The Unicode Standard Version 4.0
Edited by Joan Aliprand, Julie Allen, Joe Becker, Mark Davis, Michael Everson, Asmus Freytag, John Jenkins, Mike Ksar, Rick McGowan, Eric Muller, Lisa Moore, Michel Suigard, and Ken Whistler.
Addison-Wesley
August, 2003.
1462 pages + 38 front pages + CD.
ISBN 0-321-18578-1

One may have had to follow the evolution of the Unicode Standard since the publication of the first version in two volumes in 1991/1992 to fully appreciate the advances made in the most recent one. The intervening dozen years have seen remarkable strides in coming to both an intellectual understanding of the problems of a unified character set, a compilation of rules for encoding it, and an exposure of some of the underlying questions of human language communication.

The sheer size of Version 4.0 is daunting. At 1462 pages, it exceeds Version 3 by 422 pages. Of this, 776 are the *Code Charts* (Chapter 16), a work of art by themselves, and endlessly fascinating as a record of the symbols and mechanisms that humankind has invented in order to cope with the need to communicate with each other. This edition contains 96,248 characters, adding 47,188 new characters for minority and historic scripts, several sets of symbols, and a very large collection of additional Han ideographs beyond that of Version 3.0. Of course, the first 400+ pages are essentially about how to use Unicode, including many of the rationales that have lead to the current form. There is no other body of work that collects so much of this information in one place. It is not usual that the statistics of a published book constitute a major element of a review, but the encyclopedic nature of this work, and the magnitude of the preparation and editing task is an aspect of the work that must be appreciated.

A brief history: the Unicode Consortium was formed late in the 1980's, to create a unified character set and encoding. At about the same time, the International Organization for Standards (ISO), and its International Electrotechnical Commission (IEC) started work on a similarly targeted project; this is the ISO/IEC 10464 standard. The political and technical peace worked between these groups is a demonstration of international cooperation from which all could learn. The Unicode Consortium has sponsored the publication of all versions of the Unicode Standard, and represents a conforming implementation of ISO/IEC 10646.

One of the most disappointing aspects of the Unicode work is the number of computer and information technology professionals who still "do not get it." The question continues to be posed: "What's wrong with ASCII?" It is impossible to argue with the success of ASCII as a "code for information interchange," for machines! In fact, the machine-readable information files describing Unicode on the included CD are themselves strictly in an ASCII representation. However, for of us who believe human written communication is richer than this, the narrowness of the "ASCII suffices" view is depressing. On the other hand, the difficulty of doing better is demonstrated by the body of knowledge captured in the Unicode Standard.

There are 15 chapters before the code charts. The *Introduction* is just that, starting the reader with Unicode and the context in which it is developed. The *General Structure* chapter includes reference to design principles, encoding, writing directions, and other structural issues. The Unicode Standard is a Standard, and Chapter 3 addresses *Conformance*. The key to using the Unicode character set are the machine-readable files, which give the *Character Properties*, described in the fourth chapter. Chapters 5 and 6 deal with questions relating to use and the relationship to how humans write, and are respectively the *Implementation Guidelines* and a

discussion of *Writing Systems and Punctuation*. The remainder of the chapters, from 7 to 15, deals with specific areas of the Standard, These are respectively *European Alphabetic Scripts*, *Middle Eastern Scripts*, *South Asian Scripts*, *Southeast Asian Scripts*, *East Asian Scripts*, *Additional Modern Scripts*, *Symbols*, and *Special Areas and Format Characters*. Chapter 17, following the code charts, is the *Han Radical-Stroke Index* (for locating Han ideographs).

The Han ideographs in the Standard bring up the question of "Han Unification," one of the more (but not the only) controversial principles in the design and construction of the Unicode character set and encoding. Briefly, "Han Unification" refers to identifying ideographs common to Asian languages, and encoding each just once, even if the actual written forms in the languages are not identical. The idea is not new with the work of the Unicode Consortium, but the Unicode Standard brings a new pervasiveness to this approach, and to some Asian language purists, a more creditable threat to language identity than before.

The distinction between character set and character encoding is subtle. Standards such as those for XML specify the use of the character set, and provide rules for representation that go beyond the canonical encoding of the Unicode Standard (for some good reasons). Then, once encoded, the encoding may be transformed, leading to the "Universal Transformation Formats," of which UTF-8 is probably the most commonly used. The modern computer and IT practitioner must appreciate, and be comfortable with, these subtleties.

The work of the Unicode Consortium is not done. A quick visit to the web site at www.unicode.org will lead to a link to "Open Issues for Public Review." Here will be found questions ranging from ones concerning individual characters (Bengali Reph and Ya-Phalaa) to much more technical questions such as "Changing General Category of Braille Patterns to 'Letter Other' " and "Unicode Regular Expressions," as well questions concerning language use in the form of "Terminal Punctuation Characters."

This reviewer is a little saddened by the fact that the Klingon alphabet did not make it into the Unicode repertoire; it would have represented a sense of adventure. To balance the accounting, it is pleasing to see the Shavian alphabet there, for those of us who remember the publication of "Androcles and The Lion" as a parallel text.

It should be mentioned that the CD included with the Standard contains not only the Unicode database files, but also the entirety of the content of the book in the form of PDF files. In fact, all these same files are available on the Unicode Consortium web site. So, is there a reason to buy the "printed-on-paper" version? Of course! This is the reference needed by all organizations having to deal with language, international communication (including software internationalization), or similar questions in social and humanity studies. The online and electronic files are invaluable, once you know for what it is you are looking. If your purpose is education, and being able to see the BIG picture, the whole context, there is still no adequate alternative to the printed book.

The next phase in the story of the Unicode character encoding is for software libraries for Unicode processing to become as universally available, and as widely understood, as those that we know so well from the C/C++ standard library. Java and C# are making progress in that direction, supported by the larger computer companies. To fully understand the challenges, and the successes, and the diversity inherent in our writing systems, ask your organization or local library to invest in this latest publication, the Unicode Standard Version 4.0, and indulge yourself.

Bruce K. Haddon