuyin fuhao* in Mandarin Chinese.

Boustrophedon. A pattern of writing seen in some ancient manuscripts and inscriptions, where alternate lines of text are laid out in opposite directions, and where right-to-left lines generally use glyphs mirrored from their left-to-right forms. Literally, "as the ox turns," referring to the plowing of a field.

Braille. A writing system using a series of raised dots to be read with the fingers by people who are blind or whose eyesight is not sufficient for reading printed material. (See *Section 15.10, Braille*.)

Braille Pattern. One of the 64 (for six-dot Braille) or 256 (for eight-dot Braille) possible tangible dot combinations.

Byte. (1) The minimal unit of addressable storage for a particular computer architecture. (2) An octet. Note that many early computer architectures used bytes larger than 8 bits in size, but the industry has now standardized almost uniformly on 8-bit bytes. The Unicode Standard follows the current industry practice in equating the term *byte* with *octet* and using the more familiar term *byte* in all contexts. (See *octet*.)

Byte Order Mark. The Unicode character U+FEFF when used to indicate the byte order of a text. (See *Section 2.13, Special Characters and Noncharacters*, and *Section 16.8, Specials*.)

Byte Serialization. The order of a series of bytes determined by a computer architecture.

Byte-Swapped. Reversal of the order of a sequence of bytes.

Canonical. (1) Conforming to the general rules for encoding—that is, not compressed, compacted, or in any other form specified by a higher protocol. (2) Characteristic of a normative mapping and form of equivalence specified in *Chapter 3, Conformance*.

Canonical Decomposable Character. A character that is not identical to its canonical decomposition. (See definition D69 in *Section 3.7, Decomposition*.)

Canonical Decomposition. Mapping to an inherently equivalent sequence—for example, mapping ä to a + ◌̈. (For a full, formal definition, see definition D68 in *Section 3.7, Decomposition*.)

Canonical Equivalent. Two character sequences are said to be canonical equivalents if their full canonical decompositions are identical. (See definition D70 in *Section 3.7, Decomposition*.)

Cantillation Mark. A mark that is used to indicate how a text is to be chanted or sung.

Capital Letter. Synonym for *uppercase letter*. (See *case*.)

Case. (1) Feature of certain alphabets where the letters have two distinct forms. These variants, which may differ markedly in shape and size, are called the *uppercase* letter (also known as *capital* or *majuscule*) and the *lowercase* letter (also known as *small* or *minuscule*). (2) Normative property of characters, consisting of uppercase, lowercase, and titlecase (Lu, Ll, and Lt). (See *Section 4.2, Case—Normative*.)

Case Mapping. The association of the uppercase, lowercase, and titlecase forms of a letter. (See *Section 5.18, Case Mappings*.)

Case-Ignorable. A character C is defined to be *case-ignorable* if C has the value MidLetter for the Word_Break property or its General_Category is one of Nonspacing_Mark (Mn), Enclosing_Mark (Me), Format (Cf), Modifier_Letter (Lm), or Modifier_Symbol (Sk). (See definition D121 in *Section 3.13, Default Case Algorithms*.)

Case-Ignorable Sequence. A sequence of zero or more case-ignorable characters. (See definition D122 in *Section 3.13, Default Case Algorithms*.)

CCS. Acronym for *coded character set*.

Cedilla. A mark originally placed beneath the letter *c* in French, Portuguese, and Spanish to indicate that the letter is to be pronounced as an *s*, as in *façade*. Obsolete Spanish diminutive of *ceda*, the letter *z*.

CEF. Acronym for *character encoding form*.

CES. Acronym for *character encoding scheme*.

Character. (1) The smallest component of written language that has semantic value; refers to the abstract meaning and/or shape, rather than a specific shape (see also *glyph*), though in code tables some form of visual representation is essential for the reader's understanding. (2) Synonym for *abstract character.* (3) The basic unit of encoding for the Unicode character encoding. (4) The English name for the ideographic written elements of Chinese origin. [See *ideograph* (2).]

Character Block. (See *block.*)

Character Class. A set of characters sharing a particular set of properties.

Character Encoding Form. Mapping from a character set definition to the actual code units used to represent the data.

Character Encoding Scheme. A *character encoding form* plus byte serialization. There are seven character encoding schemes in Unicode: UTF-8, UTF-16, UTF-16BE, UTF-16LE, UTF-32, UTF-32BE, and UTF-32LE.

Character Name. A unique string used to identify each abstract character encoded in the standard. (See definition D4 in *Section 3.3, Semantics.*)

Character Name Alias. An additional unique string identifier, other than the character name, associated with an encoded character in the standard. (See definition D5 in *Section 3.3, Semantics.*)

Character Properties. A set of property names and property values associated with individual characters. (See *Chapter 4, Character Properties.*)

Character Repertoire. The collection of characters included in a character set.

Character Sequence. Synonym for *abstract character sequence.*

Character Set. A collection of elements used to represent textual information.

Charset. (See *coded character set.*)

Chillu. Abbreviation for *chilaaksharam* (singular) (*cillakṣaram*). Refers to any of a set of sonorant consonants in Malayalam, when appearing in syllable-final position with no inherent vowel.

Choseong. A sequence of one or more leading consonants in Korean.

Chữ Hán. The name for Han characters used in Vietnam; derived from *hànzì.*

Chữ Nôm. A demotic script of Vietnam developed from components of Han characters. Its creators used methods similar to those used by the Chinese in creating Han characters.

CJK. Acronym for Chinese, Japanese, and Korean. A variant, *CJKV,* means Chinese, Japanese, Korean, and Vietnamese.

CLDR. (See *Common Locale Data Repository.*)

Coded Character. (See *encoded character.*)

Coded Character Sequence. An ordered sequence of one or more code points. Normally, this consists of a sequence of encoded characters, but it may also include noncharacters or reserved code points. (See definition D12 in *Section 3.4, Characters and Encoding*.)

Coded Character Representation. Synonym for *coded character sequence*.

Coded Character Set. A character set in which each character is assigned a numeric code point. Frequently abbreviated as *character set*, *charset,* or *code set*.

Code Page. A coded character set, often referring to a coded character set used by a personal computer—for example, PC code page 437, the default coded character set used by the U.S. English version of the DOS operating system.

Code Point. Any value in the Unicode codespace; that is, the range of integers from 0 to $10FFFF_{16}$. (See definition D10 in *Section 3.4, Characters and Encoding*.)

Code Position. Synonym for *code point*. Used in ISO character encoding standards.

Code Set. (See *coded character set*.)

Codespace. (1) A range of numerical values available for encoding characters. (2) For the Unicode Standard, a range of integers from 0 to $10FFFF_{16}$. (See definition D9 in *Section 3.4, Characters and Encoding*.)

Code Unit. The minimal bit combination that can represent a unit of encoded text for processing or interchange. The Unicode Standard uses 8-bit code units in the UTF-8 encoding form, 16-bit code units in the UTF-16 encoding form, and 32-bit code units in the UTF-32 encoding form. (See definition D77 in *Section 3.9, Unicode Encoding Forms*.)

Code Value. Obsolete synonym for *code unit*.

Collation. The process of ordering units of textual information. Collation is usually specific to a particular language. Also known as *alphabetizing* or *alphabetic sorting*. Unicode Technical Report #10, "Unicode Collation Algorithm," defines a complete, unambiguous, specified ordering for all characters in the Unicode Standard.

Combining Character. A character with the General Category of Combining Mark (M). (See definition D52 in *Section 3.6, Combination*.) (See also *nonspacing mark*.)

Combining Character Sequence. A maximal character sequence consisting of either a base character followed by a sequence of one or more characters where each is a combining character, ZERO WIDTH JOINER, or ZERO WIDTH NON-JOINER; or a sequence of one or more characters where each is a combining character, ZERO WIDTH JOINER, or ZERO WIDTH NON-JOINER. (See definition D56 in *Section 3.6, Combination*.)

Combining Class. A numeric value in the range 0..255 given to each Unicode code point, formally defined as the property Canonical_Combining_Class. (See definition D104 in *Section 3.11, Canonical Ordering Behavior*.)

Common Locale Data Repository. The repository of locale data in XML format maintained by the Unicode Consortium (http://www.unicode.org/cldr/). This repository provides information needed in the localization of software products into a wide variety of lan-

guages, supplying (among other things): date, time, number, and currency formats; sorting, searching, and matching information; and translated names for languages, territories, scripts, currencies, and time zones. (See also *Locale Data Markup Language*.)

Compatibility. (1) Consistency with existing practice or preexisting character encoding standards. (2) Characteristic of a normative mapping and form of equivalence specified in *Section 3.7, Decomposition*.

Compatibility Character. A character that would not have been encoded except for compatibility and round-trip convertibility with other standards. (See *Section 2.3, Compatibility Characters*.)

Compatibility Composite Character. Synonym for *compatibility decomposable character*.

Compatibility Decomposable Character. A character whose compatibility decomposition is not identical to its canonical decomposition. (See definition D66 in *Section 3.7, Decomposition*.)

Compatibility Decomposition. Mapping to a roughly equivalent sequence that may differ in style. (For a full, formal definition, see definition D65 in *Section 3.7, Decomposition*.)

Compatibility Equivalent. Two character sequences are said to be compatibility equivalents if their full compatibility decompositions are identical. (See definition D67 in *Section 3.7, Decomposition*.)

Compatibility Precomposed Character. Synonym for *compatibility decomposable character*.

Compatibility Variant. A character that generally can be remapped to another character without loss of information other than formatting.

Composite Character. (See *decomposable character*.)

Composite Character Sequence. (See *combining character sequence*.)

Conformance. Adherence to a specified set of criteria for use of a standard. (See *Chapter 3, Conformance*.)

Conjunct Form. A ligated form representing a *consonant conjunct*.

Consonant Cluster. A sequence of two or more consonantal sounds. Depending on the writing system, a consonant cluster may be represented by a single character or by a sequence of characters. (Contrast *digraph*.)

Consonant Conjunct. A sequence of two or more adjacent consonantal letterforms, consisting of a sequence of one or more dead consonants followed by a normal, live consonant letter. A consonant conjunct may be ligated into a single conjunct form, or it may be represented by graphically separable parts, such as subscripted forms of the consonant letters. Consonant conjuncts are associated with the Brahmi family of Indic scripts. (See *Section 9.1, Devanagari*.)

Contextual Variant. A text element can have a presentation form that depends on the textual context in which it is rendered. This presentation form is known as a *contextual variant*.

Control Codes. The 65 characters in the ranges U+0000..U+001F and U+007F..U+009F. Also known as *control characters*.

Cursive. Writing where the letters of a word are connected.

DBCS. Acronym for *double-byte character set*.

Dead Consonant. An Indic consonant character followed by a *virama* character. This sequence indicates that the consonant has lost its inherent vowel. (See *Section 9.1, Devanagari*.)

*Decimal Digit*s. Digits that can be used to form decimal-radix numbers.

Decomposable Character. A character that is equivalent to a sequence of one or more other characters, according to the decomposition mappings found in the Unicode Character Database, and those described in *Section 3.12, Conjoining Jamo Behavior*. It may also be known as a *precomposed* character or a *composite* character. (See definition D63 in *Section 3.7, Decomposition*.)

Decomposition. (1) The process of separating or analyzing a text element into component units. These component units may not have any functional status, but may be simply formal units—that is, abstract shapes. (2) A sequence of one or more characters that is equivalent to a decomposable character. (See definition D64 in *Section 3.7, Decomposition*.)

Decomposition Mapping. A mapping from a character to a sequence of one or more characters that is a canonical or compatibility equivalent and that is listed in the character names list or described in *Section 3.12, Conjoining Jamo Behavior*. (See definition D62 in *Section 3.7, Decomposition*.)

Defective Combining Character Sequence. A combining character sequence that does not start with a base character. (See definition D57 in *Section 3.6, Combination*.)

Demotic Script. (1) A script or a form of a script used to write the vernacular or common speech of some language community. (2) A simplified form of the ancient Egyptian hieratic writing.

Dependent Vowel. A symbol or sign that represents a vowel and that is attached or combined with another symbol, usually one that represents a consonant. For example, in writing systems based on Arabic, Hebrew, and Indic scripts, vowels are normally represented as dependent vowel signs.

Deprecated. Of a coded character or a character property, strongly discouraged from use. (Not the same as *obsolete*.)

Deprecated Character. A coded character whose use is strongly discouraged. Such characters are retained in the standard, but should not be used. (See definition D13 in *Section 3.4, Characters and Encoding*.)

Designated Code Point. Any code point that has either been assigned to an abstract character (*assigned characters*) or that has otherwise been given a normative function by the standard (surrogate code points and noncharacters). This definition excludes reserved code points. Also known as *assigned code point*. (See *Section 2.4, Code Points and Characters*.)

Diacritic. (1) A mark applied or attached to a symbol to create a new symbol that represents a modified or new value. (2) A mark applied to a symbol irrespective of whether it changes the value of that symbol. In the latter case, the diacritic usually represents an independent value (for example, an accent, tone, or some other linguistic information). Also called *diacritical mark* or *diacritical*. (See also *combining character* and *nonspacing mark*.)

Diaeresis. Two horizontal dots over a letter, as in *naïve*. The diaeresis is not distinguished from the *umlaut* in the Unicode character encoding. (See *umlaut*.)

Digits. (See *Arabic digits*, *European digits*, and *Indic digits*.)

Digraph. A pair of signs or symbols (two graphs), which together represent a single sound or a single linguistic unit. The English writing system employs many digraphs (for example, *th*, *ch*, *sh*, *qu*, and so on). The same two symbols may not always be interpreted as a digraph (for example, *cathode* versus *cathouse*). When three signs are so combined, they are called a *trigraph*. More than three are usually called an *n-graph*.

Dingbats. Typographical symbols and ornaments.

Diphthong. A pair of vowels that are considered a single vowel for the purpose of phonemic distinction. One of the two vowels is more prominent than the other. In writing systems, diphthongs are sometimes written with one symbol and sometimes with more than one symbol (for example, with a *digraph*).

Direction. (See *paragraph direction*.)

Directionality Property. A property of every graphic character that determines its horizontal ordering as specified in Unicode Standard Annex #9, "The Bidirectional Algorithm." (See *Section 4.4, Directionality—Normative*.)

Display Cell. A rectangular region on a display device within which one or more glyphs are imaged.

Display Order. The order of glyphs presented in text rendering.

Double-Byte Character Set. One of a number of character sets defined for representing Chinese, Japanese, or Korean text (for example, JIS X 0208-1990). These character sets are often encoded in such a way as to allow double-byte character encodings to be mixed with single-byte character encodings. Abbreviated DBCS. (See also *multibyte character set*.)

Ductility. The ability of a cursive font to stretch or compress the connective baseline to effect text justification.

Dynamic Composition. Creation of composite forms such as accented letters or Hangul syllables from a sequence of characters.

EBCDIC. Acronym for Extended Binary-Coded Decimal Interchange Code. A group of coded character sets used on mainframes that consist of 8-bit coded characters. EBCDIC coded character sets reserve the first 64 code positions (x00 to x3F) for control codes, and reserve the range x41 to xFE for graphic characters. The English alphabetic characters are in discontinuous segments with uppercase at xC1 to xC9, xD1 to xD9, xE2 to xE9, and lowercase at x81 to x89, x91 to x99, xA2 to xA9.

Embedding. A concept relevant to bidirectional behavior. (See Unicode Standard Annex #9, "The Bidirectional Algorithm," for detailed terminology and definitions.)

Encapsulated Text. (1) Plain text surrounded by formatting information. (2) Text recoded to pass through narrow transmission channels or to match communication protocols.

Enclosing Mark. A nonspacing mark with the General Category of Enclosing Mark (Me). (See definition D54 in *Section 3.6, Combination*.) Enclosing marks are a subclass of nonspacing marks that surround a base character, rather than merely being placed over, under, or through it.

Encoded Character. An association (or mapping) between an *abstract character* and a *code point*. (See definition D11 in *Section 3.4, Characters and Encoding*.) By itself, an abstract character has no numerical value, but the process of "encoding a character" associates a particular code point with a particular abstract character, thereby resulting in an "encoded character."

Encoding Form. (See *character encoding form*.)

Encoding Scheme. (See *character encoding scheme*.)

Equivalence. In the context of text processing, the process or result of establishing whether two text elements are identical in some respect.

Equivalent Sequence. (See *canonical equivalent*.)

Escape Sequence. A sequence of bytes that is used for code extension. The first byte in the sequence is *escape* (hex 1B).

EUDC. Acronym for end-user defined character. A character defined by an end user, using a private-use code point, to represent a character missing in a particular character encoding. These are common in East Asian implementations.

European Digits. Forms of decimal digits first used in Europe and now used worldwide. Historically, these digits were derived from the Arabic digits; they are sometimes called "Arabic numerals," but this nomenclature leads to confusion with the real *Arabic digits*.

Fancy Text. (See *rich text*.)

Fixed Position Class. A subset of the range of numeric values for combining classes—specifically, any value in the range 10..199. (See definition D105 in *Section 3.11, Canonical Ordering Behavior*.)

Floating (*diacritic, accent, mark*). (See *nonspacing mark*.)

Folding. An operation that maps similar characters to a common target, such as uppercasing or lowercasing a string. Folding operations are most often used to temporarily ignore certain distinctions between characters.

Font. A collection of glyphs used for the visual depiction of character data. A font is often associated with a set of parameters (for example, size, posture, weight, and serifness), which, when set to particular values, generate a collection of imagable glyphs.

Format Character. A character that is inherently invisible but that has an effect on the surrounding characters.

Format Code. Synonym for *format character.*

Formatted Text. (See *rich text*.)

FSS-UTF. Acronym for *File System Safe UCS Transformation Format*, published by the X/Open Company Ltd., and intended for the UNIX environment. Now known as *UTF-8*.

Fullwidth. Characters of East Asian character sets whose glyph image extends across the entire character display cell. In legacy character sets, fullwidth characters are normally encoded in two or three bytes. The Japanese term for fullwidth characters is *zenkaku*.

GCGID. Acronym for Graphic Character Global Identifier. These are listed in the IBM document *Character Data Representation Architecture, Level 1, Registry SC09-1391*.

General Category (GC). Partition of the characters into major classes such as letters, punctuation, and symbols, and further subclasses for each of the major classes. (See *Section 4.5, General Category—Normative.*)

Generative. Synonym for *productive.*

Glyph. (1) An abstract form that represents one or more glyph images. (2) A synonym for *glyph image*. In displaying Unicode character data, one or more glyphs may be selected to depict a particular character. These glyphs are selected by a rendering engine during composition and layout processing. (See also *character.*)

Glyph Code. A numeric code that refers to a glyph. Usually, the glyphs contained in a font are referenced by their glyph code. Glyph codes may be local to a particular font; that is, a different font containing the same glyphs may use different codes.

Glyph Identifier. Similar to a glyph code, a glyph identifier is a label used to refer to a glyph within a font. A font may employ both local and global glyph identifiers.

Glyph Image. The actual, concrete image of a glyph representation having been rasterized or otherwise imaged onto some display surface.

Glyph Metrics. A collection of properties that specify the relative size and positioning along with other features of a glyph.

Grapheme. (1) A minimally distinctive unit of writing in the context of a particular writing system. For example, ‹b› and ‹d› are distinct graphemes in English writing systems because there exist distinct words like big and dig. Conversely, ‹*a*› and ‹a› are not distinct graphemes

because no word is distinguished on the basis of these two different forms. (2) What a user thinks of as a character.

Grapheme Base. A character with the property Grapheme_Base, or any standard Korean syllable block. (See definition D58 in *Section 3.6, Combination*.)

Grapheme Cluster. A maximal character sequence consisting of a grapheme base followed by zero or more grapheme extenders or, alternatively, by the sequence <CR, LF>. (See definition D60 in *Section 3.6, Combination*.) A grapheme cluster represents a horizontally segmentable unit of text, consisting of some grapheme base (which may consist of a Korean syllable) together with any number of nonspacing marks applied to it.

Grapheme Extender. A character with the property Grapheme_Extend. (See definition D59 in *Section 3.6, Combination*.) Grapheme extender characters consist of all nonspacing marks, ZERO WIDTH JOINER, ZERO WIDTH NON-JOINER, and a small number of spacing marks.

Graphic Character. A character with the General Category of Letter (L), Combining Mark (M), Number (N), Punctuation (P), Symbol (S), or Space Separator (Zs). (See definition D50 in *Section 3.6, Combination*.)

Guillemet. Punctuation marks resembling small less-than and greater-than signs, used as quotation marks in French and other languages. (See "Language-Based Usage of Quotation Marks" in *Section 6.2, General Punctuation*.)

Halant. A preferred Hindi synonym for a *virama*. It literally means *killer*, referring to its function of *killing* the inherent vowel of a consonant letter. (See *virama*.)

Half-Consonant Form. In the Devanagari script and certain other scripts of the Brahmi family of Indic scripts, a dead consonant may be depicted in the so-called half-form. This form is composed of the distinctive part of a consonant letter symbol without its vertical stem. It may be used to create conjunct forms that follow a horizontal layout pattern. Also known as *half-form*.

Halfwidth. Characters of East Asian character sets whose glyph image occupies half of the character display cell. In legacy character sets, halfwidth characters are normally encoded in a single byte. The Japanese term for halfwidth characters is *hankaku*.

Han Characters. Ideographic characters of Chinese origin. (See *Section 12.1, Han*.)

Hangul. The name of the script used to write the Korean language.

Hangul Syllable. (1) Any of the 11,172 encoded characters of the Hangul Syllables character block, U+AC00..U+D7A3. Also called a *precomposed Hangul syllable* to clearly distinguish it from a Korean syllable block. (2) Loosely speaking, a *Korean syllable block*.

Hanja. The Korean name for Han characters; derived from the Chinese word *hànzì*.

Hankaku. (See *halfwidth.*)

Han Unification. The process of identifying Han characters that are in common among the writing systems of Chinese, Japanese, Korean, and Vietnamese.

Hànzì. The Mandarin Chinese name for Han characters.

Harakat. Marks that indicate vowels or other modifications of consonant letters in Arabic script.

Hasant. The Bangla name for *halant*. (See *virama*.)

Higher-Level Protocol. Any agreement on the interpretation of Unicode characters that extends beyond the scope of this standard. Note that such an agreement need not be formally announced in data; it may be implicit in the context. (See definition D16 in *Section 3.4, Characters and Encoding.*)

High-Surrogate Code Point. A Unicode code point in the range U+D800 to U+DBFF. (See definition D71 in *Section 3.8, Surrogates*.)

High-Surrogate Code Unit. A 16-bit code unit in the range $D800_{16}$ to $DBFF_{16}$, used in UTF-16 as the leading code unit of a surrogate pair. Also known as a *leading surrogate*. (See definition D72 in *Section 3.8, Surrogates.*)

Hiragana. One of two standard syllabaries associated with the Japanese writing system. Hiragana syllables are typically used in the representation of native Japanese words and grammatical particles.

HTML. HyperText Markup Language. A text description language related to SGML; it mixes text format markup with plain text content to describe formatted text. HTML is ubiquitous as the source language for Web pages on the Internet. Starting with HTML 4.0, the Unicode Standard functions as the reference character set for HTML content. (See also *SGML*.)

IANA. Acronym for Internet Assigned Numbers Authority.

ICU. Acronym for International Components for Unicode, an Open Source set of C/C++ and Java libraries for Unicode and software internationalization support. For information, see http://icu.sourceforge.net/

Ideograph. (1) Any symbol that primarily denotes an idea (or meaning) in contrast to a sound (or pronunciation)—for example, ☎ and 💣. (2) An English term commonly used to refer to Han characters, equivalent to the borrowings *hànzì*, *kanji*, and *hanja*.

Ideographic Property. Informative property of characters that are ideographs. (See *Section 4.10, Letters, Alphabetic, and Ideographic.*)

IICore. A subset of common-use CJK unified ideographs, defined as the fixed collection 370 IICore in ISO/IEC 10646. This subset contains 9,810 ideographs and is intended for common use in East Asian contexts, particularly for small devices that cannot support the full range of CJK unified ideographs encoded in the Unicode Standard.

Ill-Formed Code Unit Sequence. A code unit sequence that does not follow the specification of a Unicode encoding form. (See definition D84 in *Section 3.9, Unicode Encoding Forms.*)

In-Band. An in-band channel conveys information about text by embedding that information within the text itself, with special syntax to distinguish it. In-band information is

encoded in the same character set as the text, and is interspersed with and carried along with the text data. Examples are XML and HTML markup.

Independent Vowel. In Indic scripts, certain vowels are depicted using independent letter symbols that stand on their own. This is often true when a word starts with a vowel or a word consists of only a vowel.

Indic Digits. Forms of decimal digits used in various Indic scripts (for example, Devanagari: U+0966 ०, U+0967 १, U+0968 २, U+0969 ३). Arabic digits (and, eventually, European digits) derive historically from these forms.

Informative. Information in this standard that is not normative but that contributes to the correct use and implementation of the standard.

Inherent Vowel. In writing systems based on a script in the Brahmi family of Indic scripts, a consonant letter symbol normally has an inherent vowel, unless otherwise indicated. The phonetic value of this vowel differs among the various languages written with these writing systems. An inherent vowel is overridden either by indicating another vowel with an explicit vowel sign or by using *virama* to create a dead consonant.

Inner Caps. Mixed case format where an uppercase letter is in a position other than first in the word—for example, "G" in the Name "McGowan."

IPA. (1) The International Phonetic Alphabet. (2) The International Phonetic Association, which defines and maintains the International Phonetic Alphabet.

IRG. Acronym for Ideographic Rapporteur Group, a subgroup of ISO/IEC JTC1/SC2/WG2. (See *Appendix E, Han Unification History*.)

ISCII. Acronym for Indian Script Code for Information Interchange.

Jamo. The Korean name for a single letter of the Hangul script. Jamos are used to form Hangul syllables.

Joiner. An invisible character that affects the joining behavior of surrounding characters. (See *Section 8.2, Arabic*, and "Cursive Connection" in *Section 16.2, Layout Controls*.)

Jongseong. A sequence of one or more trailing consonants in Korean.

JTC1. The Joint Technical Committee 1 of the International Organization for Standardization and the International Electrotechnical Commission responsible for information technology standardization.

Jungseong. A sequence of one or more vowels in Korean.

Kana. The name of a primarily syllabic script used by the Japanese writing system. It comes in two forms: *hiragana* and *katakana*. The former is used to write particles, grammatical affixes, and words that have no *kanji* form; the latter is used primarily to write foreign words.

Kanji. The Japanese name for Han characters; derived from the Chinese word *hànzì*. Also romanized as *kanzi*.

Katakana. One of two standard syllabaries associated with the Japanese writing system. Katakana syllables are typically used in representation of borrowed vocabulary (other than that of Chinese origin), sound-symbolic interjections, or phonetic representation of "difficult" kanji characters in Japanese.

Kerning. (1) Changing the space between certain pairs of letters to improve the appearance of the text. (2) The process of mapping from pairs of glyphs to a positioning offset used to change the space between letters.

Korean Syllable Block. A sequence of Korean jamos, consisting of one or more leading consonants followed by one or more vowels followed by zero or more trailing consonants, or any canonically equivalent sequence including a precomposed Hangul syllable. In regular expression notation: *L L* V V* T**. Also called a *standard Korean syllable block*. (See *Section 3.12, Conjoining Jamo Behavior*.)

LDML. (See *Locale Data Markup Language*.)

Leading Consonant. (1) In Korean, a jamo character with the Hangul_Syllable_Type property value Leading_Jamo (in the range U+1100..U+1159 or U+115F HANGUL CHOSEONG FILLER). Abbreviated as *L*. (See definition D107 in *Section 3.12, Conjoining Jamo Behavior*.) (2) Any initial consonant in a syllable.

Leading Surrogate. Synonym for *high-surrogate code unit*.

Letter. (1) An element of an alphabet. In a broad sense, it includes elements of syllabaries and ideographs. (2) Informative property of characters that are used to write words.

Ligature. A glyph representing a combination of two or more characters. In the Latin script, there are only a few in modern use, such as the ligatures between "f" and "i" (= fi) or "f" and "l" (= fl). Other scripts make use of many ligatures, depending on the font and style.

Little-endian. A computer architecture that stores multiple-byte numerical values with the least significant byte (LSB) values first.

Locale Data Markup Language. The XML specification for the exchange of locale data, defined by Unicode Technical Standard #35, "Locale Data Markup Language (LDML)." (See also *Common Locale Data Repository.*)

Logical Order. The order in which text is typed on a keyboard. For the most part, logical order corresponds to phonetic order. (See *Section 2.2, Unicode Design Principles*.)

Logical Store. Memory representation.

Logosyllabary. A writing system in which the units are used primarily to write words and/or morphemes of words, with some subsidiary usage to represent just syllabic sounds. The best example is the Han script.

Lowercase. (See *case*.)

Low-Surrogate Code Point. A Unicode code point in the range U+DC00 to U+DFFF. (See definition D73 in *Section 3.8, Surrogates*.)

Low-Surrogate Code Unit. A 16-bit code unit in the range $DC00_{16}$ to $DFFF_{16}$, used in UTF-16 as the trailing code unit of a surrogate pair. Also known as a *trailing surrogate*. (See definition D74 in *Section 3.8, Surrogates*.)

LSB. Acronym for *least significant byte*.

LZW. Acronym for *Lempel-Ziv-Welch*, a standard algorithm widely used for compression of data.

Majuscule. Synonym for *uppercase*. (See *case*.)

Mathematical Property. Informative property of characters that are used as operators in mathematical formulae.

Matra. A dependent vowel in an Indic script. It is the name for vowel letters that follow consonant letters in logical order. A matra often has a completely different letterform from that for the same phonological vowel used as an independent letter.

MBCS. Acronym for *multibyte character set*.

MIME. Multipurpose Internet Mail Extensions. MIME is a standard that allows the embedding of arbitrary documents and other binary data of known types (images, sound, video, and so on) into e-mail handled by ordinary Internet electronic mail interchange protocols.

Minuscule. Synonym for *lowercase*. (See *case*.)

Mirrored Property. The property of characters whose images are mirrored horizontally in text that is laid out from right to left (versus from left to right). (See *Section 4.7, Bidi Mirrored—Normative*.)

Missing Glyph. (See *replacement glyph*.)

Modifier Letter. A character with the Lm General Category in the Unicode Character Database. Modifier letters, which look like letters or punctuation, modify the pronunciation of other letters (similar to diacritics). (See *Section 7.8, Modifier Letters*.)

Monotonic. Modern Greek written with the basic accent, the *tonos*.

MSB. Acronym for *most significant byte*.

Multibyte Character Set. A character set encoded with a variable number of bytes per character, often abbreviated as MBCS. Many large character sets have been defined as MBCS so as to keep strict compatibility with the ASCII subset and/or ISO/IEC 2022.

Named Unicode Algorithm. A Unicode algorithm that is specified in the Unicode Standard or in other standards published by the Unicode Consortium and that is given an explicit name for ease of reference. (See definition D18 in *Section 3.4, Characters and Encoding*. See also *Table 3-1,* "Named Unicode Algorithms," for a list of named Unicode algorithms.)

Namespace. (1) A set of names, no two of which are identical. (2) A set of names together with name matching rules, so that all names are distinct under the matching rules. (See definition D6 in *Section 3.3, Semantics*.) Character names are distinct if they do not match under the name matching rules in effect for the standard.

Nekudot. Marks that indicate vowels or other modifications of consonantal letters in Hebrew.

Neutral Character. A character that can be written either right to left or left to right, depending on context. (See Unicode Standard Annex #9, "The Bidirectional Algorithm.")

NFC. (See *Normalization Form C.*)

NFD. (See *Normalization Form D.*)

NFKC. (See *Normalization Form KC.*)

NFKD. (See *Normalization Form KD.*)

Noncharacter. A code point that is permanently reserved for internal use and that should never be interchanged. Noncharacters consist of the values U+*n*FFFE and U+*n*FFFF (where *n* is from 0 to 10_{16}), and the values U+FDD0..U+FDEF.

Non-joiner. An invisible character that affects the joining behavior of surrounding characters. (See *Section 8.2, Arabic*, and "Cursive Connection" in *Section 16.2, Layout Controls*.)

Non-overridable. A characteristic of a Unicode character property that cannot be changed by a higher-level protocol.

Nonspacing Diacritic. A diacritic that is a nonspacing mark.

Nonspacing Mark. A combining character with the General Category of Nonspacing Mark (Mn) or Enclosing Mark (Me). (See definition D53 in *Section 3.6, Combination*.) The position of a nonspacing mark in presentation depends on its base character. It generally does not consume space along the visual baseline in and of itself. (See also *combining character.*)

Normalization. A process of removing alternate representations of equivalent sequences from textual data, to convert the data into a form that can be binary-compared for equivalence. In the Unicode Standard, normalization refers specifically to processing to ensure that canonical-equivalent (and/or compatibility-equivalent) strings have unique representations. For more information, see "Equivalent Sequences" in *Section 2.2, Unicode Design Principles*, and Unicode Standard Annex #15, "Unicode Normalization Forms."

Normalization Form. One of the four Unicode normalization forms defined in Unicode Standard Annex #15, "Unicode Normalization Forms"—namely, NFC, NFD, NFKC, and NFKD.

Normalization Form C (NFC). The normalization form that results from the canonical decomposition of a Unicode string, followed by the replacement of all decomposed sequences by primary composites where possible.

Normalization Form D (NFD). The normalization form that results from the canonical decomposition of a Unicode string.

Normalization Form KC (NFKC). The normalization form that results from the compatibility decomposition of a Unicode string, followed by the replacement of all decomposed sequences by primary composites where possible.

Normalization Form KD (NFKD). The normalization form that results from the compatibility decomposition of a Unicode string.

Normative. Required for conformance with the Unicode Standard.

NSM. Acronym for *nonspacing mark*.

Numeric Value Property. A property of characters used to represent numbers. (See *Section 4.6, Numeric Value—Normative.*)

Obsolete. Applies to a character that is no longer in current use, but that has been used historically. Whether a character is obsolete depends on context. For example, the Cyrillic letter *big yus* is obsolete for Russian, but is used in modern Bulgarian. (Not the same as *deprecated*.)

Octet. An ordered sequence of eight bits considered as a unit. The Unicode Standard follows current industry practice in referring to an octet as a *byte*. (See *byte*.)

Out-of-Band. An out-of-band channel conveys additional information about text in such a way that the textual content, as encoded, is completely untouched and unmodified. This is typically done by separate data structures that point into the text.

Overridable. A characteristic of a Unicode character property that may be changed by a higher-level protocol to create desired implementation effects.

Paragraph Direction. The default direction (*left* or *right*) of the text of a paragraph. This direction does not change the display order of characters within an Arabic or English word. However, it *does* change the display order of adjacent Arabic and English words and the display order of neutral characters, such as punctuation and spaces. For more details, see Unicode Standard Annex #9, "The Bidirectional Algorithm," especially definitions BD2–BD5.

Paragraph Embedding Level. The embedding level that determines the default bidirectional orientation of the text in that paragraph.

Phoneme. A minimally distinct sound in the context of a particular spoken language. For example, in American English, /p/ and /b/ are distinct phonemes because *pat* and *bat* are distinct; however, the two different sounds of /t/ in *tick* and *stick* are not distinct in English, even though they are distinct in other languages such as Thai.

Pinyin. Standard system for the romanization of Chinese on the basis of Mandarin pronunciation.

Pivot Conversion. The use of a third character encoding to serve as an intermediate step in the conversion between two other character encodings. The Unicode Standard is widely used to support pivot conversion, as its character repertoire is a superset of most other coded character sets.

Plain Text. Computer-encoded text that consists *only* of a sequence of code points from a given standard, with no other formatting or structural information. Plain text interchange is commonly used between computer systems that do not share higher-level protocols. (See also *rich text*.)

Plane. A range of 65,536 (10000_{16}) contiguous Unicode code points, where the first code point is an integer multiple of 65,636 (10000_{16}). Planes are numbered from 0 to 16, with the number being the first code point of the plane divided by 65,536. Thus Plane 0 is U+0000..U+FFFF, Plane 1 is U+**1**0000..U+**1**FFFF, ..., and Plane 16 (10_{16}) is U+**10**0000..**10**FFFF. (Note that ISO/IEC 10646 uses hexadecimal notation for the plane numbers—for example, Plane B instead of Plane 11.) (See *Basic Multilingual Plane* and *supplementary planes*.)

Points. (1) The nonspacing vowels and other signs of written Hebrew. (2) A unit of measurement in typography.

Polytonic. Ancient Greek written with several contrastive accents.

Precomposed Character. (See *decomposable character*.)

Presentation Form. A ligature or variant glyph that has been encoded as a character for compatibility. (See also *compatibility character* (1).)

Primary Composite. A character that has a canonical decomposition mapping in the Unicode Character Database (or is a canonical Hangul decomposition) but that is not in the Composition Exclusion Table. (See Unicode Standard Annex #15, "Unicode Normalization Forms.")

Private Use. Refers to designated code points in the Unicode Standard or other character encoding standards whose interpretations are not specified in those standards and whose use may be determined by private agreement among cooperating users.

Private Use Area (PUA). Any one of the three blocks of private-use code points in the Unicode Standard.

Private-Use Code Point. Code points in the ranges U+E000..U+F8FF, U+F0000..U+FFFFD, and U+100000..U+10FFFD. (See definition D49 in *Section 3.5, Properties*.) These code points are designated in the Unicode Standard for private use.

Productive. Said of a feature or rule that can be employed in novel combinations or circumstances, rather than being restricted to a fixed list. In the Unicode Standard, combining marks—particularly the accents—are productive. In contrast, variation selectors are deliberately not productive. Also known as *generative*.

Property. (See *character properties*.)

Property Alias. A unique identifier for a particular Unicode character property. (See definition D47 in *Section 3.5, Properties*.)

Property Value Alias. A unique identifier for a particular enumerated value for a particular Unicode character property. (See definition D48 in *Section 3.5, Properties*.)

Provisional. A property or feature that is unapproved and tentative and that may be incomplete or otherwise not in a usable state.

PUA. Acronym for *Private Use Area*.

Puḷḷi. The Tamil name for *virama*. (See *virama*.)

Radical. A structural component of a Han character conventionally used for indexing. The traditional number of such radicals is 214.

Rendering. (1) The process of selecting and laying out glyphs for the purpose of depicting characters. (2) The process of making glyphs visible on a display device.

Repertoire. (See *character repertoire*.)

Replacement Character. A character used as a substitute for an uninterpretable character from another encoding. The Unicode Standard uses U+FFFD REPLACEMENT CHARACTER for this function.

Replacement Glyph. A glyph used to render a character that cannot be rendered with the correct appearance in a particular font. It often is shown as an open □ or black ■ rectangle. Also known as a *missing glyph*. (See *Section 5.3, Unknown and Missing Characters*.)

Reserved Code Point. Any code point of the Unicode Standard that is reserved for future assignment. Also known as an *unassigned code point*. (See definition D15 in *Section 3.4, Characters and Encoding*.)

Rich Text. Also known as *styled text*. The result of adding information to plain text. Examples of information that can be added include font data, color, formatting information, phonetic annotations, interlinear text, and so on. The Unicode Standard does not address the representation of rich text. It is expected that systems and applications will implement proprietary forms of rich text. Some public forms of rich text are available (for example, ODA, HTML, and SGML). When everything except primary content is removed from rich text, only plain text should remain.

Row. A range of 256 contiguous Unicode code points, where the first code point is an integer multiple of 256. Two code points are in the same row if they share all but the last two hexadecimal digits. (See *plane*.)

SAM. Acronym for Syriac abbreviation mark.

SBCS. Acronym for *single-byte character set*. Any one-byte character encoding. This term is generally used in contrast with DBCS and/or MBCS.

Scalar Value. (See *Unicode scalar value*.)

Script. A collection of letters and other written signs used to represent textual information in one or more writing systems. For example, Russian is written with a subset of the Cyrillic script; Ukranian is written with a different subset. The Japanese writing system uses several scripts.

SCSU. Acronym for Standard Compression Scheme for Unicode. See Unicode Technical Standard #6, "A Standard Compression Scheme for Unicode."

SGML. Standard Generalized Markup Language. A standard framework, defined in ISO 8879, for defining particular text markup languages. The SGML framework allows for mixing structural tags that describe format with the plain text content of documents, so that

fancy text can be fully described in a plain text stream of data. (See also *HTML*, *XML*, and *rich text*.)

Shaping Characters. Characters that assume different glyphic forms depending on the context.

Shift-JIS. A shifted encoding of the Japanese character encoding standard, JIS X 0208, that is widely deployed in PCs.

Signature. An optional code sequence at the beginning of a stream of coded characters that identifies the *character encoding scheme* used for the following text. (See *Unicode signature*.)

Sinogram. Chinese character. (See *ideograph*.)

SJIS. Acronym for *Shift-JIS*.

Small Letter. Synonym for *lowercase letter.* (See *case*.)

Sorting. (See *collation.*)

Spacing Mark. A combining character that is not a nonspacing mark. (See definition D55 in *Section 3.6, Combination*.) (See *nonspacing mark*.)

Standard Korean Syllable Block. (See *Korean syllable block*.)

Static Form. (See *decomposable character.*)

Styled Text. (See *rich text.*)

Subtending Mark. A format character whose graphic form extends under a sequence of following characters—for example, U+0600 ARABIC NUMBER SIGN.

Supplementary Character. A Unicode encoded character having a supplementary code point.

Supplementary Code Point. A Unicode code point between U+10000 and U+10FFFF.

Supplementary Planes. Planes 1 through 16, consisting of the supplementary code points.

Surrogate Character. A misnomer. It would be an encoded character having a surrogate code point, which is impossible. Do not use this term.

Surrogate Code Point. A Unicode code point in the range U+D800..U+DFFF. Reserved for use by UTF-16, where a pair of surrogate code units (a high surrogate followed by a low surrogate) "stand in" for a supplementary code point.

Surrogate Pair. A representation for a single abstract character that consists of a sequence of two 16-bit code units, where the first value of the pair is a *high-surrogate code unit* and the second value is a *low-surrogate code unit*. (See definition D75 in *Section 3.8, Surrogates*.)

Syllabary. A type of writing system in which each symbol typically represents both a consonant and a vowel, or in some instances more than one consonant and a vowel.

Syllable. (1) An element of a syllabary. (2) A basic unit of articulation that corresponds to a pulmonary pulse.

Syllable Block. A sequence of Korean characters that should be grouped into a single square cell for display. (See *Section 3.12, Conjoining Jamo Behavior*.)

Symmetric Swapping. The process of rendering a character with a mirrored glyph when its resolved directionality is right-to-left in a bidirectional context. (See *mirrored property* and Unicode Standard Annex #9, "The Bidirectional Algorithm.")

Tagging. The association of attributes of text with a point or range of the primary text. The value of a particular tag is not generally considered to be a part of the "content" of the text. A typical example of tagging is to mark the language or the font for a portion of text.

Tailorable. A characteristic of an algorithm for which a higher-level protocol may specify different results than those specified in the algorithm. A tailorable algorithm without actual tailoring is also known as a default algorithm, and the results of an algorithm without tailoring are known as the default results.

TES. Acronym for *transfer encoding syntax*.

TEX. Computer language designed for use in typesetting—in particular, for typesetting math and other technical material. (According to Knuth, TEX rhymes with the word *blecchhh*.)

Text Element. A minimum unit of text in relation to a particular text process, in the context of a given writing system. In general, the mapping between text elements and code points is many-to-many. (See *Chapter 2, General Structure*.)

Titlecase. Uppercased initial letter followed by lowercase letters in words. A casing convention often used in titles, headers, and entries, as exemplified in this glossary.

Tonal Sandhi. A phonological process whereby the tone associated with one syllable in a tonal language influences the realization of a tone associated with a neighboring syllable.

Tone Mark. A diacritic or nonspacing mark that represents a phonemic tone. Tone languages are common in Southeast Asia and Africa. Because tones always accompany vowels (the syllabic nucleus), they are most frequently written using functionally independent marks attached to a vowel symbol. However, some writing systems such as Thai place tone marks on consonant symbols; Chinese does not use tone marks (except when it is written phonemically).

Tonemic. Refers to the underlying, distinctive units of a tonal system in a language. Tones of a tonal language are often referred to by numbers ("tone 1," "tone 2," and so on), and each tone has an idealized, specific tone level or contour that is considered to be its tonemic value. The term was created by analogy with *phonemic*.

Tonetic. Refers to the surface, actual pitch realization of tones in a tonal system. Tonetic values are what can be directly measured by tracking pitch contours in actual speech recordings. The term was created by analogy with *phonetic*.

Tonos. The basic accent in modern Greek, having the form of an acute accent.

Trailing Consonant. (1) In Korean, a jamo character with the Hangul_Syllable_Type property value Vowel_Jamo (in the range U+11A8..U+11F9). Abbreviated as *T*. (See definition D113 in *Section 3.12, Conjoining Jamo Behavior*.) (2) Any final consonant in a syllable.

Trailing Surrogate. Synonym for *low-surrogate code unit*.

Transcoding. Conversion of character data between different character sets.

Transfer Encoding Syntax. A reversible transformation applied to text and other data to allow it to be transmitted—for example, Base64, uuencode.

Transformation Format. A mapping from a coded character sequence to a unique sequence of code units (typically bytes).

Triangulation. (See *pivot conversion*.)

Typographic Interaction. Graphical application of one nonspacing mark in a position relative to a grapheme base that is already occupied by another nonspacing mark, so that some rendering adjustment must be done (such as default stacking or side-by-side placement) to avoid illegible overprinting or crashing of glyphs. (See definition D106 in *Section 3.11, Canonical Ordering Behavior*.)

UAX. Acronym for *Unicode Standard Annex*.

UCA. Acronym for *Unicode Collation Algorithm*.

UCD. Acronym for *Unicode Character Database*. (See *Section 4.1, Unicode Character Database*.)

UCS. Acronym for Universal Character Set, which is specified by International Standard ISO/IEC 10646, which is equivalent in repertoire to the Unicode Standard.

UCS-2. ISO/IEC 10646 encoding form: Universal Character Set coded in 2 octets. (See *Appendix C, Relationship to ISO/IEC 10646*.)

UCS-4. ISO/IEC 10646 encoding form: Universal Character Set coded in 4 octets. (See *Appendix C, Relationship to ISO/IEC 10646*.)

Umlaut. Two horizontal dots over a letter, as in German *Köpfe*. The umlaut is not distinguished from the *diaeresis* in the Unicode character encoding. (See *diaeresis*.)

Unassigned. Code points that either are reserved for future use or are never to be used.

Unassigned Character. Synonym for *not assigned to an abstract character*. This refers to surrogate code points, noncharacters, and reserved code points. (See *Section 2.4, Code Points and Characters*.)

Unassigned Code Point. (See *undesignated code point*.)

Undesignated Code Point. Synonym for *reserved code point*. These code points are reserved for future assignment and have no other designated normative function in the standard. (See *Section 2.4, Code Points and Characters*.)

Unicameral. A script that has no *case* distinctions. Most often used in the context of European alphabets.

Unicode. The universal character encoding, maintained by the Unicode Consortium. This encoding standard provides the basis for processing, storage, and interchange of text data in any language in all modern software and information technology protocols.

Unicode Algorithm. The logical description of a process used to achieve a specified result involving Unicode characters. (See definition D17 in *Section 3.4, Characters and Encoding*.)

Unicode Collation Algorithm. Tailorable text comparison mechanism used for searching, sorting, and matching Unicode strings. See Unicode Technical Standard #10, "Unicode Collation Algorithm."

Unicode Character Database. A collection of files providing normative and informative Unicode character properties and mappings. (See *Chapter 4, Character Properties*, and the UnicodeCharacterDatabase.html documentation file.)

Unicode Encoding Form. A character encoding form that assigns each Unicode scalar value to a unique code unit sequence. The Unicode Standard defines three Unicode encoding forms: UTF-8, UTF-16, and UTF-32. (See definition D79 in *Section 3.9, Unicode Encoding Forms*.)

Unicode Encoding Scheme. A specified byte serialization for a Unicode encoding form, including the specification of the handling of a *byte order mark* (BOM), if allowed. (See definition D94 in *Section 3.10, Unicode Encoding Schemes*.)

Unicode Scalar Value. Any Unicode code point except high-surrogate and low-surrogate code points. In other words, the ranges of integers 0 to $D7FF_{16}$ and $E000_{16}$ to $10FFFF_{16}$, inclusive. (See definition D76 in *Section 3.9, Unicode Encoding Forms*.)

Unicode Signature. An implicit marker to identify a file as containing Unicode text in a particular encoding form. An initial *byte order mark* (BOM) may be used as a Unicode signature.

Unicode Standard Annex. An integral part of the Unicode Standard published as a separate document.

Unicode String. A code unit sequence containing code units of a particular Unicode encoding form. (See definition D80 in *Section 3.9, Unicode Encoding Forms*.)

Unicode Technical Note. Informative publication containing information of possible interest concerning the Unicode Standard or related topics.

Unicode Technical Report. Formally approved Unicode Consortium publication containing informative technical analysis of a topic related to the Unicode Standard.

Unicode Technical Standard. Formally approved specification published by the Unicode Consortium that is related to, but not part of, the Unicode Standard.

Unicode Transformation Format. An ambiguous synonym for either *Unicode encoding form* or *Unicode encoding scheme*. The latter terms are now preferred.

Unification. The process of identifying characters that are in common among writing systems.

UPA. Acronym for Uralic Phonetic Alphabet.

Uppercase. (See *case*.)

URO. Acronym for Unified Repertoire and Ordering, the original set of CJK unified ideographs used in the Unicode Standard.

User-Defined Character. (See *EUDC*.)

User-Perceived Character. What everyone thinks of as a character in their script.

UTF. Acronym for *Unicode* (or *UCS*) *Transformation Format*.

UTF-2. Obsolete name for *UTF-8*.

UTF-7. Unicode (or UCS) Transformation Format, 7-bit encoding form, specified by *RFC-2152*.

UTF-8. (1) The UTF-8 encoding form. (2) The UTF-8 encoding scheme. (3) "UCS Transformation Format 8," defined in Annex D of ISO/IEC 10646:2003; technically equivalent to the definitions in the Unicode Standard.

UTF-8 Encoding Form. The Unicode encoding form that assigns each Unicode scalar value to an unsigned byte sequence of one to four bytes in length, as specified in *Table 3-6.* (See definition D92 in *Section 3.9, Unicode Encoding Forms.*)

UTF-8 Encoding Scheme. The Unicode encoding scheme that serializes a UTF-8 code unit sequence in exactly the same order as the code unit sequence itself. (See definition D95 in *Section 3.10, Unicode Encoding Schemes*.)

UTF-16. (1) The UTF-16 encoding form. (2) The UTF-16 encoding scheme. (3) "Transformation format for 16 planes of Group 00," defined in Annex C of ISO/IEC 10646:2003; technically equivalent to the definitions in the Unicode Standard.

UTF-16 Encoding Form. The Unicode encoding form that assigns each Unicode scalar value in the ranges U+0000..U+D7FF and U+E000..U+FFFF to a single unsigned 16-bit code unit with the same numeric value as the Unicode scalar value, and that assigns each Unicode scalar value in the range U+10000..U+10FFFF to a surrogate pair, according to *Table 3-5,* "UTF-16 Bit Distribution." (See definition D91 in *Section 3.9, Unicode Encoding Forms*.)

UTF-16 Encoding Scheme. The UTF-16 encoding scheme that serializes a UTF-16 code unit sequence as a byte sequence in either big-endian or little-endian formats. (See definition D98 in *Section 3.10, Unicode Encoding Schemes.*)

UTF-16BE. The Unicode encoding scheme that serializes a UTF-16 code unit sequence as a byte sequence in big-endian format. (See definition D96 in *Section 3.10, Unicode Encoding Schemes*.)

UTF-16LE. The Unicode encoding scheme that serializes a UTF-16 code unit sequence as a byte sequence in little-endian format. (See definition D97 in *Section 3.10, Unicode Encoding Schemes*.)

UTF-32. (1) The UTF-32 encoding form. (2) The UTF-32 encoding scheme.

UTF-32 Encoding Form. The Unicode encoding form that assigns each Unicode scalar value to a single unsigned 32-bit code unit with the same numeric value as the Unicode scalar value. (See definition D90 in *Section 3.9, Unicode Encoding Forms.*)

UTF-32 Encoding Scheme. The Unicode encoding scheme that serializes a UTF-32 code unit sequence as a byte sequence in either big-endian or little-endian formats. (See definition D101 in *Section 3.10, Unicode Encoding Schemes*.)

UTF-32BE. The Unicode encoding scheme that serializes a UTF-32 code unit sequence as a byte sequence in big-endian format. (See definition D99 in *Section 3.10, Unicode Encoding Schemes*.)

UTF-32LE. The Unicode encoding scheme that serializes a UTF-32 code unit sequence as a byte sequence in little-endian format. (See definition D100 in *Section 3.10, Unicode Encoding Schemes*.)

UTN. Acronym for *Unicode Technical Note*.

UTR. Acronym for *Unicode Technical Report*.

UTS. Acronym for *Unicode Technical Standard*.

Virama. From Sanskrit *virāma*. The name of a sign used in many Indic and other Brahmi-derived scripts to suppress the inherent vowel of the consonant to which it is applied, thereby generating a *dead consonant*. (See *Section 9.1, Devanagari*.) The sign varies in shape from script to script and may be known by other names in various languages. For example, in Hindi it is known as *hal* or *halant*, in Bangla it is called *hasant*, and in Tamil it is called *puḷḷi*.

Visual Ambiguity. A situation arising from two characters (or sequences of characters) being rendered indistinguishably.

Visual Order. Characters ordered as they are presented for reading. (Contrast with *logical order*.)

Vocalization. Marks placed above, below, or within consonants to indicate vowels or other aspects of pronunciation. A feature of Middle Eastern scripts.

Vowel. In Korean, a jamo character with the Hangul_Syllable_Type property value Vowel_Jamo (in the range U+1161..U+11A2 or U+1160 HANGUL JUNGSEONG FILLER). Abbreviated as *V*. (See definition D110 in *Section 3.12, Conjoining Jamo Behavior*.)

Vowel Mark. In many scripts, a mark used to indicate a vowel or vowel quality.

W3C. Acronym for World Wide Web Consortium.

wchar_t. The ANSI C defined *wide character* type, usually implemented as either 16 or 32 bits. ANSI specifies that `wchar_t` be an integral type and that the C-language source character set be mappable by simple extension (zero- or sign-extension).

Writing Direction. The direction or orientation of writing characters within lines of text in a writing system. Three directions are common in modern writing systems: left to right, right to left, and top to bottom.

Writing System. A set of rules for using one or more scripts to write a particular language. Examples include the American English writing system, the British English writing system, the French writing system, and the Japanese writing system.

XML. eXtensible Markup Language. A subset of SGML constituting a particular text markup language for interchange of structured data. The Unicode Standard is the reference character set for XML content. (See also *SGML* and *rich text*.) XML is a trademark of the World Wide Web Consortium.

Y-variant. Two CJK unified ideographs with identical semantics and non-unifiable shapes—for example, U+732B and U+8C93. (See *Z-variant*.)

Z-variant. Two CJK unified ideographs with identical semantics and unifiable shapes—for example, U+8AAA and U+8AAC. (See *Y-variant*.)

Zenkaku. (See *fullwidth*.)

Zero Width. Characteristic of some spaces or format control characters that do not advance text along the horizontal baseline. (See *nonspacing mark*.)

