1.0 Introduction

The Unicode character encoding standard is a fixed-width, uniform text and character encoding scheme. It includes characters from the world's scripts, as well as technical symbols in common use. The Unicode standard is modeled on the ASCII character set. Since ASCII's 7-bit character size is inadequate to handle multilingual text, the Unicode Consortium adopted a 16-bit architecture which extends the benefits of ASCII to multilingual text. Unicode characters are consistently 16 bits wide, regardless of language, so no escape sequence or control code is required to specify any character in any language. Unicode character encoding treats symbols, alphabetic characters, and ideographic characters identically, so that they can be used simultaneously and with equal facility. Computer programs that use Unicode character encoding to represent characters but do not display or print text can (for the most part) remain unaltered when new scripts or characters are introduced.

1.1 Background

The primary goal of the Unicode project was to remedy serious problems common to most multilingual computer programs: overloading of the font mechanism when encoding characters, and use of multiple, inconsistent character codes caused by conflicting national character standards. Few national standards allowed for special purpose characters, such as proprietary or typographical characters. The ASCII character set and its extensions, although widely used and accepted as standard in most computing systems, are limited to 256 characters. ASCII is therefore inadequate in an increasingly complex global computing environment.

The groups most affected by the lack of a consistent international character standard are the publishers of scientific and mathematical software, newspaper and book publishers, bibliographic information services, and academic researchers.

Designers of the Unicode standard envisioned a uniform method of character identification that would be more efficient and flexible than current encoding systems. Their system would be complete enough to satisfy the needs of technical and multilingual computing, as well as text publishing. Their main goals were to eliminate the special case systems and complex application codes currently in use in many character encoding standards, and to make a larger range of characters

available in order to meet the requirements of professional quality typesetting and desktop publishing internationally.

Research and analysis revealed that an efficient character code standard would meet the following requirements:

- *Completeness.* The coded character set would be large enough to encompass all characters that were likely to be used in general text interchange.
- *Efficiency.* Plain text, composed of a sequence of fixed-width characters, provided an extremely useful model because it was simple to parse: software would not have to maintain state, look for special escape sequences, or search forward or backward through text to identify characters.
- *Uniformity.* For efficient sorting, searching, display, and editing of text, a fixed character code size would be preferable to the more complex run-length encoding schemes in current use. Although a wide character code is not always necessary, particularly in the case of scripts that contain a limited number of characters, the many benefits of a uniform character width outweigh the argument in favor of codespace economy. Text compression should not be defined by the character code standard; rather, it should be independent of the character code standard.

1.2 Conformance

There is a set of unambiguous criteria to which a Unicode-conformant implementation must adhere, to ensure that it can interoperate with other conformant implementations.

An application may be considered to conform to the Unicode standard if it makes use of independent fixed-width 16-bit characters and uses Unicode code points to represent Unicode-defined characters. Code conversion from other standards to the Unicode standard will be considered conformant if the matching table produces accurate conversions in both directions. Explicit rules for conformance are found in Section 2.6 of this volume. Information on handling missing characters is found in Appendix C.

1.3 Coverage

This first edition of The Unicode Standard contains over 28,000 characters from the world's scripts. These characters are more than sufficient for modern communication, as well as classical forms of languages such as Greek, Hebrew, Latin, Pali, Sanskrit and literary Chinese. Over 20,000 unique characters defined by national and industry standards of China, Japan, Korea, and Taiwan are included. The Unicode standard also includes math operators and technical symbols, geometric shapes and dingbats. Figure 1-1 shows the scripts included in version 1.0 of the Unicode standard and the code range for each.

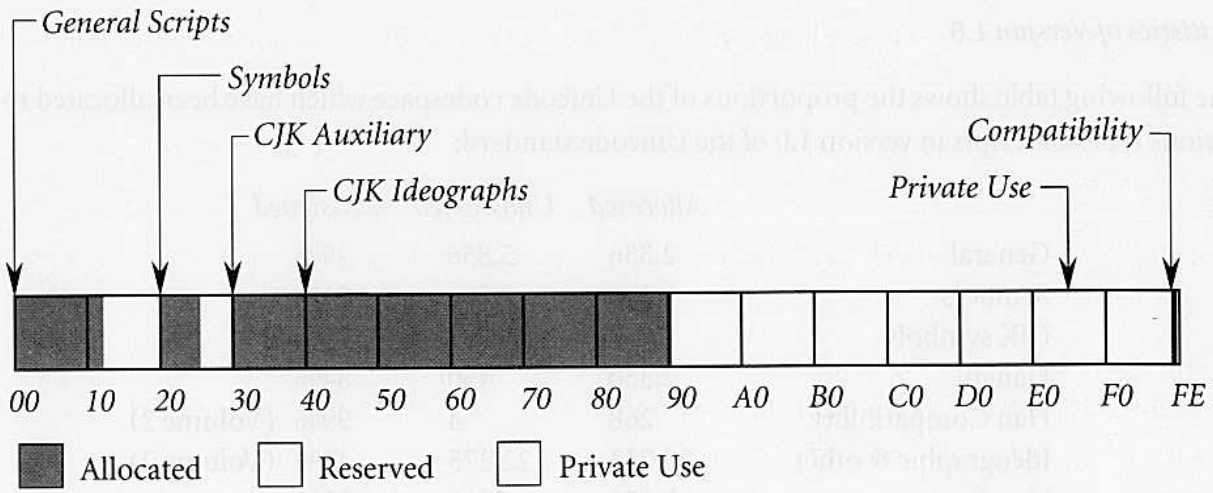

Figure 1-1. Table of Zones and Code Ranges

To define the content of the Unicode standard, the Unicode Technical Committee relied primarily on existing standards. Many characters have been included solely because they are part of an existing standard in widespread use, despite the fact that they violate the general principles of the Unicode standard in some instances.

The Unicode standard includes the character content of all major International Standards approved and published before December 31, 1990, in particular, the ISO *International Register of Character Sets*, and the ISO 6937 and ISO 8859 families of standards, as well as ISO 8879 (SGML). Characters from other standards have also been included, specifically, from bibliographic standards used in libraries (ANSI Z39.47-1985 [Roman], and ANSI Z39.64-1990 [East Asian]), and from important national standards (ISCII 1988 [India], GB 2312-1980 [China], JIS X 0208-1990 and JIS X 0212-1990 [Japan], and CNS 11643-1986 [Taiwan]). Also included are characters from certain draft standards (such as ISO DIS 6861.2, Glagolitic, Old Cyrillic and Romanian Cyrillic for bibliographic information interchange), and from various industry standards in common use (such as code pages and character sets from Adobe, Apple, IBM, Lotus, Microsoft, WordPerfect, Xerox and others).

Another source of characters is from numerous papers and national bodies' contributions to the ISO SC2/WG2 committee on character encoding.

The Unicode standard version 1.0 does not encode rare, obsolete, idiosyncratic, personal, novel, rarely exchanged or private-use characters, nor does it encode logos or graphics. Artificial entities, whose sole function is to serve transiently in the input of text, are also excluded from the Unicode standard. Graphologies unrelated to text, for example, musical and dance notations, are outside the scope of the Unicode standard. Braille symbols were not encoded, since Braille is an alternative way to present text (it can be considered a font variant).

Statistics of Version 1.0

The following table shows the proportions of the Unicode codespace which have been allocated to various types of scripts in version 1.0 of the Unicode standard:

	Allocated	*Unassigned*	*% Assigned*	
General	2,336	5,856	29%	
Symbols	1,290	2,806	31%	
CJK symbols	763	261	75%	
Hangul	2,350	450	84%	
Han Compatibility	268	4	99%	(Volume 2)
Ideographic & other	20,733	22,275	48%	(Volume 2)
User Space	5,632	N/A	N/A	
Compatibility Zone	362	133	73%	
Special	1	13		
FEFF	1	0		
FFFE, FFFF	N/A	2		
Totals	28,706 (assigned)			
	+ 5,632(private use)			
	= 34,338(allocated) 52%			

With over 30,000 unallocated character positions, the Unicode character encoding provides sufficient space for forseeable future expansion.

Future Plans

Less common and archaic scripts will be added to future versions of the Unicode Standard. Scripts of this type were not included in the initial release because of the difficulty of evaluating their content. For many of these scripts, extensive research will be necessary to produce an agreed-upon encoding. The five scripts that are included here in draft form (Ethiopian, Burmese, Khmer, Sinhala, and Mongolian) will be added to the Unicode standard when reliable information has been obtained. (See Appendix E.) Other scripts that are being considered for possible addition to the Unicode standard are:

- *Inuktitut/Cree Syllabary.* The Department of Communications, Canada, is pursuing standardization of the several variant syllabaries and computer encodings now in use for Cree and/or Inuktitut.
- *Egyptian Hieroglyphics.* A uniform standard for computer encoding exists and is being investigated.
- *Korean Hangul Syllables.* There may also be a number of additional Korean Hangul syllables added.

Interest has also been expressed in including Cuneiform, the Cherokee syllabary, the Maldivian and Syriac scripts, and Glagolitic.

The Unicode Consortium welcomes the submission of new characters for possible inclusion in the Unicode standard. For instructions on how to submit characters to the Unicode Consortium, see Appendix D.

1.4 The Unicode Consortium

The Unicode Consortium was formalized in January 1991 to promote the Unicode standard as an international encoding system for information interchange, to aid in its implementation, and to maintain quality control over future revisions. The Unicode Consortium was incorporated as a non-profit organization under the name *Unicode, Inc.*, to provide a central focus and contact point for conducting these activities. Membership is open to organizations anywhere in the world that support the Unicode standard in principle and that would like to assist in its widespread implementation. The consortium is supported through the volunteer efforts of its members (and their companies), and financially through the membership dues.

The Consortium's board of directors and officers come from a variety of organizations and represent a wide spectrum of text-encoding and computing applications. The Consortium's activities are conducted by the Unicode Technical Committee.

The Unicode Technical Committee

The Unicode Technical Committee (UTC) is the working group within the Consortium responsible for the creation, maintenance, and quality of the Unicode standard. The UTC controls all input to the standard and makes associated content decisions. Voting members of the UTC are representatives from Consortium members. However, visitors are welcome to participate in the discussions, since the intent of the UTC is to act as an open forum for free exchange of technical ideas.

The predecessor of the UTC was the less formal Unicode Working Group. Engineering teams from Apple and Xerox, who had been very active in the area of multilingual operating systems, realized their own character encoding methods could be improved. Together they produced the original Unicode design. They were joined by representatives of other companies who were experiencing similar problems. The participants worked for companies and institutions whose businesses required multilingual information systems, including Go, IBM, Metaphor, Claris, Microsoft, NeXT, Sun Microsystems, and The Research Libraries Group. The potential for mutual benefit encouraged them to share the results of past efforts and to explore the opportunities for collaboration.

The *Unicode Standard, Draft 1*, issued in September 1989, contained a preliminary repertoire of characters. A second draft, the *Unicode Preview*, issued in October 1990, contained a much more extensive repertoire of characters, as well as the framework for Han character unification. The

Unicode 1.0 Draft Standard Final Review Document, issued in December 1990, was sent out for worldwide review. Feedback on the final review document was incorporated into the standard during the first part of 1991. This book, *The Unicode Standard: Worldwide Character Encoding, Version 1.0*, finalizes the character repertoire, and provides an implementable standard on which vendors and software developers may base their applications.