Chinese, Japanese, Korean & Vietnamese Computing

CJKV

Information Processing

O'REILLY®

Ken Lunde

CJKV Information Processing

CJKV Information Processing

Ken Lunde

O'REILLY™

Beijing · Cambridge · Farnham · Köln · Paris · Sebastopol · Taipei · Tokyo

CJKV Information Processing
by Ken Lunde

Portions of this book previously appeared in *Understanding Japanese Information Processing*,

Published by O'Reilly & Associates, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

Editors: Tim O'Reilly, Peter Mui, and Gigi Estabrook

Production Editors: Ken Lunde and Jane Ellin

Printing History:

January 1999:	First Edition.
October 2002:	Minor corrections.

This book is printed on acid-free paper with 85% recycled content, 15% post-consumer waste. O'Reilly & Associates is committed to using paper with the highest recycled content available consistent with high quality.

ISBN: 1-56592-224-7
[M]

This book is dedicated to the countless people who have touched my life, either personally or professionally. I have become the person I am today—for better or for worse—from their companionship, criticism, encouragement, friendship, generosity, guidance, humor, influence, inspiration, kindness, patience, strength, support, and wisdom. I shall be forever in their debt.

Table of Contents

Foreword

In September 1993, an important event in the history of Japanese computing took place: the publication of Ken Lunde's *Understanding Japanese Information Processing*. Even today, no other book brings together such a wealth of information on Japanese data processing. I and my colleague Takeo Suzuki had the unique honor to work on the Japanese translation of this book, a work that made worldwide impact.

Today, an event of even greater importance—a major milestone in the history of East Asian computing—is unfolding: the publication of Ken Lunde's new book: *CJKV Information Processing*. No other book even pretends to approach it in either content or quality. Though the range of issues covered is broad, the treatment of each topic is comprehensive and in-depth.

With the recent spread of Unicode as an international character set and the growing importance of East Asia in the world economy, new CJK applications are appearing at an increasingly rapid pace. Many software publishers are "jumping on the CJK wagon," with the hopes of tapping the potentially huge East Asian market, which has mostly remained closed to non-Asian software developers.

An important reason for the poor market penetration of these products is their low quality. Many of these applications are simply too primitive to adequately meet the practical needs of users.

This state of affairs can be explained by three interrelated factors. First, this is a young market, and developers have not yet acquired sufficient skills and experience. For example, some developers of Japanese software, who have little or no knowledge of the language, hire outside help such as students, often with disastrous results.

Second, developers often do not have access to high-quality data, especially dictionary data, required for the all-important input method (also known as "front-end processor" or FEP) development. It is easy enough to find such data by surfing the Web, but it requires much skill to eliminate the countless errors and adapt it to specific needs.

Third is the lack of good information on CJKV processing. The issues are complex. The developer must contend with multiple, mostly incompatible encoding systems, different character sets, a bewildering variety of locale-dependent input methods, code conversion between incompatible character sets, and support for Unicode, to mention but a few.

How does one acquire reliable and detailed information on these issues? Until the appearance of the present work, this was well-nigh impossible. For the first time, Ken Lunde's pioneering work provides nothing less than an inexhaustible source of accurate and complete information on every aspect of CJKV data processing.

Let me illustrate how useful this information was to the dictionary projects of the Kanji Dictionary Publishing Society (KDPS). We have recently completed *The Kodansha Kanji Learner's Dictionary* (Kodansha International, 1998), based on the *New Japanese-English Character Dictionary* (Kenkyusha, 1990; NTC, 1993) of which I am the chief editor. At the same time, we have been developing DESK, a comprehensive CJK database from which dozens of dictionaries, CJK FEP data, and learning aids are being developed.

Though we are dedicated CJK specialists, the information in Ken Lunde's book was invaluable every step of the way. For example, when we created outline fonts for some 1,350 user-defined characters, it helped us decide on encoding ranges and methods. When we switched to a new platform, it helped us write code conversion routines and build function libraries. When we developed our dictionary page composition system, it guided us in the purchase of software and taught us in depth about typography and font technology. And so on and so on.

As can be seen from our example, the aim of this book is *highly practical*. The author has a very full grasp of the real needs of such diverse users as software developers, lexicographers, and language learners, and provides detailed information for each need with great clarity and precision. I am fully confident that this book shall become an invaluable source of information to everyone interested in CJKV information processing.

Jack Halpern (春遍雀來)
Editor in Chief, CJK Dictionary Publishing Society (CDPS)
http://www.cjk.org/

Preface

Close to six years have passed since *Understanding Japanese Information Processing* was published, and a lot has changed since then. One reason for the delay in getting this new book published was that I decided not only to revise the Japanese portions, but also to include significantly more information about Chinese and Korean, plus add information about Vietnamese—the title was changed accordingly. I was inspired to undertake this significant "CJKV" expansion sometime in 1996, during a lengthy conversation I had at a Togo's near the UC Berkeley campus with Peter Mui, my editor for *Understanding Japanese Information Processing*.

Join me in reading this thick tome of a book, and you shall find that "CJKV" (Chinese, Japanese, Korean, and Vietnamese) will become a standard term in your arsenal of knowledge. But, before we dive in, allow me to use some terms with which you are no doubt familiar. Otherwise, you probably would have no need to continue reading.

Known to more and more people, "internationalization" and "localization" seem to have become household or "buzz" words in the field of computing, and have also become very hot topics among high-tech firms and researchers due to the expansion of software markets to include virtually all parts of the globe. This book is specifically about CJKV-enabling, which is the adaptation of software for one or more CJKV markets. It is my intention that readers will find relevant and useful CJKV-enabling information within these pages.

Virtually every book on internationalization or localization includes information on character sets and encodings, but this book provides much more. In summary, this book provides a brief description of the writing systems, a thorough background of the history and current state of character sets, detailed information on encoding methods, code conversion techniques, input methods, keyboard arrays, font

formats, typography, output methods, algorithms with sample source code, tools that perform useful information processing tasks, and how to handle CJKV text with email and in the context of the Web. Expect to find plenty of platform-independent information and discussions about character sets, how CJKV text is encoded and handled on a number of computer systems, and basic guidelines and tips for developing software targeted for CJKV markets.

Now, let me tell you what this book is *not* about. Don't expect to find out how to design your own word processor, how to design your own fonts for use on your computer (I give sources for tools, though), or how to properly handle formats for CJKV numerals, currency, dates, or times. This book is not by any stretch of the imagination a complete reference manual for internationalization or localization, but should serve well as a companion to such reference works (which are, fortunately, slowly becoming more abundant).

It is my intention for this book to become the definitive source for information relating to CJKV information processing issues (*Understanding Japanese Information Processing*, which concentrated on Japanese issues, apparently became the definitive source for that field). Thus, this book focuses heavily on how CJKV text is handled on computer systems in a very platform-independent way. Everything presented in this book can be programmed, categorized, or easily referenced.

This book was written to fill the gap in information relating to CJKV information processing. I first attempted to do this over the course of several years by maintaining an online document that I named *JAPAN.INF* (*Electronic Handling of Japanese Text*). This document had been made publicly available through a number of FTP sites worldwide, and had gained international recognition as *the* source for information relating to Japanese text handling on computer systems. *Understanding Japanese Information Processing* excerpted and further developed key information contained in *JAPAN.INF*. However, since the publication of *Understanding Japanese Information Processing* in 1993, *JAPAN.INF*, well, uh, sort of died. Not a horrible death, mind you, but rather to prepare for its reincarnation as a totally new online document that I have entitled *CJK.INF* (the CJK analog to *JAPAN.INF*). The work I did on *CJK.INF* helped to prepare me to write this new book, which provides updated material plus *significantly* more information about Chinese, Korean, and Vietnamese (to the point that granting the book a new title was deemed appropriate and necessary). I hope that this book becomes as widely accepted as the original.

While I have expended great efforts to provide sufficient amounts of information for Chinese, Japanese, Korean, and Vietnamese computing, you will notice that there is still some bias toward Japanese in many parts of this book. But, almost everything discussed in this book can apply equally to all of these languages.

However, the details of Vietnamese computing in the context of using Chinese characters are still emerging, so its coverage is somewhat limited.

Audience

Anyone interested in how CJKV text is processed on a computer will find this book useful, including those who wish to enter the field of CJKV information processing, and those who are already in the field, but need additional reference material. This book will also be useful for people using any kind of computer and any type of computer operating system: MacOS, MS-DOS, Unix, and Windows.

Although this book is specifically about CJKV information processing, anyone with an interest in creating multilingual software or a general interest in I18N (internationalization) or L10N (localization) will learn a great deal about the issues involved in handling complex writing systems on computers. This is particularly true for people interested in working with CJKV text. Information relating to CJKV-enabling is still, unfortunately, relatively scarce.

I assume that readers have little or no knowledge of a CJKV language (Chinese, Japanese, Korean, or Vietnamese) and its writing system. In Chapter 2, *Writing Systems*, I include material that should provide a good introduction to CJKV languages and their writing systems. If you only know one CJKV language, Chapter 2 should prove to be quite useful.

Conventions Used in this Book

Kanji, *hanzi*, *hanja*, *hangul*, *kana*, *hiragana*, *katakana*, and other terms come up time and time again throughout this book. You will also encounter abbreviations and acronyms, such as ANSI, ASCII, CNS, EUC, GB, ISO, JIS, KS, and TCVN. Terms, abbreviations, and acronyms—along with many others—are usually explained in the text and again in the glossary (Appendix X, *Glossary*, which I encourage you to study).

Hexadecimal values, when used in text, are prefixed with 0x, such as 0x8080. Every two hexadecimal digits beyond 0x represent a single byte. For example, 0x20 represents a one-byte value, but 0x0020 represents a two-byte value. Decimal values appear as themselves. You can use Appendix B, *Notation Conversion Table*, to convert between notations.

Throughout this book I generically use short suffixes such as “J,” “K,” “S,” “T,” “V,” and “CJKV” to denote locale-specific or CJKV-capable versions of software products. I use these suffixes for the sake of consistency, and because software manufacturers often change the way in which they denote CJKV versions of their products. In practice, you may instead encounter the suffix 日本語版 (*nihon-*

goban, meaning "Japanese version"), the prefix "Kanji," or the prefix 日本語 (*nihongo*, meaning "Japanese") in Japanese product names. For Chinese software, 中文 (*zhōngwén*, meaning "Chinese") is a common prefix. I also refrain from using version numbers for software described in this book (as you know, this sort of information becomes outdated very quickly). I use version numbers only when they represent a significant advancement or development stage in a product.

References to "China" in this book refer to the *People's Republic of China* (PRC; 中华人民共和国 *zhōnghuá rénmín gònghé guó*), also commonly known as Mainland China. References to "Taiwan" in this book refer to the *Republic of China* (ROC; 中華民國 *zhōnghuá mínguó*). Quite often this distinction is necessary.

Name ordering in this book, when transliterated in Latin characters, follows the convention that is used in the West—the given name appears first, followed by the surname. When the name is written using CJKV characters—in parentheses following the transliterated version—the surname appears first, followed by the given name.

"ISO 10646-1:1993" and "Unicode" are used interchangeably throughout this book. Only in some specific contexts are they different.

Italic is used for pathnames, filenames, program names, new terms where they are defined, newsgroup names, and Internet addresses, such as domain names, URLs, and email addresses.

`Constant width` is used in examples to illustrate output from commands, the contents of files, or the text of email messages.

`Constant bold` is used in examples to indicate commands or other text that should be typed literally by the user; occasionally, it is also used to distinguish parts of an example.

`Constant oblique` is used in code fragments and examples to show variables for which a context-specific substitution should be made. The variable *`email address`*, for example, would be replaced by an actual email address.

The `%` (percent) character is used to represent the Unix shell prompt in Unix command lines.

Footnotes are used for parenthetical remarks. Lies are sometimes spoken to simplify or shorten the discussion (especially in Chapter 2 where I introduce the many CJKV writing systems), and the footnotes—usually, but not always—restore the truth.

Organization

Let's now preview the contents of each chapter in this book. Don't feel compelled to read this book linearly, but feel free to jump around from section to section and into the appendixes. Also, the index is there for you to use.

Chapter 1, *CJKV Information Processing Overview*, contains an overview of the issues that are addressed by this book, and will give you an idea of what you can expect to learn. This establishes the context in which this book will become useful in your work or research.

Chapter 2, *Writing Systems*, contains information directly relating to CJKV writing systems. Here you will learn about the various types of characters that compose CJKV texts. This chapter is intended for readers who are not familiar with the Chinese, Japanese, Korean, or Vietnamese languages (or who are familiar with only one or two of those languages). Everyone is bound to learn something new here.

Chapter 3, *Character Set Standards*, describes the two classes of CJKV character set standards: *coded* and *non-coded*. Coded character set standards are further divided into two classes: *national* and *international*. Comparisons are also drawn between CJKV character set standards.

Chapter 4, *Encoding Methods*, contains information on how the character set standards described in Chapter 3 are encoded on computer systems. Encoding is a complex but important step in representing and manipulating human-language text in a computer. Other topics include software for converting from one CJKV encoding to another, and instructions on how to repair damaged CJKV text files.

Chapter 5, *Input Methods*, contains information on how CJKV text is input. First I discuss CJKV input in general terms, then describe several specific methods for inputting CJKV characters on computer systems. Next, we move on to the hardware necessary for CJKV input, specifically keyboard arrays. These range from common keyboard arrays, such as the QWERTY array, to Chinese character tablets containing thousands of individual keys.

Chapter 6, *Font Formats*, contains information about bitmapped and outline font formats as they relate to CJKV. The information presented in this chapter represents my daily work at Adobe Systems, so some sections may suffer from excruciating detail.

Chapter 7, *Typography*, contains information about how CJKV text is properly laid out on a printed page. Having CJKV fonts is not enough—there are rules that govern where characters can and cannot be used, and how different character

classes are handled when in proximity. The chapter ends with a description of software programs that provide advanced line layout functionality.

Chapter 8, *Output Methods*, contains information about how to display, print, or otherwise output CJKV text. Here you will find information relating to the latest printing and display technologies.

Chapter 9, *Information Processing Techniques*, contains information and algorithms relating to CJKV code conversion and text handling techniques. The actual mechanics are described in detail, and, where appropriate, include algorithms written in C, Java, and other programming languages. The chapter ends with a brief description of three Japanese code processing tools that I have written and maintained over a period of several years. These tools show how the algorithms can be applied in the context of Japanese.

Chapter 10, *Operating Systems, Text Editors, and Word Processors*, contains information about operating systems, text editors, and word processors that are CJKV-capable, meaning that they support one or more CJKV locale.

Chapter 11, *Dictionaries and Dictionary Software*, contains information about dictionaries, both printed and electronic, that are useful when dealing with CJKV text. Also included are tips on how to more efficiently make use of the various indexes used to locate Chinese characters in dictionaries.

Chapter 12, *The Internet*, contains information on how CJKV text is best handled electronically over networks such as email systems and news readers. Included are tips on how to ensure that what you send is received intact as well as information about the Internet domains that cover the CJKV locales.

Chapter 13, *The World Wide Web*, contains information on displaying CJKV text using various web browsers, and provides instructions for creating your own HTML (HyperText Markup Language) and XML (Extensible Markup Language) documents containing CJKV text. The role of Adobe Acrobat, PDF (Portable Document Format), and CGI (Common Gateway Interface) programming is also discussed in detail.

Appendix A, *Code Conversion Tables*, provides a code conversion table between decimal Row-Cell, hexadecimal ISO-2022, hexadecimal EUC, and hexadecimal Shift-JIS (Japanese-specific) codes. Also included is an extension that handles the Shift-JIS user-defined range.

Appendix B, *Notation Conversion Table*, lists all 256 eight-bit byte values in four common notations: binary, octal, decimal, and hexadecimal.

Appendix C, *Vendor Character Set Standards*, is reference material for those interested in vendor-specific extensions to CJKV character set standards.

Appendix D, *Vendor Encoding Methods*, is reference material for those interested in how the vendor character sets in Appendix C are encoded.

Appendix E, *GB 2312-80 Table*, is a code table for the characters defined in GB 2312-80 (along with the additions and corrections stipulated by GB 6345.1-86), indexed by decimal Row-Cell codes.

Appendix F, *GB/T 12345-90 Table*, is a code table for the characters defined in GB/T 12345-90, indexed by decimal Row-Cell codes.

Appendix G, *CNS 11643-1992 Table*, is a code table for the characters defined in all seven planes of CNS 11643-1992, indexed by decimal Row-Cell codes. Also included are tables that are specific to CNS 11643-1986, such as Plane 15. This is a long appendix, listing well over 50,000 unique hanzi!

Appendix H, *Big Five Table*, is a code table for the characters defined in Big Five, indexed by hexadecimal Big Five codes.

Appendix I, *Hong Kong GCCS Table*, contains a table for the complete set of 3,049 hanzi promulgated by the Hong Kong government, indexed by hexadecimal Big Five codes. Also included is an additional set of 145 hanzi defined by Hong Kong's Department of Judiciary.

Appendix J, *JIS X 0208:1997 Table*, is a code table for the characters defined in JIS X 0208:1997, indexed by decimal Row-Cell codes.

Appendix K, *JIS X 0212-1990 Table*, is a code table for the characters defined in JIS X 0212-1990, indexed by decimal Row-Cell codes. Also included are four additional katakana characters that *may* be added to this standard in the future (but this seems unlikely).

Appendix L, *KS X 1001:1992 Table*, is a code table for the characters defined in KS X 1001:1992, indexed by decimal Row-Cell codes.

Appendix M, *KS X 1002:1991 Hanja Table*, is a code table for *only* the 2,856 hanja defined in KS X 1002:1991, indexed by decimal Row-Cell codes.

Appendix N, *Hangul Reading Table*, contains a complete reading index for all 2,350 modern hangul defined in the KS X 1001:1992 character set standard.

Appendix O, *TCVN 6056:1995 Table*, is a code table for the characters defined in TCVN 6056:1995, indexed by decimal Row-Cell codes.

Appendix P, *Code Table Indexes*, provides various Chinese character indexes—reading, radical, and stroke-count—to be used in conjunction with various appendixes in this book.

Appendix Q, *Character Lists and Mapping Tables*, contains lists of characters and mapping tables referred to throughout this book.

Appendix R, *Chinese Character Lists*, provides a printout of various Chinese character lists—based on non-coded character set standards—as described in Chapter 3, *Character Set Standards*.

Appendix S, *Single-Byte Code Tables*, provides complete ASCII, EBCDIC, EBCDIK, ISO 8859-1:1998, CJKV-Roman, half-width katakana, and half-width jamo code tables, indexed by hexadecimal codes.

Appendix T, *Software and Document Sources*, provides addresses and contact information for software and documents mentioned throughout the book.

Appendix U, *Mailing Lists*, provides information on various (email-based) mailing lists that may be of interest to readers.

Appendix V, *Professional Organizations*, includes information on organizations that deal with CJKV information processing issues.

Appendix W, *Perl Code Examples*, provides Perl equivalents of algorithms found in Chapter 9—along with other goodies.

Appendix X, *Glossary*, defines many of the concepts and terms used throughout this book (and other books).

Finally, the *Bibliography* lists many useful references, some of which I used to write this book.

Acknowledgments

To write a reference work this thick requires interaction with people from around the globe. It is quite impossible for me to list all the people who have helped me over the years—there are literally hundreds.

In some cases, people simply come to me for help on a particular subject (that's what happens, I guess, when people are aware of your email address—to ensure that I will receive a ton of email, in various parts of this book you will find that my email address is *lunde@oreilly.com*). Sometimes I may not know the answer to a particular question, but the question usually inspires me to seek out the truth. The truth *is* out there.

1998 marks seven wonderful years at Adobe Systems, a company that provides me with daily CJKV-related challenges. Its advanced font technology and commitment to customers is what initially attracted me, and this is what keeps me there. Besides, they let me keep a limited-edition three-foot bright-orange super melt-down Godzilla in my office, along with a similarly-sized Gamera figure, as a memorial and tribute to *The King* (who quite sadly passed away in the 1995 film

Godzilla versus Destroyah[*]). Speaking of tributes, all aspects of the production of this book are a tribute to Adobe Systems' publishing technology.

To all the people who have read my previous writings, put up with my sometimes dull or otherwise annoying personality at work, pointed out errors in my work, exchanged email with me for whatever reason, or otherwise helped me to become a better person: *thank you!* You should know who you are.

Special thanks go to Tim O'Reilly (the president and founder of O'Reilly & Associates) and Peter Mui for believing in my first book, *Understanding Japanese Information Processing*. It was Peter who encouraged me to expand it to cover the complete CJKV framework. (I am sorry that it took so long to get done—it was a long and painful experience.) Thanks go to Edie Freedman for sticking with my idea of a blowfish for the cover.[†] Mike Sierra graciously helped me through the layout of the book, and nurtured my desire to learn Adobe FrameMaker's particular paradigm. Chris Reilley also deserves a lot of credit for turning my poorly designed figures into works of fine art—for a second time. Gigi Estabrook, the editor, continually pushed me to get this book done. Ellie Fountain Maden performed the copyedit, discovering various errors and oddities.

The following were responsible for reviewing various parts of this book, during the various stages of its prolonged development: Joe Becker, Jim Breen, Robert Bringhurst, Woohyong Choi (최 우형), James Davis, L. Peter Deutsch, James Đỗ (杜伯福), Terry Dowling, Martin Dürst, Jeff Engelman, Gus Fernandez, Jeffrey Friedl, David Gourley, Jerry Hall, Jack Halpern (春遍雀來), Ken'ichi Handa (半田剣一), Dennis Hanks, Ted Harrison, Patty Hay (許珮婷), Carl Hoffman, Chiaki Ishikawa (石川千秋), Matt Jacobs, David Kelly, Hoon Kim (김 훈), Kyongsok Kim (김 경석), Kazuo Koike (小池和夫), Akira Komatsu (小松章 or 郑褚璋), Norbert Lindenberg, Toshiaki Maeda (前田年昭), Dirk Meyer, Charles Muller, Terry O'Donnell, Glen Perkins, Etsuko Obata Reiman (エツコ・オバタ・ライマン), Craig Rublee, Limin Shi (施利民), Kohji Shibano (芝野耕司), Jungshik Shin (신 정식), Frank (Yung-Fong) Tang (譚永鋒), Ngô Trung Việt (吳中越), Taro Yamamoto (山本太郎), Koichi Yasuoka (安岡孝一), and Haifeng Zhu (朱海峰). I am grateful to all of them for providing me with useful insights and inspiring ideas. I am, however, responsible for any errors, omissions, or oddities that you may encounter.

---

* ゴジラ対デストロイア (*gojira tai desutoroia*) in Japanese. As a side-note to this footnote, the creator of Godzilla, Tomoyuki Tanaka, passed away in April of 1997 at the age of 86. The first American-made Godzilla film was released in 1998. Like this book, size *does* matter.

† Michael Slinn made the astute observation that the Babel Fish would have been more appropriate as a cover creature for this book—according to Douglas Adams' *The Hitch Hiker's Guide to the Galaxy*, you simply stick a Babel Fish in your ear, it connects with your brain, and you can suddenly understand all languages. Perhaps the blowfish is still used for this book's cover because there were no nineteenth-century Babel Fish engravings in the Dover Pictorial Archive…

Finally, I wish to thank my wonderful parents, Vernon Delano Lunde and Jeanne Mae Lunde, for all of their support throughout the years; my son, Edward Dharmputra Lunde; my step-son, Ryuho Kudo (工藤龍芳); my beautiful daughter, Ruby Mae Lunde (工藤瑠美); and my beloved wife and partner, Hitomi Kudo (工藤仁美).

Errors, Omissions, and Updates

A book containing this much highly technical information is bound to contain some errors. No doubt, these errors will be corrected in future printings or editions of this book. In the meantime, any errors will be maintained at the following URL:

ftp://ftp.oreilly.com/pub/examples/nutshell/cjkv/errata/

If you happen to find any errors or notice any omissions, please send them to the following address:

O'Reilly & Associates, Incorporated
1005 Gravenstein Highway North
Sebastopol, CA 95472 USA
800-998-9938 (in the USA or Canada)
+1-707-829-0515 (international or local)
+1-707-829-0104 (facsimile)

You can also send messages electronically. To be put on O'Reilly's mailing list or to request a catalog, send email to:

info@oreilly.com

To ask technical questions or comment on this book, send email to:

bookquestions@oreilly.com

Because this book is filled with hundreds of URLs, I am providing the following web page, which arranges them all by chapter/appendix then by page number, for easier, clickable access (and as a way for me to keep them all up-to-date):

http://www.oreilly.com/~lunde/cjkv-urls.html

In this chapter:

- *Multiple Writing Systems*
- *Character Set Standards*
- *Encoding Methods*
- *Input Methods*
- *Typography*
- *Basic Concepts and Terminology*

1

CJKV Information Processing Overview

A lot of mystique and intrigue surrounds how CJKV—Chinese, Japanese, Korean, and Vietnamese—text is handled on computer systems. Although I agree with there being intrigue, there is far too much mystique, in my opinion. Much of this mystery is due to a lack of information, or simply a lack of information written in a language other than Chinese, Japanese, Korean, or Vietnamese. Nevertheless, many fine folks, like you, would like to know how this all works. To confirm some of your worst fears and speculations, CJKV text *does* require special handling on computer systems. However, it should not be very mysterious after having read this book. You need only break the so-called *one-byte-equals-one-character* barrier—most CJKV characters are represented by more than a single byte (or, to put it in another way, more than eight bits).*

English information processing was a reality soon after the introduction of early computer systems, which were first developed in England and the United States. Adapting software to handle more complex writing systems such as those used to represent CJKV text is a more recent phenomenon. This adaptation developed in various stages, and continues today.

There are several key issues that make CJKV text a challenge to process on computer systems:

- CJKV writing systems use a mixture of different, but sometimes related, writing systems
- CJKV character set standards enumerate thousands or tens of thousands of characters, which is orders of magnitude more than used in the West

* For a greater awareness of (and appreciation for) some of the complexities of dealing with multiple-byte text, you might consider glancing now at the section entitled "Byte Versus Character Handling" in Chapter 9, *Information Processing Techniques*, beginning on page 433.

- There is no universally recognized or accepted CJKV character set standard such as ASCII for writing English—although Unicode can be considered a good first attempt
- There is no universally recognized or accepted CJKV encoding system such as ASCII encoding—again, the various Unicode encodings can be considered an attempt at accomplishing this
- There is no universally recognized or accepted input device such as the QWERTY keyboard array—although this same keyboard array, through a method of transliteration, can be used to input most CJKV text through reading or other means
- CJKV text can be written horizontally or vertically, and requires special typographic rules not found in Western typography, such as spanning tabs and unique line-breaking rules

You will learn that the ASCII character set standard is not as universal as most people think—different flavors of ASCII exist, as do different ASCII encoding methods. You will begin to wonder why so many developers assume that everyone uses ASCII.

This chapter also includes several sections that explain and illustrate some very basic yet important computing concepts, such as notation and byte order, that relate to material in the remainder of this book. Even if you consider yourself a seasoned software engineer or expert programmer, you may still find value in those sections because they carry much more importance in the context of CJKV information processing. That is, how these concepts relate to CJKV information processing may be slightly different than what you previously learned.

Multiple Writing Systems

CJKV text is typically composed of a mixture of different writing systems. Japanese, as an example, is unique in that it uses four different writing systems. Others, such as Chinese and Korean, use less than four writing systems. Japanese is one of the few, if not the only, languages that exhibit this characteristic of so many writing systems being used together, even in the same sentence (as you will see very soon). This makes Japanese quite complex, orthographically speaking, and poses several problems.* The four Japanese writing systems are Latin characters, hiragana, katakana, and kanji (collectively referred to as "Chinese characters" regardless of the language). You are already familiar with Latin characters because the English language is written with these—this writing system consists of the

* Orthography is a linguistic term that refers to the writing system of a language.

upper- and lowercase Latin alphabet, which are the characters often found on typewriter keys. Hiragana and katakana are native Japanese syllabaries (see Appendix X, *Glossary*, for a definition of "syllabary"). Both hiragana and katakana represent the same set of 108 syllables, and are collectively known as *kana*. Kanji are characters that the Japanese borrowed from China over 1,600 years ago—Chinese characters number in the thousands, and encompass meaning, reading, and shape.

Now let's look at an example sentence composed of these four writing systems. This should serve to illustrate how the different Japanese writing systems can be effectively mixed.

> EUC 等のエンコーディング方法は日本語と英語が混交しているテキストをサポートします。

In case you are curious, this sentence means "Encoding methods such as EUC can support texts that mix Japanese and English." Let's look at this sentence again, but with the Latin characters underlined.

> EUC 等のエンコーディング方法は日本語と英語が混交しているテキストをサポートします。

In this case there is a single abbreviation, EUC (short for *Extended Unix Code*, which refers to a locale-independent encoding method, a topic to be covered in Chapter 4, *Encoding Methods*, of this book). It is quite common to find Latin characters used for abbreviations in CJKV texts. Latin characters used to transliterate Japanese text are called ローマ字 (*rōmaji*) in Japanese.

Now let's underline the katakana characters.

> EUC 等のエンコーディング方法は日本語と英語が混交しているテキストをサポートします。

Each katakana character represents one syllable, typically a lone vowel or a consonant-plus-vowel combination. Katakana characters are commonly used for writing words borrowed from other languages, such as English, French, or German. Table 1-1 lists these three underlined katakana words, along with their meanings and readings.

Table 1-1: Sample Katakana

Katakana	Meaning	Reading[a]
エンコーディング	*encoding*	enkōdingu
テキスト	*text*	tekisuto
サポート	*support*	sapōto

[a] The macron is used to denote long vowel sounds.

Note how their readings closely match that of their English counterparts, from which they were derived. This is no coincidence: it is common for the Japanese readings to be spelled out with katakana characters to closely match the borrowed words.

Next we underline the hiragana characters.

EUC 等のエンコーディング方法は日本語と英語が混交しているテキストをサポートします。

Hiragana characters, like katakana described above, represent syllables. Hiragana characters are mostly used for writing grammatical words and inflectional endings. Table 1-2 illustrates the usage or meaning of the hiragana in the above sentence.

Table 1-2: Sample Hiragana

Hiragana	Meaning or Usage	Reading
の	possessive marker	no
は	topic marker	wa[a]
と	*and* (conjunction)	to
が	subject marker	ga
している	*doing…* (verb)	shite-iru
を	object marker	o
します	*do…* (verb)	shimasu

[a] This hiragana character is normally read *ha*, but when used as a topic marker, it becomes *wa*.

That's a lot of grammatical stuff! Japanese is a postpositional language, meaning that grammatical markers, such as prepositions as used in English, come after the nouns that they modify. These grammatical markers are called *particles* (助詞 *joshi*) in Japanese.

Finally, we underline the Chinese characters (called *hànzì* in Chinese, *kanji* in Japanese, *hanja* in Korean, and *chữ Hán* in Vietnamese):

EUC 等のエンコーディング方法は日本語と英語が混交しているテキストをサポートします。

At first glance, Chinese characters appear to be more complex than the other characters in the sentence. This happens to be true most of the time. Chinese characters represent meanings, and are often called *ideographs*, *pictographs*, or *logographs*.* Chinese characters are also assigned one or more readings (pronunci-

* Being a widespread convention, this is beyond critique. However, linguists use these terms for different classes of Chinese characters, depending on their etymology.

ations), each of which is determined by context. While their readings differ depending on the language (Chinese, Japanese, Korean, or Vietnamese), Chinese characters often have the same meaning. This makes it possible for Japanese to understand (but not necessarily to pronounce) some very basic Chinese, Korean, and Vietnamese texts. Table 1-3 provides a listing of the underlined Chinese characters and compounds (words composed of two or more Chinese characters) thereof, along with their meanings and readings.

Table 1-3: Sample Chinese Characters and Chinese Character Compounds

Chinese Characters	Meaning	Reading
等	*such as …*	nado
方法	*method*	hōhō
日本語	*Japanese (language)*	nihongo
英語	*English (language)*	eigo
混交	(*to*) *mix*	konkō

Of course, this example includes only those types of characters that are used in Japanese—other locales use different types of characters. Table 1-4 lists the four CJKV locales, along with what writing systems they use.

Table 1-4: CJKV Locales and Their Writing Systems

Locale	Writing Systems
China	Latin and hanzi (simplified)
Taiwan	Latin, zhuyin, and hanzi (traditional)
Japan	Latin, hiragana, katakana, and kanji
Korea[a]	Latin, jamo, hangul, and hanja
Vietnam	Latin (Quốc ngữ), chữ Nôm, and chữ Hán

[a] Jamo are the alphabet-like components that make up hangul.

Table 1-5 lists some sample characters from each of the writing systems used in CJKV locales. We discuss these writing systems in greater detail in Chapter 2, *Writing Systems*.

Table 1-5: Sample CJKV Characters

Character Class	Sample Characters
Latin Characters	A B C D E F G H I J … q r s t u v w x y z
Zhuyin	ㄅ ㄆ ㄇ ㄈ ㄉ ㄊ ㄋ ㄌ ㄍ ㄎ … ㄠ ㄡ ㄢ ㄣ ㄤ ㄥ ㄦ ㄧ ㄨ ㄩ
Hiragana	ぁ あ ぃ い ぅ う ぇ え ぉ お … り る れ ろ ゎ わ ゐ ゑ を ん
Katakana	ァ ア ィ イ ゥ ウ ェ エ ォ オ … ロ ヮ ワ ヰ ヱ ヲ ン ヴ ヵ ヶ

Table 1-5: Sample CJKV Characters (continued)

Character Class	Sample Characters
Jamo	ㄱ ㄲ ㄳ ㄴ ㄵ ㄶ ㄷ ㄸ ㄹ ㄺ … ㅀ ㅇ ㅘ ㅙ ㅚ ㅝ ㅞ ㅟ ㆍ ㆍㅣ
Hangul	가 각 간 갇 갈 갉 갊 감 갑 값 … 횝 횡 히 힉 힌 힐 힘 힙 힛 힝
Hanzi (simplified)	啊 阿 埃 挨 哎 唉 哀 皑 癌 蔼 … 黪 黯 黩 鼬 黫 黮 黭 鼹 鼾 齾
Hanzi (traditional)	一 乙 丁 七 乃 九 了 二 人 儿 … 龘 龜 鸝 灩 灪 爨 鬱 鬻 龖 龘
Kanji	亜 唖 娃 阿 哀 愛 挨 姶 逢 葵 … 齶 龕 龜 龠 堯 槇 遙 瑤 凜 熙
Hanja	伽 佳 假 價 加 可 呵 哥 嘉 嫁 … 晞 曦 熙 熹 熺 犧 禧 稀 羲 詰

But, how frequently used are each of these character classes? Given an average sampling of Japanese writing, one normally finds 30 percent kanji, 60 percent hiragana, and 10 percent katakana. Actual percentages depend on the nature of the text. For example, you may find a higher percentage of kanji in technical literature, and a higher percentage of katakana in fields such as fashion and cosmetics, which make extensive use of loan words written in katakana. Most Korean texts consist of nothing but hangul, and most Chinese texts are composed of only hanzi.* Latin characters are used the least, except in Vietnam.

So, how many characters do you need to learn in order to read and write CJKV languages effectively? Here are some *very* basic guidelines:

- You must learn hiragana and katakana if you plan to deal with Japanese—this constitutes approximately 200 characters
- Learning hangul is absolutely necessary for Korean, but you can get away with not learning hanja
- You need to have general knowledge of about 1,000 kanji to read over 90 percent of the kanji in typical Japanese texts—more are required for reading Chinese texts because only hanzi are used

If you have not already learned Chinese, Japanese, Korean, or Vietnamese, I encourage you to learn one of them so that you can better appreciate the complexity of their writing systems. Although I discuss character dictionaries (and learning aids to a lesser extent) in Chapter 11, *Dictionaries and Dictionary Software*, they are no substitute for a human teacher.

Character Set Standards

A character set simply provides a common *bucket* of characters. You may have never thought of it this way, but the English alphabet is an example of a character

* Well, you will also find symbol-like characters, such as punctuation marks.

set standard. It specifies 52 upper- and lowercase letters. Character set standards are used to ensure that we learn a minimum number of characters in order to communicate with others in society. In effect, they limit the number of characters we need to learn. There are only a handful of characters in the English alphabet, so nothing is really being limited, and as such, there really is no character set standard *per se*. In the case of languages that use Chinese characters, however, character set standards play an especially vital role. They specify which Chinese characters—out of the tens of thousands in existence—are the most important to learn. The current Japanese set, called Jōyō Kanji (常用漢字 *jōyō kanji*), although advisory, limits the number of Chinese characters to 1,945.* There are similar character sets in China, Taiwan, and Korea. These character set standards were designed with education in mind, and are referred to as *non-coded* character sets.

Character set standards designed for use on computer systems are almost always larger than those used for the purpose of education, and are referred to as *coded* character sets. Establishing coded character set standards for use with computer systems is a way to ensure that everyone is able to view documents created by someone else. ASCII is a Western character set standard, and ensures that their computer systems can communicate with each other. But, as you will soon learn, ASCII is not sufficient for the purpose of professional publishing (neither is its most common extension, ISO 8859-1:1998).

Coded character set standards typically contain characters above and beyond those found in non-coded ones. For example, the ASCII character set standard contains 94 printable characters—42 more than the upper- and lowercase alphabet. In the case of Japanese, there are thousands of characters in the coded character sets in addition to the 1,945 in the basic non-coded character set. The basic coded Japanese character set standard, in its most current form, enumerates 6,879 characters, and is designated JIS X 0208:1997. There are four versions of this character set, each designated by the year in which it was established: 1978, 1983, 1990, and 1997. There are two typical compatibility problems that you may encounter when dealing with different versions of the same character set standard:

- Some of these versions contain different numbers of characters—later versions generally add characters
- Some of these versions are not 100 percent compatible with each other due to changes

In addition, there may be an extended character set standard, such as Japan's JIS X 0212-1990, that defines 6,067 additional characters (most of which are kanji).

* The predecessor of this character set, Tōyō Kanji (当用漢字 *tōyō kanji*), was prescriptive.

Additional incompatibility occurs because operating system developers take these coded character set standards one step further by defining their own extensions. These vendor character set standards are largely, but not completely, compatible, and almost always use one of the national standards as their base. When you factor in vendor character set standards, things appear to be a big mess. This book documents these character sets, making it easier to grapple with such confusion.

Encoding Methods

Encoding is the process of mapping a character to a numeric value. By doing this, you create the ability to uniquely identify a character through its associated numeric value. Ultimately, the computer needs to manipulate the character as a numeric value. Independent of any CJKV language or computerized implementations thereof, indexing encoded values allows a numerically enforced ordering to be mapped onto what might otherwise be a randomly ordered natural language. While there is no universally recognized encoding method, many are commonly used. For example, ISO-2022-KR, EUC-KR, Johab, and Unified Hangul Code (UHC) for Korean.

First, before describing these encoding methods, here's a short explanation of how memory is allocated on computer systems. Computer systems process data called bits. These are the most basic units of information, and can hold one of two possible values: on or off. These are usually mapped to the values 1 or 0, respectively. Bits are strung together into units called bytes. Bytes are usually composed of seven or eight bits. Seven bits in an array allow for up to 128 unique combinations, or values; eight bits allow for up to 256. While these numbers are sufficient for representing most characters in Western writing systems, it does not even come close to accommodating large character sets whose characters number in the thousands, such as those required by the CJKV locales.

The first attempt to encode Chinese characters on computer systems involved the use of Japanese half-width katakana characters. This is a limited set of 63 characters that constitutes a minimal set for representing Japanese text. But there was no support for kanji. The solution to this problem, at least for Japanese, was formalized in 1978, and employed the notion of using two bytes to represent a single character. This did not eliminate the need for one-byte characters, though. The Japanese solution was to extend the notion of one-byte character encoding to include two-byte characters. This allows for text with mixed one- and two-byte characters. How one- and two-byte characters are distinguished depends on the encoding method. Two bytes equal 16 bits, and thus can provide up to 65,536 unique values. This is best visualized as a 256×256 matrix. See Figure 1-1 for an illustration of such a matrix.

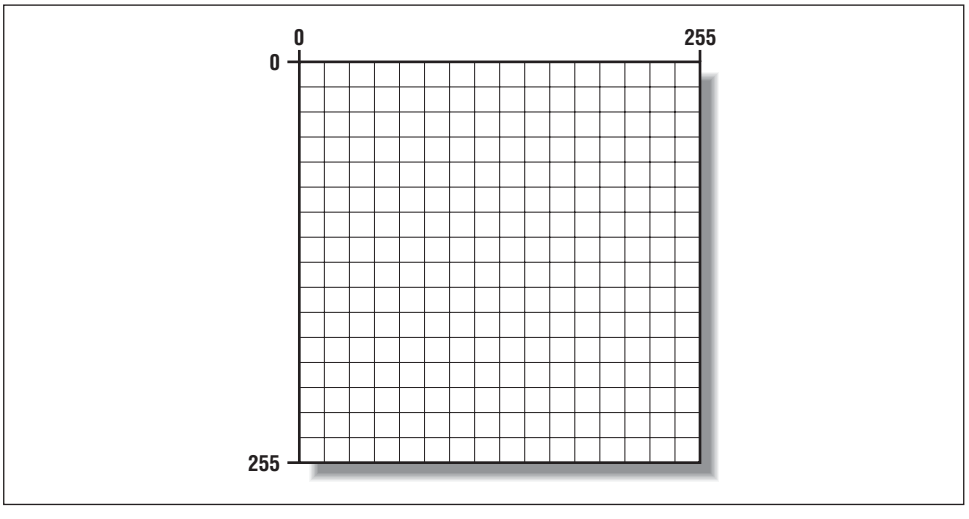


Figure 1-1: 256×256 encoding matrix

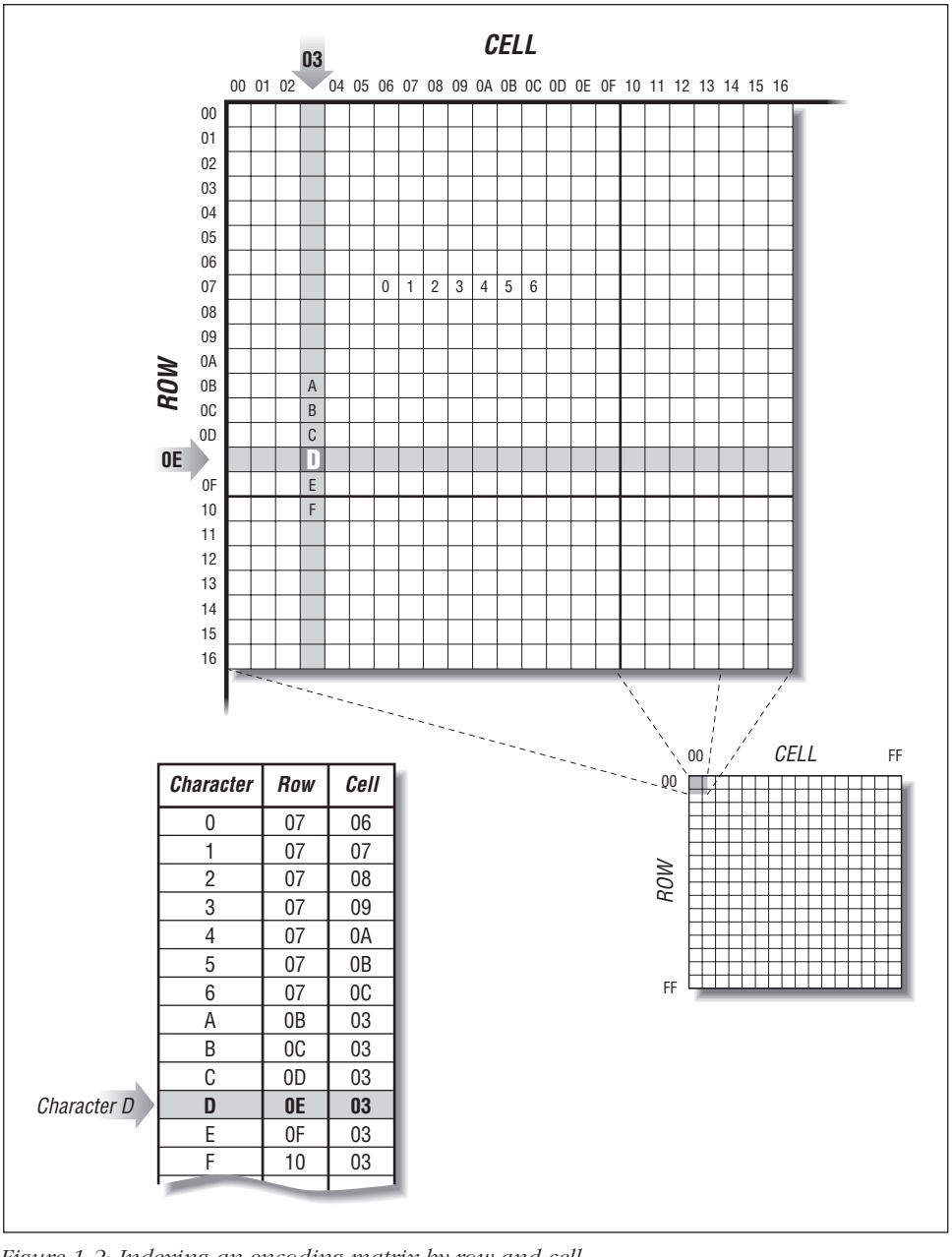
However, not all of these 65,536 cells can be used for representing displayable characters. To enable the mixture of one- and two-byte characters within a single text stream, some characters needed to be reserved as control characters, some of which then serve as the characters that signify when a text stream shifts between one- and two-byte modes. In the case of ISO-2022-JP encoding, the upper limit of displayable characters was set at 8,836, which is the size of the code space made from a 94×94 matrix.*

But why do you need to mix one- and two-byte characters anyway? It is to support existing one-byte encoding standards, such as ASCII, within a two-byte encoding system. One-byte encoding methods are here to stay, and it is still a rather efficient means to encode the characters necessary to write English and many other languages. However, languages with large character sets—those spoken in the CJKV locales—require two bytes to encode characters. A mixed one- and two-byte character stream efficiently represents a mixture of English and Chinese text.

Along with discussions about character sets and encodings, you will encounter the terms "row" and "cell" again and again in this book. These refer to the axes of a matrix used to hold and encode characters. A matrix is composed of rows, and a row is made up of cells. The first byte of the character specifies the row, and the second byte specifies the cell within the row. Figure 1-2 illustrates a matrix and how characters' positions correspond to row and cell values.

---

* *Code space* refers to the area within the (usual) 256×256 encoding matrix that can be used for encoding characters. Most of the figures in Chapter 4 and Appendix D, *Vendor Encoding Methods*, illustrate code spaces that fall within this 256×256 matrix.



Character	Row	Cell
0	07	06
1	07	07
2	07	08
3	07	09
4	07	0A
5	07	0B
6	07	0C
A	0B	03
B	0C	03
C	0D	03
D	**0E**	**03**
E	0F	03
F	10	03

Figure 1-2: Indexing an encoding matrix by row and cell

In an attempt to allow for a mixture of one- and two-byte characters, several CJKV encoding methods have been developed. As you will learn in Chapter 4, these encoding methods are largely, but not completely, compatible. You will also see

that there are encoding methods that use three or even four bytes to represent a single character!

The most common Japanese encoding methods are ISO-2022-JP, Shift-JIS, and EUC-JP. ISO-2022-JP, the most basic, uses seven-bit bytes (or, seven bits of a byte) to represent characters, and requires special characters or sequences of characters (called *shifting characters* or *escape sequences*) to shift between one- and two-byte modes. Shift-JIS and EUC-JP encodings make generous use of eight-bit characters, and use the value of the first byte as the way to distinguish one- and multiple-byte characters.

Input Methods

Those who type English text have the luxury of using keyboards that can hold all the keys to represent a sufficient number of characters. CJKV characters number in the thousands, though, so how does one type CJKV text? Large keyboards that hold thousands of individual keys exist, but they require special training and are difficult to use. This has led to software solutions: *input methods* and *conversion dictionaries*.

Most CJKV text is typically input in two stages:

1. The user types raw keyboard input, which the computer interprets using the input method and the conversion dictionary to display a list of *candidate* characters (*candidate* here refers to the character or characters that are mapped to the input string in the conversion dictionary).
2. The user selects one choice from the list of candidate characters, or requests more choices.

How well each stage is handled on your computer depends greatly on the quality (and vintage) of the input software you are using.

Software called an *input method* handles both of these input stages: it is so named because it grabs the user's keyboard input before any other software can use it (specifically, it is the first software to process keyboard input).

The first stage of input requires keyboard input, and can take one of two usual forms:

- Transliteration using Latin characters (type "k" plus "a" to get か, and so on)
- Native-script input (zhuyin for Chinese [in Taiwan], hiragana for Japanese, hangul for Korean, and so on)

The form used depends on user preference and the type of keyboard in use. For Japanese, the input method converts transliterated Japanese into hiragana on the

fly, so it doesn't really matter which keyboard you are using. In fact, studies show that over 70 percent of Japanese computer users prefer transliterated Japanese input.

Once the input string is complete, it is then parsed in one of two ways: either by the user during input, or by a parser built into the input method. Finally, each segment is run through a conversion process that consists of a lookup into a conversion dictionary. This is very much like a *key-value* lookup. Typical conversion dictionaries have tens of thousands of entries. It seems that the more entries, the better the conversion quality. However, if the conversion dictionary is too large, users are shown a far too lengthy list of candidates. This reduces input efficiency.

Can Chinese characters be input one at a time? While single Chinese-character input is possible, there are three basic units that can be used. These units allow you to limit the number of candidates from which you must choose. Typically, the larger the input unit, the fewer candidates. The units are as follows:

- Single Chinese character
- Chinese character compound
- Chinese character phrase

Early input programs required that each Chinese character be input individually (single Chinese character). Nowadays it is much more efficient to input Chinese characters as they appear in compounds or even phrases. This means that you may input two or more Chinese characters at once by virtue of inputting their combined reading. For example, the Chinese character compound 漢字 (the two Chinese characters for writing the word meaning "Chinese character") can be input as two separate characters, 漢 (read *kan* in Japanese) and 字 (read *ji* in Japanese). Table 1-6 shows the two target Chinese characters, along with other Chinese characters with the same reading.

Table 1-6: Single Chinese Character Input—Japanese

Character	Reading	Chinese Characters with Identical Readings
漢	K A N	乾侃冠寒刊勘勧巻喚堪姦完官寛干幹患感慣憾換敢柑桓棺款歓汗漢澗潅環甘監看竿管簡緩缶翰肝艦莞観諌貫鑑鑑間閑関陥韓館舘
字	J I	事似侍児字寺慈持時次滋治爾璽痔磁示而耳自蒔辞

You can see that there are many other Chinese characters with those readings, so you may have to wade through a long list of candidate Chinese characters before you find the correct one. A more efficient way is to input them as one unit, called a Chinese character compound. This produces a much shorter list of candidates

from which to choose. Table 1-7 illustrates the two Chinese characters input as a compound, along with candidate compounds with the same reading.

Table 1-7: Chinese Character Compound Input—Japanese

Compound	Reading	Compounds with Identical Readings
漢字	K A N J I	漢字 感じ 幹事 監事 完司

Note how the list of Chinese character compounds is much shorter in this case. There is an even higher-level input unit called a Chinese character phrase. This is similar to inputting two or more Chinese characters as a single compound, but adds another element, similar to a preposition in English, that makes the whole string into a phrase. An example of a Chinese character phrase is 漢字は, which means "the Chinese character" in Japanese. Because Chinese-language text is composed solely of hanzi, Chinese character phrase applies only to Japanese, and possibly Korean.

Some of you may know of input software that claims to let you convert whole sentences at once. This is not really true. Such software allows you to input whole sentences, but the sentence is then parsed into smaller units, usually Chinese character phrases, then converted. Inputting whole sentences before any conversion is merely a convenience for the user.

Korean input has some special characteristics that are related to how their most widely used writing system, hangul, is composed. Whether input is by a QWERTY or a Korean keyboard array, Korean input involves the entering of hangul elements called *jamo*. As the jamo are input, the operating system or input software attempts to compose hangul using an automaton. Because of how hangul are composed of jamo, the user may have up to three alternatives for deleting characters:

- Delete entire hangul
- Delete by jamo
- Delete by word

This particular option is specific to Korean, and depends on the input method.

Typography

CJKV text can usually be written or set in one of two orientations: left to right, top to bottom (horizontal setting, as in this book); and top to bottom, right to left (vertical setting). Chapter 7, *Typography*, provides plenty of examples of horizontal versus vertical writing. Vertical writing orientation, more often than not, causes problems with Western-language software. Luckily, it is generally accept-

able to set CJKV text in the same horizontal orientation as most Western languages. Traditional novels and short stories are often set vertically, but technical materials, such as science textbooks and the like, are set horizontally.

Vertically set CJKV text is not a simple matter of changing writing direction. Some characters require special handling, such as a different orientation (90-degree clockwise rotation) or a different position within the *em-square*.* Chapter 7 provides some sample text set both horizontally and vertically, and illustrates some characters that require special treatment.

In addition to the two writing directions for CJKV text, there are other special text formatting considerations, such as special rules for wrapping characters at the ends of lines, special justification, metrics adjustment, and a way to annotate characters.

Basic Concepts and Terminology

Now I'll define some basic concepts which will help carry you through this entire book. These concepts are posed as questions. After all, these are questions you might raise as you read this book. If at any time you encounter a new term, please glance at the glossary toward the back of the book: new terms are included and explained there.

What Are All Those Abbreviations and Acronyms?

Most technical fields are flooded with abbreviations and acronyms, and CJKV information processing is no exception. Some of the more important (and confusing) ones are explained in the following section, but when in doubt, consult Appendix X.

What is the difference between GB and GB/T?

Most references to "GB" mean the GB 2312-80 character set standard, which represents the most widely implemented character set for Chinese.

GB stands for "Guo Biao" (国标 *guóbiāo*), which is short for "Guojia Biaozhun" (国家标准 *guójiā biāozhǔn*), and means "National Standard."

Because GB/T character set standards are traditional analogs of existing GB character set standards, some naturally think that the "T" stands for "Traditional." Yet another myth to blow out of the water. The "T" in "GB/T" actually stands for "Tui" (推 *tuī*), which is short for "Tuijian" (推荐 *tuījiàn*), and means "recommended." It

* The term *em-square* refers to a square-shaped space whose height and width roughly correspond to the width of the letter "M." The term *design space* is actually a more accurate way to represent this typographic concept.

means "recommended" in the sense that it is the opposite of "forced" or "mandatory."

The "K" in GBK (an extension to GB 2312-80) comes from the Chinese word 扩展 (*kuòzhǎn*), which means "extension."

What are JIS, JISC, and JSA? How are they related?

In much of the literature in the field of Japanese information processing, you will quite often see references to JISC, JIS, and JSA. The most common of these is JIS, the least JISC. What these refer to can sometimes be confusing, and is often contradicted in reference works.

JIS stands for *Japanese Industrial Standard* (日本工業規格 *nihon kōgyō kikaku*), the name given to the standards used in Japanese industry.* The character ㊫ is the symbol for JIS. JIS can refer to several things: the character set standards established by JISC, the encoding method specified in these character set standards, and even the keyboard arrays described in JIS manuals. Context should usually make its meaning clear. The term JIS appears frequently in this book.

JISC stands for *Japanese Industrial Standards Committee* (日本工業標準調査会 *nihon kōgyō hyōjun chōsakai*). This is the name of the governing body that establishes JIS standards and publishes manuals through JSA. The committee that develops and writes each JIS manual is composed of people from Japanese industry who have a deep technical background in the topic to be covered by the JIS manual. Committee members are listed at the end of each JIS manual.

JSA stands for *Japanese Standards Association* (日本規格協会 *nihon kikaku kyōkai*). This organization publishes the manuals for the JIS standards established by JISC, and generally oversees the whole process.

JIS is often used as a blanket term covering JIS, JISC, and JSA, but now you know what they *really* mean.

Several JIS "C" series standards changed designation to "X" series standards on March 1, 1987. Table 1-8 lists the JIS standards—mentioned in this book—that changed designation from "C" to "X" series.

Table 1-8: JIS Standard Designation Changes

JIS "C" Series	JIS "X" Series
JIS C 6220	JIS X 0201
JIS C 6228	JIS X 0202
JIS C 6225	JIS X 0207

* There are even JIS standards for manufacturing toilet paper!

Table 1-8: JIS Standard Designation Changes (continued)

JIS "C" Series	JIS "X" Series
JIS C 6226	JIS X 0208
JIS C 6233	JIS X 6002
JIS C 6235	JIS X 6003
JIS C 6236	JIS X 6004
JIS C 6232	JIS X 9051
JIS C 6234	JIS X 9052

Because these changes took place well over a decade ago, they have long been reflected in software and other documentation.

What is KS?

KS simply stands for "Korean Standard" (한국 공업 규격/韓國工業規格 *hangug gongeob gyugyeog*). All Korean character set standard designations begin with "KS." The character ㉾ is the symbol for KS.

All KS standards also include another letter in their designation. Those that are discussed in this book all include the letter "X," which now indicates electric or electronic standards.*

Several KS "C" series standards changed designation to "X" series standards on August 20, 1997. Table 1-9 lists the KS standards—mentioned in this book—that changed designation from the "C" to "X" series.

Table 1-9: KS Standard Designation Changes

KS "C" Series	KS "X" Series
KS C 5601	KS X 1001
KS C 5657	KS X 1002
KS C 5636	KS X 1003
KS C 5620	KS X 1004
KS C 5700	KS X 1005-1
KS C 5861	KS X 2901
KS C 5715	KS X 5002

Because these changes are very recent, it may take years until they are reflected in software and documentation.

* Other letter designations for KS standards include "B" (mechanical), "D" (metallurgy), and "A" (general guidelines).

Are VISCII and VSCII identical? What about TCVN?

While both VISCII and VSCII are short for *Vietnamese Standard Code for Information Interchange*, they represent completely different character sets and encodings. VISCII is defined in RFC 1456, and VSCII is derived from TCVN 5712:1993 (specifically, VN2), which is a Vietnamese national standard. VSCII is also known as ISO IR 180. The differences among VISCII and VSCII are described in Chapter 3, *Character Set Standards*, beginning on page 78. Appendix S, *Single-Byte Code Tables*, provides complete encoding tables for VISCII and VSCII, which better illustrate their differences.

TCVN stands for *Tiêu Chuẩn Việt Nam*, which means "Vietnamese Standard" in Vietnamese. Like GB, JIS, and KS, it represents the first portion of Vietnamese standard designations.

What Are Internationalization and Localization?

Internationalization (often abbreviated as I18N—the initial letter "I" followed by the middle 18 letters followed by the final letter "N") is a blanket term referring to the process of preparing software so that it can be used by more than one culture, region, or locale.* Localization (often abbreviated as L10N) is the process of adapting software to one specific culture, region, or locale. Japanization (often abbreviated as J10N) is a specific instance of L10N. While this book does not necessarily address all of these issues, you will find information pertinent to internationalization and localization.

Either way, I18N or L10N are often desired by users because they provide menus and documentation written in the language of the target locale. They often require special character set handling because so many non-Latin character sets require more than one byte to represent all their characters.

What Are the Multilingual and Locale Models?

There are two basic models for internationalization: the *locale model* and the *multilingual model*. The locale model implements a set of attributes for specific locales. The user must explicitly switch from one locale to another. The character sets implemented by the locale model are specific to a given culture, region, or locale.

The multilingual model goes one step further by not requiring you to flip between locales—multilingual systems use a character set that contains all the characters

* Quiz time. Guess what CJKV6N, C10N, G11N, K11N, M17N, S32S, and V12N stand for?

necessary for several cultures or regions. But still, there are cases when it is impossible to correctly render characters without knowing the target locale.

What Is Row-Cell?

Row-Cell is the translated form of the Japanese word 区点 (*kuten*), which literally means "ward [and] point" (or, more intuitively, "row [and] cell").* This idea serves as an encoding-independent method for referring to characters in CJKV character set standards. A Row-Cell code usually consists of four decimal digits—the "Row" portion consists of a two-digit number with a range from 1 to 94; likewise, the "Cell" portion also consists of a two-digit number with a range from 1 to 94. For example, the first character in most CJKV character set standards is 01-01 in Row-Cell notation, and is more often than not a "space" character.

When I provide lists of characters throughout this book, I usually include Row-Cell codes. These are useful for future reference of these data (so that you don't have to hunt for the codes yourself!).

Characters and Glyphs—What Is the Difference?

Now here's a topic that is usually beaten to death! The term "character" is an abstract notion indicating a class of shapes declared to have the same meaning or form. The term "glyph" is a specific instance of a character. Sometimes, more than one character can constitute a single glyph, such as the two characters f and i, which can be fused together as the single entity "fi." This glyph "fi" is called a *ligature*. The dollar sign is a good example of a character with several glyphs. There are at least four glyphs for the dollar sign, listed as follows:

- An "S" shape with a single vertical bar: $
- An "S" shape with a single broken vertical bar: $
- An "S" shape with two vertical bars: $
- An "S" shape with two broken vertical bars: $

The differences among these four glyphs are minor, but you cannot deny that they still represent the same character, specifically the "dollar sign." Quite often you see a difference in glyph as a difference in typeface. However, there are some characters that have a dozen or more variant forms. Consider the kanji 辺 (*hen*, used in the Japanese family name 渡辺 *watanabe*), which has only two variant forms that are generally available in Japanese fonts (including Unicode-based fonts): 邊 and

* In Chinese, Row-Cell is expressed as 区位 (*qūwèi*); in Korean as 행렬/行列 (*haengryeol*). Note that if the "Cell" portion of "Row-Cell" is in isolation in Korean that it is expressed instead with the hangul 열 (*yeol*), not 렬 (*ryeol*).

邉. DTP center Biblos, a developer of "Gaiji" fonts, offers fonts that provide the following eight additional variant forms of this kanji:

邊 邊 邊 邉 邉 邉 邉 邉

Enfour Media, another developer of "Gaiji" fonts, offers fonts that provide the following 18 variant forms of this kanji:

邊 邊 邊 邊 邊 邊 邊 邊 邉 邉 邉 邉 邉 邉 邉 邉 邉 邉

Clearly, these variant forms all appear to represent that same character, but are simply different glyphs.

You will find that CJKV character set standards do not define the glyph for the characters contained within their pages. Unfortunately (or, fortunately, as the case may be), many think that the glyphs that appear in these manuals are the official ones. Note, however, that the official Jōyō Kanji Table *does* define the glyph shape, at least for the 1,945 kanji contained within. JSA published two manuals that do, in fact, define glyph shapes: JIS X 9051-1984* and JIS X 9052-1983.† They were designed for the JIS X 0208-1983 standard. However, these glyphs have not been widely accepted in industry. It seems as though JSA has no intention of ever revising these documents—this may be their way of not enforcing glyphs.

The one Japanese organization that had a chance in establishing a definitive Japanese glyph standard in Japan is called FDPC, which is short for Font Development and Promotion Center (文字フォント開発・普及センター *moji fonto kaihatsu fukyū sentā*). FDPC was a MITI- (Ministry of International Trade and Industry—通商産業省 *tsūshō sangyō shō*) funded organization, and has since been folded in with JSA. This government organization, with the help of developing members, developed a series of Japanese outline fonts called "Heisei" (平成 *heisei*) typefaces. The first two Heisei typefaces that were released are Heisei Mincho W3 (平成明朝W3 *heisei minchō W3*) and Heisei Kaku (squared) Gothic W5 (平成角ゴシックW5 *heisei kaku goshikku W5*). In fact, the standard Japanese typeface used in the production of this book is Heisei Mincho W3. A total of seven weights of both designs have been produced, weights 3 (W3) through 9 (W9). Two weights of Heisei Maru (rounded) Gothic (平成丸ゴシック *heisei maru goshikku*), 4 and 8, have also been developed. The Heisei typefaces have become somewhat commonplace in the Japanese market.

China takes glyph issues *very* seriously, and expended the effort to develop a series of standards, published in a single manual entitled *32×32 Dot Matrix Font Set and Data Set of Chinese Ideograms for Information Interchange* (信息交换用汉字32×32点阵字模集及数据集 *xìnxī jiāohuàn yòng hànzì 32×32 diǎnzhèn zìmújí*

* Previously designated JIS C 6232-1984.

† Previously designated JIS C 6234-1983.

jí shújùjí), that explicitly define glyphs for the GB 2312-80 character set standard in various typeface styles. These standards are listed in Table 1-10.

Table 1-10: Chinese Glyph Standards

Standard	Page Numbers in Manual	Title (in English)
GB 6345.1-86	1–27	*32×32 Dot Matrix Font Set of Chinese Ideograms for Information Interchange*
GB 6345.2-86	28–31	*32×32 Dot Matrix Font Data Set of Chinese Ideograms for Information Interchange*
GB 12034-89	32–55	*32×32 Dot Matrix Fangsongti Font Set and Data Set of Chinese Ideograms for Information Interchange*
GB 12035-89	56–79	*32×32 Dot Matrix Kaiti Font Set and Data Set of Chinese Ideograms for Information Interchange*
GB 12036-89	80–103	*32×32 Dot Matrix Heiti Font Set and Data Set of Chinese Ideograms for Information Interchange*

Songti (specified in GB 6345.1-86), Fangsongti, Kaiti, and Heiti are the most common typeface styles used in Chinese. When the number of available pixels is reduced, it is impossible to completely represent all of a Chinese character's strokes. These standards are useful because they establish bitmapped patterns that offer a compromise between accuracy and legibility. The recent GB 16794.1-1997 standard (信息技术—通用多八位编码字符集48点阵字形 *xīnxì jìshù—tōngyòng duōbāwèi biānmǎ zìfújí 48 diǎnzhèn zìxíng*) is similar to the GB standards listed in Table 1-10, but covers the complete GBK character set and provides 48×48 bitmapped patterns. An older set of GB standards, GB 5007.1-85 (信息交换用汉字24×24点阵字模集 *xìnxī jiāohuàn yòng hànzì 24×24 diǎnzhèn zìmújí*) and GB 5007.2-85 (信息交换用汉字24×24点阵字模数据集 *xìnxī jiāohuàn yòng hànzì 24×24 diǎnzhèn zìmú shújùjí*), provided 24×24 bitmapped patterns for a single design.

Exactly how character and glyph are defined can differ depending on the source. Table 1-11 provides the ISO (International Organization for Standardization) and The Unicode Consortium definitions for the terms character, glyph, and glyph image.

What Is the Difference Between Typeface and Font?

The term *typeface* refers to the printed style of a glyph or character set. A *font*, on the other hand, refers to a single instance of a typeface, such as a specific point size. This is why the commonly used term *outline font* is a misnomer—the outlines are scalable, which means that they are not specific to any one point size. A better term is *outline font instance*.

Table 1-11: Character, Glyph, and Glyph Image Definitions—ISO and Unicode

Terminology	ISO	Unicode[a]
Character	A member of a set of elements used for the organisation, control, or representation of data.[b] An atom of information with an individual meaning, defined by a character repertoire.[c]	(1) The smallest component of written language that has semantic value; refers to the 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.
Glyph	A recognizable abstract graphical symbol which is independent of any specific design.[c]	(1) An abstract form that represents one or more glyph images. (2) A synonym for *glyph image*.
Glyph image	An image of a glyph, as obtained from a glyph representation displayed on a presentation surface.[c]	The actual, concrete image of a glyph representation having been rasterized or otherwise imaged onto some display surface.

[a] *The Unicode Standard, Version 2.0* (Addison-Wesley, 1996).
[b] ISO 10646-1:1993.
[c] ISO 9541-1:1991.

Western typography commonly uses serif, sans serif, and script typeface styles. Table 1-12 lists the common CJKV typeface styles, along with correspondences across locales.

Table 1-12: CJKV Typeface Styles

Western	Chinese[a]	Japanese	Korean
Serif	Song (宋体 *sòngtǐ*)	Mincho (明朝体 *minchōtai*)	Myeongjo (명조체/明朝體 *myeongjoce*)
Sans serif	Hei (黑体 *hēitǐ*)	Gothic (ゴシック体 *goshikkutai*)	Gothic (고딕체/고딕體 *godigce*)
Script	Kai (楷体 *kǎitǐ*)	Kaisho (楷書体 *kaishotai*) Gyosho (行書体 *gyōshotai*) Sosho (草書体 *sōshotai*)	Haeseo (해서체/楷書體 *haeseoce*) Haengseo (행서체/行書體 *haengseoce*) Choseo (초서체/草書體 *coseoce*)
Other	Fangsong (仿宋体 *fǎngsòngtǐ*)	Kyokasho (教科書体 *kyōkashotai*)	

[a] Replace 体 with 體 in these typeface style names for Traditional Chinese.

Table 1-12 by no means constitutes a complete list of CJKV typeface styles—there are numerous typeface styles for hangul, for example.

What Are Half- and Full-Width Characters?

The terms half- and full-width refer to the relative glyph size of characters. These terms are referred to as *hankaku* (半角 *hankaku*) and *zenkaku* (全角 *zenkaku*), respectively, in Japanese.* Half-width is relative to full-width. Full-width refers to the glyph size of standard CJKV characters, such as zhuyin, kana, hangul, and Chinese characters. Latin characters, which appear to take up approximately half the display width of CJKV characters, are considered to be half-width by this standard. The very first Japanese characters to be processed on computer systems were half-width katakana. They have the same approximate display width as Latin characters. There are now full-width Latin and katakana characters. Table 1-13 shows the difference in display width between half- and full-width characters (the katakana character used as the example is read *ka*).

Table 1-13: Half- and Full-Width Characters

	Katakana	Latin
Half-width	ｶｶｶｶｶ	12345
Full-width	カカカカカ	１２３４５

As you can see, full-width characters occupy twice the display width as their half-width versions. At one point in time there was a clear-cut relationship between display width of a glyph and number of bytes used to encode it (the encoding length)—the number of bytes simply determined the display width. Half-width katakana characters were originally encoded with one byte. Full-width ones were encoded with two bytes. Now that there is a much richer choice of encoding methods available, this relationship no longer holds true. Table 1-14 lists several popular encoding methods, along with the number of bytes required to represent half- and full-width characters.

Table 1-14: Half- and Full-Width Character Representations

	ASCII	ISO-2022-JP	Shift-JIS	EUC-JP	ISO 10646-1:1993
Full-width					
Katakana	…	2 bytes	2 bytes	2 bytes	2 or 4 bytes
Latin	…	2 bytes	2 bytes	2 bytes	2 or 4 bytes
Half-width					
Katakana	…	1 byte	1 byte	2 bytes	2 or 4 bytes
Latin	1 byte	1 byte	1 byte	1 byte	2 or 4 bytes

* In Chinese, these terms are 半形 (*bànxíng*) and 全形 (*quánxíng*), respectively. In Korean, perhaps 반각/半角 (*bangag*) and 전각/全角 (*jeongag*), respectively.

Latin Versus Roman Characters

Many people debate whether the 26 letters of the English alphabet should be referred to as *Roman* or *Latin* characters. While some standards, such as those published by ISO, prefer the term Latin, others prefer the term Roman. This book will prefer the term Latin over Roman. Readers of this book should treat both terms synonymously.

When speaking of typeface designs, the use of the term Roman is used in contrast with the term italic.

What Is Notation?

The term notation refers to a method of representing units. A given distance, whether expressed in miles or kilometers, is, after all, the same distance. In computer science, common notations for representing the value of bytes are listed in Table 1-15, and all correspond to a different numeric base.

Table 1-15: Decimal 100 in Common Notations

Notation	Base	Range	Example
Binary	2	0 and 1	`01100100`
Octal	8	0–7	`144`
Decimal	10	0–9	`100`
Hexadecimal	16	0–9 and A–F	`64`

While the numbers in the "Example" column all have the same underlying value, they have been expressed using different notations, and thus take on a different form. Most people (that is, non-nerds) think in decimal notation; however, computers (and some nerds) process information using binary notation (as discussed above, computers process bits, which have two possible values). Below you will find that hexadecimal notation does, however, have distinct advantages when dealing with computers.

What Is an Octet?

We have already discussed the terms bits and bytes. But what about the term octet? At a glance, you can tell it has something to do with the number eight. An octet represents eight bits, and is an eight-bit byte. This becomes confusing when dealing with 16-bit encodings. 16 bits can be broken down into two eight-bit bytes, or two octets. 32 bits, likewise, can be broken down into four eight-bit bytes, or four octets.

Given 16 bits in a row:

```
0110010001011111
```

This string of bits can be broken down into two eight-bit units, specifically octets (bytes):

```
01100100
01011111
```

The first eight bits represent 100 (0x64), and the second 95 (0x5F). All 16 bits together as one unit are usually equal to 25695 in decimal or 0x645F in hexadecimal—it may be different depending on a computer's specific architecture. Divide 25695 by 256 to get the first byte's value as a decimal octet, which results in 100 in this case—the remainder from this division is the value of the second byte, which, in this case, is 95. Table 1-16 lists representations of two octets (bytes) and their 16-bit unit equivalent. This is done for you in different notations.

Table 1-16: Octets and 16-Bit Units in Various Notations

Notation	First Octet	Second Octet	16-Bit Unit
Binary	01100100	01011111	0110010001011111
Octal	144	137	62137
Decimal	100	95	25695
Hexadecimal	64	5F	645F

Note how going from two octets to a 16-bit unit is a simple matter of concatenation in the case of binary and hexadecimal notation. Not so with decimal notation, which requires multiplication of the first octet by 256, then addition of the second octet. The ease of going between different representations (octets versus 16-bit units) depends on the notation that you are using. Of course, string concatenation is easier than two mathematical processes. This is why hexadecimal is used so frequently in computers.

In some cases, the order in which byte concatenation takes place matters, such as when the byte order (endianness) differs depending on the underlying computing architecture. Guess what the next section is about?

What Are Little and Big Endian?

There are two basic computer architectures when it comes to the issue of byte order: little endian and big endian. That is, the order in which the bytes of multiple-byte storage units (such as integers, floats, doubles, and so on) appear.* One-byte storage units, such as char, do not need this special treatment (that is,

* A derivation of little and big endian came from *Gulliver's Travels*, in which there were civil wars fought over which end of a boiled egg to crack.

unless your particular machine or implementation represents them with more than one byte).

- Little endian machines use computing architectures supported by Vax and Intel processors. This typically means that MS-DOS and Windows machines are little endian.
- Big endian machines use computing architectures supported by Motorola and Sun processors. This typically means MacOS and most Unix workstations. Big endian is also known as "network byte order."

Table 1-17 provides an example two-byte value as encoded on little and big endian machines.

Table 1-17: Little and Big Endian Representation

Notation	High Byte	Low Byte	Little Endian	Big Endian
Binary	01100100	01011111	0101111101100100	0110010001011111
Hexadecimal	64	5F	5F64	645F

A four-byte example, such as 0x64, 0x5F, 0x7E, and 0xA1, becomes 0xA17E5F64 on little endian machines, and 0x645F7EA1 on big endian machines. Note how the bytes themselves (not the underlying bits of each byte) are reversed depending on endianness. This is precisely why endianness is also referred to as byte order. The term endian is used to describe what impact the byte at the end has on the overall value. The Unicode value for a "space" character is 0x0020 for big-endian, and 0x2000 for little-endian.

Now that you understand the concept of endianness, the real question that needs answering is when endianness matters. Keep reading…

What Are Multiple-Byte and Wide Characters?

If you have ever read comprehensive books and materials about ANSI C, you more than likely came across the terms multiple-byte and wide characters. Those documents don't do those terms justice. Here you'll get a definitive answer.

When dealing with encodings that are processed on a per-byte basis, endianness is irrelevant. These encodings support what are known as *multiple-byte* characters. So, what encodings are these? Table 1-18 provides an incomplete yet informative list of these encodings.

There are some encodings that require special endian treatment, and cannot be treated on a per-byte basis. These encodings include what are known as *wide* characters, and almost always provide a facility for indicating the byte order. Table 1-19 lists some encodings that use wide characters.

Table 1-18: Multiple-Byte Character Encodings

Encoding	Encoding Length	Locale
ASCII	one-byte	*not applicable*
ISO-2022	one- and two-byte	CJKV
EUC	one- through four-byte, depending on locale	CJKV
GBK	one- and two-byte	China
Big Five	one- and two-byte	Taiwan
Big Five Plus	one- and two-byte	Taiwan
Shift-JIS	one- and two-byte	Japan
Johab	one- and two-byte	Korea
UHC	one- and two-byte	Korea
UTF-8	one- through six-byte	*not applicable*

Table 1-19: Wide Character Encodings

Encoding	Encoding Length
UCS-2	16-bit fixed
UCS-4	32-bit fixed
UTF-16	16-bit variable-length
Unicode Version 2.0	*Same as UTF-16*

It is with endianness that we can more easily distinguish multiple-byte from wide characters. Multiple-byte characters have the same byte order regardless of the underlying processor architecture; the byte order of wide characters is determined by the underlying processor architecture.

In this chapter:

- *Latin Characters and Transliteration*
- *Zhuyin*
- *Kana*
- *Hangul*
- *Chinese Characters*
- *Non-Chinese Chinese Characters*

2

Writing Systems

Now that you have had a taste of what to expect to learn about CJKV information processing, let's begin with a thorough description of the various CJKV writing systems. We've already touched briefly upon this subject in the introductory material, but you need to learn a bit more. After reading this chapter you should have an understanding of the types of characters used to write CJKV text, specifically the following:

- Latin characters
- Zhuyin
- Kana (*hiragana* and *katakana*)
- Hangul (and jamo)
- Chinese characters
- Non-Chinese Chinese characters (Japanese *kokuji*, Korean *gugja*, and Vietnamese *chữ Nôm*)

Each of these types of characters exhibits its own special characteristics, and often has locale-specific usages. This information is absolutely crucial for understanding discussions elsewhere in this book.

Latin Characters and Transliteration

Latin characters (拉丁字母 *lādīng zìmǔ* in Chinese, ラテン文字 *raten moji* or ローマ字 *rōmaji* in Japanese, 로마자 *romaja* in Korean, and Quốc ngữ/國語 in Vietnamese) used in CJKV texts are the same as those used in Western texts, specifically the 52 upper- and lowercase letters of the Latin alphabet, sometimes decorated with accents to indicate length or tone. Also included are the ten

numerals 0 through 9. Accented characters, usually vowels, are often required for transliteration purposes. Table 2-1 lists the basic set of Latin characters.

Table 2-1: Latin Characters

Lowercase	abcdefghijklmnopqrstuvwxyz
Uppercase	ABCDEFGHIJKLMNOPQRSTUVWXYZ
Numerals	0123456789

There is really nothing special about these characters. Latin characters are most often used in tables (numerals), in abbreviations (alphabet), or for transcription or transliteration purposes (sometimes with accented characters).

Commonly used transliteration systems for CJKV text that use characters beyond the standard set of Latin characters illustrated above include Pinyin (Chinese), Hepburn (Japanese), Kunrei (Japanese), and Ministry of Education (Korean). These and other CJKV transliteration systems are covered in the following sections.

Chinese Transliteration Methods

Chinese uses two primary transliteration methods: Pinyin (拼音 *pīnyīn*) and Wade-Giles (韋氏 *wéishì*). There is also the Yale method, which is not covered in this book. While there are many similarities between these two transliteration methods, they mainly differ in where they are used. Pinyin is used in China, while Wade-Giles is popular in Taiwan. Historically speaking, Wade-Giles was the originally recognized Chinese transliteration system during the nineteenth century.

Table 2-2 lists the consonant sounds as transliterated by Pinyin and Wade-Giles—zhuyin symbols (described later in this chapter) are included for the purpose of cross-reference.

Table 2-2: Chinese Transliteration—Consonants

Zhuyin	Pinyin	Wade-Giles
ㄅ	B	P
ㄆ	P	P′
ㄇ	M	M
ㄈ	F	F
ㄉ	D	T
ㄊ	T	T′
ㄋ	N	N
ㄌ	L	L
ㄍ	G	K
ㄎ	K	K′

Table 2-2: Chinese Transliteration—Consonants (continued)

Zhuyin	Pinyin	Wade-Giles
ㄏ	H	H
ㄐ	J	CH[a]
ㄑ	Q	CH'[a]
ㄒ	X	HS[a]
ㄓ	ZH	CH
ㄔ	CH	CH'
ㄕ	SH	SH
ㄖ	R	J
ㄗ	Z	TS
ㄘ	C	TS'
ㄙ	S	S

[a] Only before i or ü.

Table 2-3 lists the vowel sounds as transliterated by Pinyin—zhuyin are again included for reference. Note that this table is constructed as a matrix that indicates what zhuyin vowel combinations are possible and how they are transliterated—the two axes themselves serve to indicate the transliterations for single zhuyin vowels.

Table 2-3: Chinese Transliteration—Vowels

	ㄧ I	ㄨ U	ㄩ Ü
ㄚ A	ㄧㄚ IA	ㄨㄚ UA	
ㄛ O		ㄨㄛ UO	
ㄜ E	ㄧㄝ IE		ㄩㄝ ÜE
ㄞ AI		ㄨㄞ UAI	
ㄟ EI		ㄨㄟ UEI	
ㄠ AO	ㄧㄠ IAO		
ㄡ OU	ㄧㄡ IOU		
ㄢ AN	ㄧㄢ IAN	ㄨㄢ UAN	ㄩㄢ ÜAN
ㄣ EN	ㄧㄣ IN	ㄨㄣ UEN	ㄩㄣ ÜN
ㄤ ANG	ㄧㄤ IANG	ㄨㄤ UANG	
ㄥ ENG	ㄧㄥ ING	ㄨㄥ UENG or ONG	ㄩㄥ IONG

The zhuyin character ㄦ, which deserves separate treatment, is usually transliterated *er*.

It is sometimes necessary to use an apostrophe to separate the Pinyin readings of individual hanzi when the result can be ambiguous. Consider the transliterations

for the words 先 and 西安, which are *xiān* and *xī'ān*, respectively. Note the use of the apostrophe.

More details about the zhuyin characters themselves appear later in this chapter, starting on page 40. Po-Han Lin (林伯翰 *lín bóhàn*) has developed a Java applet that can convert between the Pinyin, Wade-Giles, and Yale transliteration systems.* He also provides additional details about Chinese transliteration.†

Chinese tone marks

Also of interest is how tone marks are rendered when transliterating Chinese text. Basically, there are two systems for indicating tone. One system, which requires the use of special fonts, employs diacritic marks that serve to indicate tone. The other system uses the numerals 1 through 4 immediately after each hanzi transliteration—no special fonts are required. Pinyin transliteration generally uses diacritic marks, but Wade-Giles uses numerals.

Table 2-4 lists the names of the Chinese tone marks, along with an example hanzi for each. Note that there are cases in which there is no tone required.

Table 2-4: Chinese Tone Mark Examples

Tone	Tone Name	Number[a]	Example	Meaning
None	轻声/輕聲 (*qīngshēng*)	*none*	ma (吗)	question particle
Flat	阴平/陰平 (*yīnpíng*)	1	ma1 or mā (妈)	*mother*
Rising	阳平/陽平 (*yángpíng*)	2	ma2 or má (麻)	*hemp, flax*
Falling-Rising	上声/上聲 (*shǎngshēng*)	3	ma3 or mǎ (马)	*horse*
Falling	去声/去聲 (*qùshēng*)	4	ma4 or mà (骂)	*cursing, swearing*

[a] Microsoft's pinyin input method uses the numeral 5 to indicate no tone.

It is also common to find reference works in which Pinyin readings have no tone marks at all, That is, no numerals and no diacritic marks. I have observed that tone marks can be effectively omitted when the corresponding hanzi are in proximity, such as on the same page—the hanzi themselves can be used to remove any ambiguity that arises from no indication of tones. Pinyin readings provided throughout this book use diacritic marks to indicate tone.

Japanese Transliteration Methods

There are four Japanese transliteration systems worth exploring in the context of this book:

---

* *http://www.edepot.com/java.html*

† *http://www.edepot.com/taoroman.html*

- The Hepburn system (ヘボン式 *hebon shiki*), developed by James Curtis Hepburn, an American missionary, in 1886—this is considered the most widely used system
- The Kunrei system (訓令式 *kunrei shiki*), developed in 1937, is considered the official transliteration system by the Japanese government
- The Nippon system (日本式 *nippon shiki*), developed in 1881 by Aikitsu Tanakadate (田中館愛橘 *tanakadate aikitsu*)—nearly identical to the Kunrei system, but the least used
- The word processor system (ワープロ式 *wāpuro shiki*) has been developed in a somewhat *ad hoc* fashion over recent years by Japanese word processor and input method manufacturers

The Japanese transliterations in this book adhere to the Hepburn system. Because the word processor system allows for a wide variety of transliteration possibilities (that is the nature of input systems), it is a topic of discussion in Chapter 5, *Input Methods*.

Table 2-5 lists the basic kana syllables (shown here and in other tables of this section using hiragana), transliterated according to the three transliteration systems. Those that are transliterated differently in the three systems have been highlighted for easy differentiation. Table 2-18 on page 43 provides similar information, but presented in a different manner.

Table 2-5: Single Syllable Japanese Transliteration

Kana	Hepburn	Kunrei	Nippon
あ	A	A	A
い	I	I	I
う	U	U	U
え	E	E	E
お	O	O	O
か	KA	KA	KA
が	GA	GA	GA
き	KI	KI	KI
ぎ	GI	GI	GI
く	KU	KU	KU
ぐ	GU	GU	GU
け	KE	KE	KE
げ	GE	GE	GE
こ	KO	KO	KO
ご	GO	GO	GO

Table 2-5: Single Syllable Japanese Transliteration (continued)

Kana	Hepburn	Kunrei	Nippon
さ	SA	SA	SA
ざ	ZA	ZA	ZA
し	SHI	SI	SI
じ	JI	ZI	ZI
す	SU	SU	SU
ず	ZU	ZU	ZU
せ	SE	SE	SE
ぜ	ZE	ZE	ZE
そ	SO	SO	SO
ぞ	ZO	ZO	ZO
た	TA	TA	TA
だ	DA	DA	DA
ち	CHI	TI	TI
ぢ	JI	ZI	DI
つ	TSU	TU	TU
づ	ZU	ZU	DU
て	TE	TE	TE
で	DE	DE	DE
と	TO	TO	TO
ど	DO	DO	DO
な	NA	NA	NA
に	NI	NI	NI
ぬ	NU	NU	NU
ね	NE	NE	NE
の	NO	NO	NO
は	HA	HA	HA
ば	BA	BA	BA
ぱ	PA	PA	PA
ひ	HI	HI	HI
び	BI	BI	BI
ぴ	PI	PI	PI
ふ	FU	HU	HU
ぶ	BU	BU	BU
ぷ	PU	PU	PU
へ	HE	HE	HE
べ	BE	BE	BE

Table 2-5: Single Syllable Japanese Transliteration (continued)

Kana	Hepburn	Kunrei	Nippon
ぺ	PE	PE	PE
ほ	HO	HO	HO
ぼ	BO	BO	BO
ぽ	PO	PO	PO
ま	MA	MA	MA
み	MI	MI	MI
む	MU	MU	MU
め	ME	ME	ME
も	MO	MO	MO
や	YA	YA	YA
ゆ	YU	YU	YU
よ	YO	YO	YO
ら	RA	RA	RA
り	RI	RI	RI
る	RU	RU	RU
れ	RE	RE	RE
ろ	RO	RO	RO
わ	WA	WA	WA
ゐ	WI	WI	WI
ゑ	WE	WE	WE
を	O	O	WO
ん	N or M[a]	N	N

[a] An *m* was once used before the consonants *b*, *p*, or *m*—an *n* is now used in all contexts.

Table 2-6 lists what are considered to be the palatalized syllables—although they represent a single syllable, they are represented with two kana characters. Those that are different in the three transliteration systems are highlighted.

Table 2-6: Japanese Transliteration—Palatalized Syllables

Kana	Hepburn	Kunrei	Nippon
きゃ	KYA	KYA	KYA
ぎゃ	GYA	GYA	GYA
きゅ	KYU	KYU	KYU
ぎゅ	GYU	GYU	GYU
きょ	KYO	KYO	KYO
ぎょ	GYO	GYO	GYO

Table 2-6: Japanese Transliteration—Palatalized Syllables (continued)

Kana	Hepburn	Kunrei	Nippon
しゃ	SHA	SYA	SYA
じゃ	JA	ZYA	ZYA
しゅ	SHU	SYU	SYU
じゅ	JU	ZYU	ZYU
しょ	SHO	SYO	SYO
じょ	JO	ZYO	ZYO
ちゃ	CHA	TYA	TYA
ぢゃ	JA	ZYA	DYA
ちゅ	CHU	TYU	TYU
ぢゅ	JU	ZYU	DYU
ちょ	CHO	TYO	TYO
ぢょ	JO	ZYO	DYO
にゃ	NYA	NYA	NYA
にゅ	NYU	NYU	NYU
にょ	NYO	NYO	NYO
みゃ	MYA	MYA	MYA
みゅ	MYU	MYU	MYU
みょ	MYO	MYO	MYO
ひゃ	HYA	HYA	HYA
びゃ	BYA	BYA	BYA
ぴゃ	PYA	PYA	PYA
ひゅ	HYU	HYU	HYU
びゅ	BYU	BYU	BYU
ぴゅ	PYU	PYU	PYU
ひょ	HYO	HYO	HYO
びょ	BYO	BYO	BYO
ぴょ	PYO	PYO	PYO
りゃ	RYA	RYA	RYA
りゅ	RYU	RYU	RYU
りょ	RYO	RYO	RYO

Table 2-7 lists what are considered to be long (or doubled) vowels—the first five rows are hiragana, and the last five are katakana. Note that only the long hiragana *i* (いい, expressed as *ii*) is common to all three systems, and that the Kunrei and Nippon systems are identical in this regard.

The only difference among these systems' long vowel transliterations is the use of a macron (Hepburn) versus a circumflex (Kunrei and Nippon). Almost all Latin

Table 2-7: Japanese Transliteration—Long Vowels

Kana	Hepburn	Kunrei	Nippon
ああ	Ā	Â	Â
いい	II	II	II
うう	Ū	Û	Û
ええ	Ē	Ê	Ê
えい	EI	EI	EI
おう	Ō	Ô	Ô
アー	Ā	Â	Â
イー	Ī	Î	Î
ウー	Ū	Û	Û
エー	Ē	Ê	Ê
オー	Ō	Ô	Ô

fonts include circumflexed vowels, but those with macroned vowels are still extremely rare.

Finally, Table 2-8 shows some examples of how to transliterate Japanese double consonants, all of which use a small つ or ツ (*tsu*).

Table 2-8: Japanese Transliteration—Double Consonants

Example	Transliteration
かっこ	*kakko*
いっしょ	*issho*
ふっそ	*fusso*
ねっちゅう	*netchū*
しって	*shitte*
ビット	*bitto*
ベッド	*beddo*
バッハ	*babha*

Korean Transliteration Methods

There are three generally accepted transliteration methods for Korean text: Ministry of Education (문교부/文敎部 *mungyobu*; derived from McCune-Reischauer), established on January 13, 1984; Korean Language Society (한글학회/한글學會 *hangeul haghoe*), established on February 21, 1984;* and ISO/TR 11941:1996 (*Information Documentation—Transliteration of Korean Script into Latin Characters*), estab-

* *http://www.hangeul.or.kr/*

lished in 1996. The Korean text in this book adheres to the ISO/TR 11941:1996 transliteration method, specifically Method 2.* Other transliteration methods, not covered in this book, include the Yale, Lukoff, and Horne methods.

Table 2-9 lists the jamo that represent consonants, along with their representation in these three transliteration methods. Also included are the ISO/TR 11941:1996 transliterations when these jamo serve as the final consonant of a syllable. ISO/TR 11941:1996 Method 1 is used for North Korea (DPRK), and Method 2 is used for South Korea (ROK). Uppercase is used solely for clarity.

Table 2-9: Korean Transliteration—Consonants

Jamo	MOE	KLS	ISO (DPRK)	Final	ISO (ROK)	Final
ㄱ	K/G	G	K	K	G	G
ㄴ	N	N	N	N	N	N
ㄷ	T/D	D	T	T	D	D
ㄹ	R/L	L	R	L	R	L
ㅁ	M	M	M	M	M	M
ㅂ	P/B	B	P	P	B	B
ㅅ	S/SH	S	S	S	S	S
ㅇ	*none*/NG	*none*/NG	*none*	NG	*none*	NG
ㅈ	CH/J	J	C	C	J	J
ㅊ	CH′	CH	CH	CH	C	C
ㅋ	K′	K	KH	KH	K	K
ㅌ	T′	T	TH	TH	T	T
ㅍ	P′	P	PH	PH	P	P
ㅎ	H	H	H	H	H	H
ㄲ	KK	GG	KK	KK	GG	GG
ㄸ	TT	DD	TT	*n/a*	DD	*n/a*
ㅃ	PP	BB	PP	*n/a*	BB	*n/a*
ㅆ	SS	SS	SS	SS	SS	SS
ㅉ	TCH	JJ	CC	*n/a*	JJ	*n/a*

Note that some of the double jamo do not occur at the end of syllables. Also, some of these transliteration methods, most notably the Ministry of Education system, have a number of rules that dictate how to transliterate certain jamo depending on their context. The ㄱ jamo, for example, is transliterated as *k* in most contexts, and *g* when between vowels.

* Notable exceptions include words such as *hangul*, which should really be transliterated as *hangeul*.

ISO/TR 11941:1996 also defines transliterations for compound consonant jamo that appear only at the end of hangul syllables, all of which are listed in Table 2-10.

Table 2-10: ISO/TR 11941:1996 Compound Jamo Transliteration

Jamo	DPRK	ROK
ㄳ	KS	GS
ㄵ	NJ	NJ
ㄶ	NH	NH
ㄺ	LK	LG
ㄻ	LM	LM
ㄼ	LP	LB
ㄽ	LS	LS
ㄾ	LTH	LT
ㄿ	LPH	LP
ㅀ	LH	LH
ㅄ	PS	BS

Table 2-11 lists the jamo that represent vowels and diphthongs, along with their representations in the three transliteration methods. Again, uppercase is used for clarity, and differences have been highlighted.

Table 2-11: Korean Transliteration—Vowels

Jamo	MOE	KLS	ISO (DPRK and ROK)
ㅏ	A	A	A
ㅑ	YA	YA	YA
ㅓ	Ŏ	EO	EO
ㅕ	YŎ	YEO	YEO
ㅗ	O	O	O
ㅛ	YO	YO	YO
ㅜ	U	U	U
ㅠ	YU	YU	YU
ㅡ	Ŭ	EU	EU
ㅣ	I	I	I
ㅐ	AE	AE	AE
ㅒ	YAE	YAE	YAE
ㅔ	E	E	E
ㅖ	YE	YE	YE
ㅘ	WA	WA	WA
ㅙ	WAE	WAE	WAE

Table 2-11: Korean Transliteration—Vowels (continued)

Jamo	MOE	KLS	ISO (DPRK and ROK)
ㅚ	OE	OE	OE
ㅝ	WǑ	WEO	WEO
ㅞ	WE	WE	WE
ㅟ	WI	WI	WI
ㅢ	ǓI	EUI	YI

Note that the ISO/TR 11941:1996 transliteration method is identical for both North and South Korea (DPRK and ROK, respectively).

As with most transliteration methods, there are countless exceptions and special cases. Tables 2-9 and 2-11 provide only the basic transliterations for jamo. It is when you start combining consonants and vowels that exceptions and special cases become an issue. In fact, a common exception is the transliteration of the hangul used for the Korean surname "Lee." More detailed information about Korean transliteration systems can be found on the Web.*

Vietnamese Romanization Methods

Writing Vietnamese using Latin characters—called *Quốc ngữ* (國語)—is considered the most acceptable method for expressing Vietnamese today. In fact, Quốc ngữ is not considered a transliteration method as with Chinese, Japanese, and Korean—it is the currently acceptable means to express Vietnamese in writing. This writing system is based on Latin script, but is decorated with additional characters and many diacritic marks. This complexity serves to account for the very rich Vietnamese sound system, complete with tones.

In addition to the English alphabet, Quốc ngữ requires two additional consonants and twelve additional base characters (that is, characters that do not indicate tone), as shown in Table 2-12.

Table 2-12: Additional Quốc Ngữ Consonants and Base Characters

Lowercase	đ ăâêôơư
Uppercase	Đ ĂÂÊÔƠƯ

The modifiers that are used for the base vowels, in the order shown in Table 2-12, are called *breve* or *short* (*trăng* or *mũ ngược* in Vietnamese), *circumflex* (*mũ* in Vietnamese), and *horn* (*móc* or *râu* in Vietnamese).

* *http://www.basistech.com/bpkim/koreanroman.html*

While these additional base characters include diacritic marks and other attachments, they do not indicate tone. There are six tones in Vietnamese, five of which are written with a tone mark. Every Vietnamese word must have a tone. These six tones are shown in Table 2-13, along with their names.

Table 2-13: The Six Vietnamese Tones

Tone Mark	Name (in Vietnamese)	Name (in English)
none	Không dấu	*none*
◌̀	Huyền	Grave
◌̉	Hỏi	Hook above, curl, or *hoi*
◌̃	Ngã	Tilde
◌́	Sắc	Acute
◌̣	Nặng	Dot below, underdot, or *nang*

All of the diacritic-annotated characters that are required for the Quốc ngữ writing system, which are combinations of base characters plus tones, are provided in Table 2-14.

Table 2-14: Quốc Ngữ Base Characters with Tone Marks

		Base Characters												
		a A	**ă Ă**	**â Â**	**e E**	**ê Ê**	**i I**	**o O**	**ô Ô**	**ơ Ơ**	**u U**	**ư Ư**	**y Y**
Tone Marks	◌̀	à À	ằ Ằ	ầ Ầ	è È	ề Ề	ì Ì	ò Ò	ồ Ồ	ờ Ờ	ù Ù	ừ Ừ	ỳ Ỳ
	◌̉	ả Ả	ẳ Ẳ	ẩ Ẩ	ẻ Ẻ	ể Ể	ỉ Ỉ	ỏ Ỏ	ổ Ổ	ở Ở	ủ Ủ	ử Ử	ỷ Ỷ
	◌̃	ã Ã	ẵ Ẵ	ẫ Ẫ	ẽ Ẽ	ễ Ễ	ĩ Ĩ	õ Õ	ỗ Ỗ	ỡ Ỡ	ũ Ũ	ữ Ữ	ỹ Ỹ
	◌́	á Á	ắ Ắ	ấ Ấ	é É	ế Ế	í Í	ó Ó	ố Ố	ớ Ớ	ú Ú	ứ Ứ	ý Ý
	◌̣	ạ Ạ	ặ Ặ	ậ Ậ	ẹ Ẹ	ệ Ệ	ị Ị	ọ Ọ	ộ Ộ	ợ Ợ	ụ Ụ	ự Ự	ỵ Ỵ

In summary, Quốc ngữ requires 134 additional characters beyond the English alphabet. 14 are additional base characters (see Table 2-12), and the remaining 120 include diacritic marks that indicate tone (see Table 2-14).

ASCII-based Vietnamese transliteration methods

When only the ASCII character set is available, it is still possible to represent Vietnamese text using well-established systems. The two most common ASCII-based transliteration methods are called VIQR (Vietnamese Quoted-Readable) and VSCII-MNEM (VSCII Mnemonic). The VIQR system is documented in RFC 1456. Table 2-15 illustrates how Quốc ngữ base characters and tones are represented in these two systems.

Table 2-15: VIQR and VSCII-MNEM Transliteration Methods

	Quốc Ngữ	VIQR	VSCII-MNEM
Base Characters	ă Ă	a(A(	a< A<
	â Â	a^ A^	a> A>
	ê Ê	e^ E^	e> E>
	ô Ô	o^ O^	o> O>
	ơ Ơ	o+ O+	o* O*
	ư Ư	u+ U+	u* U*
	đ Đ	dd DD	dd DD
Tones	à À	a` A`	a! A!
	ả Ả	a? A?	a? A?
	ã Ã	a~ A~	a" A"
	á Á	a' A'	a' A'
	ạ Ạ	a. A.	a. A.

Table 2-16 illustrates how base characters and tones are combined in each system. Note how the base character's ASCII-based annotation comes before the ASCII-based tone mark.

Table 2-16: Base Character Plus Tones Using VIQR and VSCII-MNEM Methods

Quốc Ngữ	VIQR	VSCII-MNEM
ờ Ờ	o+` O+`	o*! O*!
ở Ở	o+? O+?	o*? O*?
ỡ Ỡ	o+~ O+~	o*" O*"
ớ Ớ	o+' O+'	o*' O*'
ợ Ợ	o+. O+.	o*. O*.

Zhuyin

Zhuyin, developed in the early 1900s, is a method for transcribing Chinese text using Chinese character elements for their reading value. It is also known as the *National Phonetic System* (注音符号 *zhùyīn fúhào*) or *bopomofo*. The name bopomofo is derived from the readings of the first four characters in the character set: *b*, *p*, *m*, and *f*. There are a total of 37 characters (representing 21 consonants and 16 vowels), along with five symbols to indicate tone (one of which has no glyph) in the zhuyin character set.

Table 2-17 illustrates each of the zhuyin characters, along with their readings—vowels are at the end of the table.

Table 2-17: Zhuyin Characters

Zhuyin	Reading (Pinyin)
ㄅ	B
ㄆ	P
ㄇ	M
ㄈ	F
ㄉ	D
ㄊ	T
ㄋ	N
ㄌ	L
ㄍ	G
ㄎ	K
ㄏ	H
ㄐ	J
ㄑ	Q
ㄒ	X
ㄓ	ZH
ㄔ	CH
ㄕ	SH
ㄖ	R
ㄗ	Z
ㄘ	C
ㄙ	S
ㄚ	A
ㄛ	O
ㄜ	E
ㄝ	EI
ㄞ	AI
ㄟ	EI
ㄠ	AO
ㄡ	OU
ㄢ	AN
ㄣ	EN
ㄤ	ANG
ㄥ	ENG
ㄦ	ER
ㄧ or 一	I

Table 2-17: Zhuyin Characters (continued)

Zhuyin	Reading (Pinyin)
ㄨ	U
ㄩ	IU

The zhuyin character set is included in character sets developed in China (GB 2312-80 and GB/T 12345-90, Row 8) and Taiwan (CNS 11643-1992, Plane 1, Row 5). This set of characters is identical across these two Chinese locales, with one exception, which is indicated in Table 2-17 with two different characters: "丨" is used in China, and "一" is used in Taiwan.

Kana

The most frequently-used writing system in Japanese text is kana. Kana is made up of two closely related writing systems:

- Hiragana
- Katakana

Although one would expect to find kana characters only in Japanese character sets, they are, in fact, part of some Chinese and Korean character sets, in particular GB 2312-80 and KS X 1001:1992. In fact, kana are encoded at the same code points in the case of GB 2312-80! Why in the world would Chinese and Korean character sets include kana? Most likely for the purposes of creating Japanese-looking text using a Chinese or Korean character set.*

The following sections provide detailed information about kana, along with how they were derived from Chinese characters.

Hiragana

Hiragana (平仮名 *hiragana*) are characters that represent sounds, specifically syllables. A syllable is generally composed of a consonant plus a vowel—sometimes a single vowel will do. In Japanese, there are five vowels: *a*, *i*, *u*, *e*, and *o*; and fourteen basic consonants: *k*, *s*, *t*, *n*, *h*, *m*, *y*, *r*, *w*, *g*, *z*, *d*, *b*, and *p*. It is important to understand that hiragana is a syllabary, not an alphabet—you cannot decompose a hiragana character into a part that represents the vowel and a part that represents the consonant. Hiragana (and katakana, covered in the next section) is one of the

* There is, however, one fatal flaw in the Chinese and Korean implementations of kana. They forgot to include five symbols used with kana, all available in row 1 of JIS X 0208:1997: ヽ (01-19), ヾ (01-20), ゝ (01-21), ゞ (01-22), and ー (01-28).

only true syllabaries still in common use today. Table 2-18 illustrates a matrix containing the basic and extended hiragana syllabary.

Table 2-18: The Hiragana Syllabary

		K	S	T	N	H	M	Y	R	W	G	Z	D	B	P
A	あ	か	さ	た	な	は	ま	や	ら	わ	が	ざ	だ	ば	ぱ
I	い	き	し	ち	に	ひ	み		り	ゐ	ぎ	じ	ぢ	び	ぴ
U	う	く	す	つ	ぬ	ふ	む	ゆ	る		ぐ	ず	づ	ぶ	ぷ
E	え	け	せ	て	ね	へ	め		れ	ゑ	げ	ぜ	で	べ	ぺ
O	お	こ	そ	と	の	ほ	も	よ	ろ	を	ご	ぞ	ど	ぼ	ぽ
N	ん														

The following are some notes to accompany Table 2-18:

- Several hiragana have smaller versions, and are as follows (in parentheses you will find the standard version): ぁ (あ), ぃ (い), ぅ (う), ぇ (え), ぉ (お), っ (つ), ゃ (や), ゅ (ゆ), ょ (よ), and ゎ (わ)
- Two hiragana, ゐ and ゑ, are no longer commonly used
- The hiragana を is read as *o*, not *wo*
- The hiragana ん is considered an independent syllable, and is pronounced approximately *ng*

Notice that some cells do not contain any characters. These sounds are no longer used in Japanese, and thus no longer need a character to represent them. Also, the first block of characters is set in a 5×10 matrix. This is sometimes referred to as the *50 Sounds Table* (50音表 *gojūon hyō*), so named because it has a capacity of 50 cells. The other blocks of characters are the same as those in the first block, but with diacritic marks.

Diacritic marks serve to annotate characters with additional information—usually a variant reading. In the West you commonly see accented characters such as *á*, *à*, *â*, *ä*, *ã*, and *å*. The accents are called diacritic marks.

In Japanese there are two diacritic marks: *dakuten* (also called *nigori*) and *handakuten* (also called *maru*). The dakuten (濁点 *dakuten*) appears as two short strokes (゛) in the upper-right corner of some kana characters. The dakuten serves to voice the consonant portion of the kana character to which it is attached.* Examples of voiceless consonants include *k*, *s*, and *t*. Their voiced counterparts are *g*, *z*, and *d*, respectively. Hiragana *ka* (か) becomes *ga* (が) with the addition of the dakuten. The *b* sound is a special voiced version of a voiced *h* in Japanese.

* Voicing is a linguistic term referring to the vibration of the vocal bands while articulating a sound.

The handakuten (半濁点 *handakuten*) appears as a small open circle (゜) in the upper-right corner of kana characters that begin with the *h* consonant. It transforms this *h* sound into a *p* sound.

Hiragana were derived by cursively writing kanji, but no longer carry the meaning of the kanji from which they were derived. Table 2-20 on page 45 lists the kanji from which the basic hiragana characters were derived.

In modern Japanese, hiragana are used to write grammatical words, inflectional endings for verbs and adjectives, and some nouns. They can also be used as a fallback (read "crutch") in case you forget how to write a kanji—the hiragana that represent the reading of a kanji are used in this case. In summary, hiragana are used to write some native Japanese words.

The following characters represent the standard hiragana character set as enumerated in the basic Japanese character set standard, JIS X 0208:1997:

ぁあぃいぅうぇえぉおかがきぎくぐけげこごさざしじすずせぜそぞただちぢっつづてでとどなにぬねのはばぱひびぴふぶぷへべぺほぼぽまみむめもゃやゅゆょよらりるれろゎわゐゑをん

Note how these characters have a cursive or calligraphic look to them (cursive and calligraphic refer to a smoother, handwritten style of characters). Keep these shapes in mind while we move on to katakana.

Katakana

Katakana (片仮名 *katakana*), like hiragana, is a syllabary, and with minor exceptions, they represent the same set of sounds as hiragana. Their usage, however, differs from hiragana. Where hiragana are used to write native Japanese words, katakana are primarily used to write words of foreign origin, called *gairaigo* (外来語 *gairaigo*), to write onomatopoeic words,* and for emphasis—similar to the use of italics to represent foreign words and to express emphasis in English. For example, the Japanese word for bread is written パン and read *pan*. It was borrowed from the Portuguese word *pão*, which is read sort of like *pown*. Katakana are also used to write foreign names. Table 2-19 illustrates the basic and extended katakana syllabary.

The following are some notes to accompany Table 2-19:

- Several katakana have smaller versions, and are as follows (in parentheses you will find the standard version): ァ (ア), ィ (イ), ゥ (ウ), ェ (エ), ォ (オ), ヵ (カ), ヶ (ケ), ッ (ツ), ャ (ヤ), ュ (ユ), ョ (ヨ), and ヮ (ワ)
- Two katakana, ヰ and ヱ, are no longer commonly used

* Words that serve to describe a sound, such as *buzz* or *biss* in English. In Japanese, for example, ブクブク (*bukubuku*) represents the sound of a balloon expanding.

- The katakana ヲ is read as *o*, not *wo*
- The katakana ン is considered an independent syllable, and is pronounced approximately *ng*

Table 2-19: The Katakana Syllabary

		K	S	T	N	H	M	Y	R	W	G	Z	D	B	P
A	ア	カ	サ	タ	ナ	ハ	マ	ヤ	ラ	ワ	ガ	ザ	ダ	バ	パ
I	イ	キ	シ	チ	ニ	ヒ	ミ		リ	ヰ	ギ	ジ	ヂ	ビ	ピ
U	ウ	ク	ス	ツ	ヌ	フ	ム	ユ	ル		グ	ズ	ヅ	ブ	プ
E	エ	ケ	セ	テ	ネ	ヘ	メ		レ	ヱ	ゲ	ゼ	デ	ベ	ペ
O	オ	コ	ソ	ト	ノ	ホ	モ	ヨ	ロ	ヲ	ゴ	ゾ	ド	ボ	ポ
N	ン														

Katakana were derived by extracting a single portion of a whole kanji, and, like hiragana, no longer carry the meaning of the kanji from which they were derived. If you compare several of these characters to some kanji, you may recognize common shapes. Table 2-20 lists the basic katakana characters, along with the kanji from which they were derived.

The following characters represent the standard katakana character set as enumerated in the basic Japanese character set standard, JIS X 0208:1997:

ァアィイゥウェエォオカガキギクグケゲコゴサザシジスズセゼソゾタダチヂッツヅテデトドナニヌネノハバパヒビピフブプヘベペホボポマミムメモャヤュユョヨラリルレロヮワヰヱヲンヴヵヶ

Katakana, unlike hiragana, have a squared, more rigid feel to them. Structurally speaking, they are quite similar in appearance to kanji, which we discuss later.

The Development of Kana

You already know that kana were derived from kanji, and Table 2-20 provides a complete listing of kana characters, along with the kanji from which they were derived.

Table 2-20: The Kanji from Which Kana Were Derived

Katakana	Kanji		Hiragana
ア	阿	安	あ
イ	伊	以	い
ウ	宇		う
エ	江	衣	え
オ	於		お

Table 2-20: The Kanji from Which Kana Were Derived (continued)

Katakana	Kanji		Hiragana
カ	加		か
キ	幾		き
ク	久		く
ケ	介	計	け
コ	己		こ
サ	散	左	さ
シ	之		し
ス	須	寸	す
セ	世		せ
ソ	曽		そ
タ	多	太	た
チ	千	知	ち
ツ	川		つ
テ	天		て
ト	止		と
ナ	奈		な
ニ	二	仁	に
ヌ	奴		ぬ
ネ	祢		ね
ノ	乃		の
ハ	八	波	は
ヒ	比		ひ
フ	不		ふ
ヘ	部		へ
ホ	保		ほ
マ	万	末	ま
ミ	三	美	み
ム	牟	武	む
メ	女		め
モ	毛		も
ヤ	也		や
ユ	由		ゆ
ヨ	與	与	よ
ラ	良		ら
リ	利		り
ル	流	留	る

Table 2-20: The Kanji from Which Kana Were Derived (continued)

Katakana	Kanji		Hiragana
レ	礼	禮	れ
ロ	呂		ろ
ワ	和		わ
ヰ	井	為	ゐ
ヱ	恵		ゑ
ヲ	乎	遠	を
ン	尓	无	ん

Note how many of the kanji from which katakana and hiragana characters were derived are the same, and how the shapes of several hiragana/katakana pairs are similar. In fact, many katakana are nearly identical to kanji, and can usually be distinguished by their smaller size. Katakana can usually be distinguished from kanji in that they are usually found in strings containing other katakana. Table 2-21 shows some examples of this phenomenon.

Table 2-21: Katakana and Kanji with Similar Shapes

Katakana	Kanji
エ	工
カ	力
タ	夕
ト	卜
ニ	二
ネ	ネ
ハ	八
ヒ	匕
ム	厶
メ	乂
ロ	口

Hangul

Hangul (한글 *hangeul*) are the characters that are used to express contemporary Korean texts in writing.* Unlike Japanese kana, hangul is not a syllabic writing system, but rather a writing system that is composed of elements that represent a pure alphabet. How does one make the distinction between an alphabet and syllabary? Each hangul character can be *easily* decomposed into hangul elements,

* The word "hangul" was coined sometime around 1910, and means "Korean script."

which in turn represent individual sounds (that is, consonants and vowels), not syllables. Hangul elements, which do not carry any meaning, are commonly referred to as *jamo* (자모/字母 *jamo*), literally meaning "alphabet."*

King Sejong (世宗 *sejong*) of the Yi Dynasty initiated the development of what is now referred to as hangul back in the year 1440, and the work was completed in 1446.† Hangul is considered to be one of the most scientific (or well-designed) writing systems due to its extremely regular and predictable structure.‡

Jamo are typically combined with one or two additional jamo to form hangul, which represent syllabic units. Table 2-22 lists a handful of hangul, along with the jamo used to build them.

Table 2-22: Decomposition of Hangul into Jamo

Hangul	Reading	Jamo	Transliterated
가	GA	ㄱ plus ㅏ	*g* plus *a*
갈	GAL	ㄱ plus ㅏ plus ㄹ	*g* plus *a* plus *l*
갉	GALG	ㄱ plus ㅏ plus ㄹ plus ㄱ	*g* plus *a* plus *l* plus *g*

There are exactly six ways to combine simple jamo into modern hangul, as illustrated in Figure 2-1, along with examples ("C" stands for consonant, "V" stands for vowel, and the order in which consonants and vowels are read is indicated with numerals).

Korean has special terms for the jamo that are used to construct hangul, depending on where in the syllable they appear:

- *Coseong* (초성/初聲 *coseong*) for the initial sound, usually a consonant
- *Jungseong* (중성/中聲 *jungseong*) for the middle sound, usually a vowel
- *Jongseong* (종성/終聲 *jongseong*) for the final sound, usually a consonant

The following characters represent the 93 jamo as enumerated in row 4 of the basic Korean character set standard, KS X 1001:1992:

ㄱ ㄲ ㄳ ㄴ ㄵ ㄶ ㄷ ㄸ ㄹ ㄺ ㄻ ㄼ ㄽ ㄾ ㄿ ㅀ ㅁ ㅂ ㅃ ㅄ ㅅ ㅆ ㅇ ㅈ ㅉ ㅊ ㅋ ㅌ ㅍ ㅎ ㅏ ㅐ ㅑ ㅒ ㅓ ㅔ ㅕ ㅖ ㅗ ㅘ ㅙ ㅚ ㅛ ㅜ ㅝ ㅞ ㅟ ㅠ ㅡ ㅢ ㅣ ㅥ ㅦ ㅧ ㅨ ㅩ ㅪ ㅫ ㅬ ㅭ ㅮ ㅯ ㅰ ㅱ ㅲ ㅳ ㅴ ㅵ ㅶ ㅷ ㅸ ㅹ ㅺ ㅻ ㅼ ㅽ ㅾ ㅿ ㆀ ㆁ ㆂ ㆃ ㆄ ㆅ ㆆ ㆇ ㆈ ㆉ ㆊ ㆋ ㆌ ㆍ ㆎ

* Sometimes referred to as *jaso* (자소/字素 *jaso*).

† The result of this work was a book entitled 訓民正音 (훈민정음 *hunmin jeongeum*).

‡ Unfortunately, its Latin transliteration methods that we explored earlier in this chapter aren't nearly as regular or predictable.

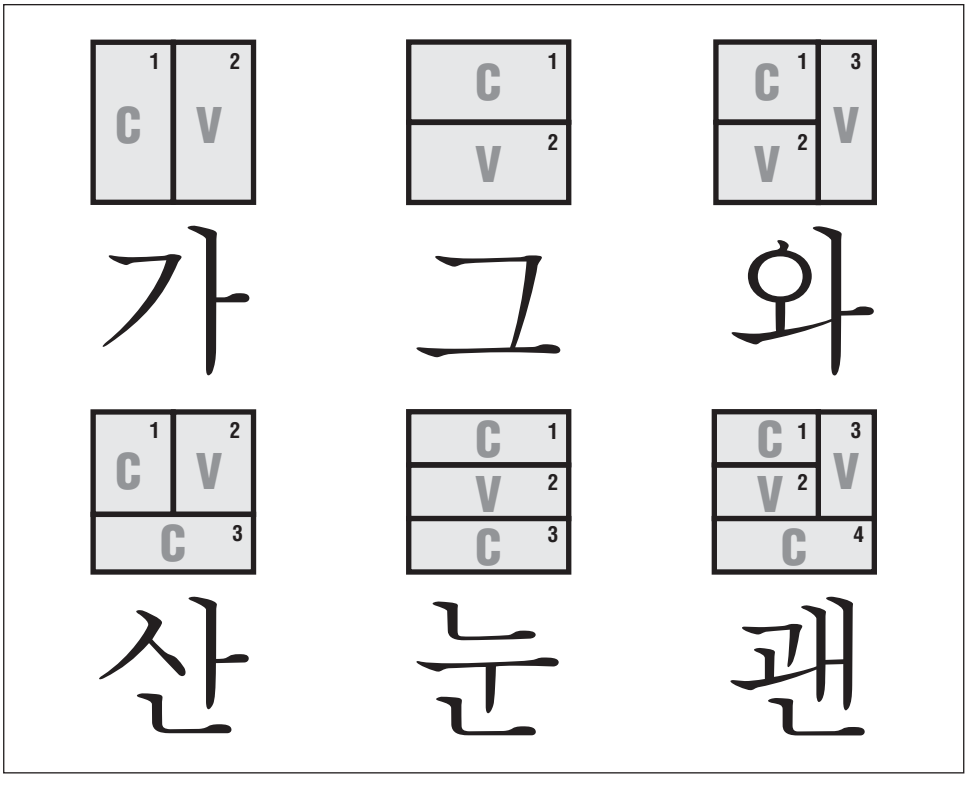


Figure 2-1: Six ways to compose jamo into modern hangul

The first 51 of these jamo are considered modern, and are used in combination to form modern hangul. The remaining 42 jamo are considered ancient.

The following characters represent the first 94 hangul as enumerated in row 16 of the basic Korean character set standard, KS X 1001:1992:

가각간갇갈갉갊감갑값갓갔강갖갗같갚갛개객갠갤갬갭갯갰갱갸갹갼걀걋걍걔걘걜
거걱건걷걸걺검겁것겄경겆겉겊겋게겐겔겜겝겟겠겡겨격겪견겯결겸겹겻겼경곁계
곈곌곕곗고곡곤곧골곪곬곯곰곱곳공곶과곽관괄괆

Appendix N, *Hangul Reading Table*, provides a complete code table, annotated with readings, for all 2,350 hangul in the KS X 1001:1992 character set standard. This volume is potentially useful for those who do not yet have a good command of hangul, or who do not have access to a Korean-capable operating system. There are also details about hangul on the Web.*

* *http://www.hangul.org/*

Chinese Characters

The single most complex type of character used in CJKV text are Chinese characters (汉字/漢字 *hànzì* in Chinese; 漢字 *kanji* in Japanese; 한자/漢字 *hanja* in Korean; and *chữ Hán*/字漢 in Vietnamese). To grasp the concept of Chinese characters, one must first understand the magnitude of such a writing system. The 26 characters of the English alphabet (52 characters, if one counts both upper- and lower case) seem quite limiting compared to the tens of thousands of Chinese characters in current use by the CJKV locales. It is well documented that the Japanese borrowed the Chinese script over the course of a millennium. What is not well known is that while the Japanese were borrowing from the Chinese, the Chinese were, themselves, adding to the total number of characters in their language by creating new characters.* This means that the Japanese were able, in essence, to capture and freeze a segment of Chinese history every time they borrowed from the Chinese. The same can be said of Korean and Vietnamese, both of whom also borrowed Chinese characters.

Before we begin discussing the history of Chinese characters, and how Chinese characters are composed, let's take some time to illustrate some Chinese characters. The following sets of characters represent the first row of 94 Chinese characters in each of the CJKV character set standards (with row number indicated in parentheses):

GB 2312-80—Row 16

啊阿埃挨哎唉哀皑癌蔼矮艾碍爱隘鞍氨安俺按暗岸胺案肮昂盎凹敖熬翱袄傲奥懊澳芭捌扒叭吧笆八疤巴拔跋靶把耙坝霸罢爸白柏百摆佰败拜稗斑班搬扳般颁板版扮拌伴瓣半办绊邦帮梆榜膀绑棒磅蚌镑傍谤苞胞包褒剥

GB/T 12345-90—Row 16

啊阿埃挨哎唉哀皚癌藹矮艾礙愛隘鞍氨安俺按暗岸胺案骯昂盎凹敖熬翱襖傲奧懊澳芭捌扒叭吧笆八疤巴拔跋靶把耙壩霸罷爸白柏百擺佰敗拜稗斑班搬扳般頒板版扮拌伴瓣半辦絆邦幫梆榜膀綁棒磅蚌鎊傍謗苞胞包褒剝

CNS 11643-1992 Plane 1—Row 36

一乙丁七乃九了二人儿入八几刀刁力匕十卜又三下丈上丫丸凡久么也乞于亡兀刃勺千叉口土士夕大女子孑孓寸小尢尸山川工己已巳巾干廾弋弓才丑丐不中丰丹之尹予云井互五亢仁什仃仆仇仍今介仄元允內六兮公冗凶

* The Chinese are still creating new characters in some locales, especially Hong Kong.

CNS 11643-1992 Plane 2—Row 1

乂乜凵匚厂万丌乇亍囗屮彳丏冇与丮亓仂仉仈冘勼卬厹圠夃夬尐巿旡殳毌气爿丱丼
仨仜仩仡仝仚刌刉匜卌圢圣夗夯宁宄尒尻屴屳帄庀庂忉戉扐氕氶汃氿氻犮犰玊禸肊
阞伎优伬仵伔仱伀价伈伝伂伅伢伓伄仴伒冱刓刐劦

CNS 11643-1992 Plane 3—Row 1

丨丶丿亅丅丄冂冖匸卩厶个亇义凢乆亐亼亾亾凣刄尣廾夂夊宀巛幺广廴彐彑彡阝丈
乣乢亣内仅仏从仌冄円冗汀凨刅办劝勽匀区协夲卝历厷会双叏収圡屮帀弍户戸攴攵
无丰生㞢仟伏仉夫伉仦仍伇参仝同同册写尻切刋刉

CNS 11643-1992 Plane 4—Row 1

𠂆乁乀乚く丂丩乃几刂㔾巜冃日勹卂卩孑孒尢屮忄扌犭刃彡三仐先冃仌刐刕早叩产
厸叱百壬矢允允币开不汔灬亓冈艹辶世丫乍广仁仢任先再㐱凧凷刌刏劶刢医厈尾㚇
丈叽谷另囚因㔯先压㚈尔㞋户产帆半化扣抖甲朩禾

CNS 11643-1992 Plane 5—Row 1

ㄣㄥ乙乚乙ナ丩乚凵马廿屯丮厷千屯心少巨玌今从仈欠㔽刂切功匃卯厄叕叉㫐吃玊
屹岂㞢朩歹毛爪王豕朱乩尹早囘夙由刎庞㓞㔟匏本夆厍产勊肉叱另妣孔宍犰㠯圳卉
纣弘礼忙戸乳扒此卣毋氵卜汉年夹犰犰而聿丫邗阩阸

CNS 11643-1992 Plane 6—Row 1

乁卜㐅丩㕚乚二三二七巳卜工丐且牛丸凡丂天个亼六刃勹力可云飞尸光弋兮且丘历
玊凧禾爪去区仃㐌兀亽冋内舟兊出凶瓜刂刅劢亡廾卬冎冎丸丞殳及収色囜囙囜史太
凬寸不屮丝王五玊㔾乐刄廾廾弍弓弓弓弓弓弓孔忄

CNS 11643-1992 Plane 7—Row 1

壼寡奲奭僊僜僆僘僴僂傀僜僚僻傀僝僸兠靐毱圖凛瀞劃劆劏劗劕劊劊劉韓寫㕓厰愿
嵫臺离㽗嗇噌噻嗲嚓嘝嗻嚹嚩嚍嘰噑噦嚐嚄嚧嚅嚳毀貫噐嚗嚕嘘嗥噁瓥疊園圜圝壘墖
墕墸墒墇墵墤冀齸墿壿复壒臺墜壁壂壽臺堝嵡橐嬎

CNS 11643-1986 Plane 15—Row 1

乙壬乣从生么厄尧仈厄凧允凡尢刈匃区匹厄邓曰叫囟个不龙弓开幺屯史乔企乍佇佀
夙有冄冗安次充凷厄叨叽叺坧㚘乔夸尥奶存刋龙尬㞋町屮岙㔾产羊幻处尨杢汉水炎
肒迂夵肙底乩亦宋仪伉关佰伂伂征攸低侂汞罕冷汼

JIS X 0208:1997—Row 16

亜唖娃阿哀愛挨姶逢葵茜穐悪握渥旭葦芦鯵梓圧斡扱宛姐虻飴絢綾鮎或粟袷安庵按
暗案闇鞍杏以伊位依偉囲夷委威尉惟意慰易椅為畏異移維緯胃萎衣謂違遺医井亥域
育郁磯一壱溢逸稲茨芋鰯允印咽員因姻引飲淫胤蔭

JIS X 0212-1990—Row 16

丐丄丌丌刃丟乖两丨丫丯丰丰丵乀乁乂毛氽乚乜乩乩乩乞乳乹乩亍三岂叁亶亹仃仐
企托仟仡们仜仌伶仳件份佤仿伀伂伢伀伋伬伙伙佞你你丗佪伾侶征佈
佶侶侐侠佔侀佘佟佣佪佬佮金佷佸侮佺佚佾侁侘侄

KS X 1001:1992—Row 42

伽佳假價加可呵哥嘉嫁家暇架枷柯歌珂痂稼苛茄街袈訶賈跏軻迦駕刻却各恪慤殼珏
脚覺角閣侃刊墾奸姦干幹懇揀杆柬桿澗癎看磵稈竿簡肝艮艱諫間乫喝曷渴碣竭葛褐
蝎鞨勘坎堪嵌感憾戡敢柑橄減甘疳監瞰紺邯鑑鑒龕

KS X 1002:1991—Row 55

仮傢咖帑坷宪孵榎樌珈笳勒舸葭諤卻咯垹搉擱栲侃慳棨榦玕秆萇術赶迁齦噶楬秸羯
蠍鬾坩堝嵁弇憨撼欿歛泔淦澉矙轗酣讌韐傋僵壇忱扛杠橿殭矼穅繈罟羌羟茫豇韁剴
匈揩檠玲磕閫硜賡鏗呿昒秬筥籧胅腒苣茞蕖躉袟裾

TCVN 6056:1995—Row 42

一丁七万丈三上下不与丐丑且丕世丘丙丞丢丫中丰串丸丹主丿乂乃久之乍乎乏乖乘
乙九乞也乱乳乾亂了予事二亍于云互五井亙些亟亡亢交亥亨享京亭人什仁仃仄仆仇
今介仍仔仕他仗付仙仝代令仰件价任份仿企伊伍伎

As you can clearly see, there is quite a variety of character sets available that include Chinese characters. Chapter 3, *Character Set Standards*, will make sense of all these character set standards.

A noteworthy characteristic of Chinese characters is that they can be quite complex (believe it or not, they can get much more complex than the sets of 94 characters illustrated above!). Chinese characters are composed of radicals and radical-like elements, which can be thought of as building-blocks of sorts. Radicals are discussed later in this chapter starting on page 54, in the section entitled "The Structure of Chinese Characters."

Chinese Character Readings

In Japanese, the typical Chinese character has at least two readings—some have more. For example, the Chinese character 生, whose meaning relates to "life," has over 200 readings in Japanese—most of which are used for Japanese given names, which are known for their unusual readings.

Chinese character readings come from two sources:

- Language-specific reading
- Borrowed (and usually approximated) reading

The native Japanese reading was how the Japanese read a word before the Chinese influenced their language and writing system. The native Japanese reading is called the Kun reading (訓読み *kun yomi*).

The borrowed Chinese reading is the Japanese-language approximation of the original Chinese reading of a kanji. These borrowed approximate readings are called On readings (音読み *on yomi*), On being the word for "sound." If a particular kanji was borrowed more than once, multiple readings can result. Table 2-23 lists several kanji, along with their respective readings.

Table 2-23: Chinese Characters and Their Readings—Japanese

Kanji	Meaning	On Readings	Kun Readings
剣	*sword*	ken	akira, haya, tsurugi, tsutomu
窓	*window*	sō	mado
車	*car*	sha	kuruma
万	*ten-thousand*	ban, man	katsu, kazu, susumu, taka, tsumoru, tsumu, yorozu
生	*life, birth*	sei, shō	ari, bu, fu, fuyu, haeru, hayasu, i, ikasu, ikeru, ikiru, iku, ki, mi, nama, nari, nori, o, oki, ou, susumu, taka, ubu, umareru, umu, yo, *and so on*
店	*store, shop*	ten	mise

So, how does one go about deciding which reading to use? Good question! As you learned earlier, the Japanese borrowed kanji as compounds of two or more kanji, and often use the On reading for such compounds. Conversely, when these same kanji appear in isolation, the Kun reading is often used. Table 2-24 provides some examples of individual kanji and kanji compounds.

Table 2-24: Kanji and Kanji Compounds—Japanese

Kanji Compound	Meaning	Readings
自動車	*automobile*	jidōsha (On readings)
車	*car*	kuruma (Kun reading)
剣道	*Kendo*	kendō (On readings)
剣	*sword*	tsurugi (Kun reading)

As with all languages, there are always exceptions to rules! Sometimes you find kanji compounds that use Kun readings for one or all kanji. You may also find kanji in isolation that use On readings. Table 2-25 lists some examples.

Table 2-25: Irregular Uses of Kanji Readings—Japanese

Kanji Compound	Meaning	Reading
重箱	*nest of boxes*	jūbako (On plus Kun reading)
湯桶	*bath ladle*	yutō (Kun plus On reading)

Table 2-25: Irregular Uses of Kanji Readings—Japanese (continued)

Kanji Compound	Meaning	Reading
窓口	*ticket window*	madoguchi (Kun plus Kun reading)
単	*simple, single*	tan (On reading)

Japanese personal names tend to use the Kun readings even though they are in compounds. For instance, 藤本 is read *fujimoto* rather than *tōhon*.

The Structure of Chinese Characters

Chinese characters are composed of smaller, primitive units called radicals, and other non-radical elements, which are used as building blocks. These elements serve as the most basic units for building Chinese characters. 214 radicals are used for indexing Chinese characters. Several radicals stand alone as single, meaningful Chinese characters. Table 2-26 provides some examples of radicals, along with several Chinese characters that can be written with them (examples are taken from Japanese).

Table 2-26: Radicals and Chinese Characters Made from Them

Radical	Variant	Stand-Alone?	Meaning	Examples
木		yes	*tree*	本札朴朶李材条杲林桙栞棚森槝
火	灬	yes	*fire*	灯灰灸災炎点無然熊熟熱燃燭爛
水	氵	yes	*water*	氷永汁江汲沢泉温測港源溢澡濯
辶	辶	no	*running*	辷込辻辺辿迄迅迎近返迚連週還

Note how each radical is placed within Chinese characters—they are stretched or squeezed so that all of the radicals that constitute a Chinese character fit into the general shape of a square. Also note how radicals are positioned within Chinese characters, specifically on the left, right, top, or bottom.

Radicals and radical-like elements, in turn, are composed of smaller units called *strokes*. A radical can consist of one or more strokes. Sometimes a single stroke is considered a radical. There exists one stroke that is considered a single Chinese character: 一, the Chinese character that represents the number one. Figure 2-2 shows how a typical Chinese character is composed of radicals and strokes.

There are many classifications of Chinese characters, but four are the most common: pictographs, simple ideographs, compound ideographs, and phonetic ideographs. Pictographs, the most basic of the Chinese characters, are little pictures, and usually look much like the object they represent.* Table 2-27 lists examples of pictographs.

* Written 象形文字 (*xiàngxíng wénzì*) in Chinese, 象形文字 (*shōkei moji*) in Japanese, and 상형문자/象形文字 (*sanghyeong munja*) in Korean.

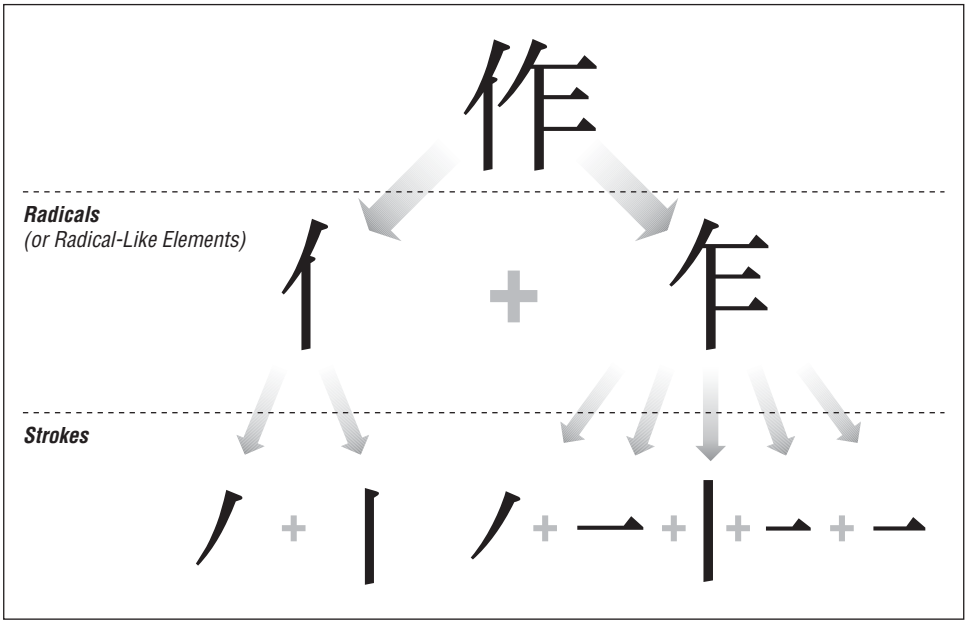


Figure 2-2: Decomposition of Chinese characters into radicals and strokes

Table 2-27: Pictographs

Chinese Character	Meaning
日	*sun*
月	*moon*
山	*mountain*
火	*fire*
木	*tree*
車	*car, cart*
口	*mouth, opening*

Whereas pictographs represent concrete objects, simple ideographs represent abstract concepts or ideas (as the name suggests), such as numbers and directions.* Table 2-28 lists examples of simple ideographs.

Table 2-28: Simple Ideographs

Chinese Character	Meaning
上	*up*
下	*down*

* Written 指事文字 (*zhǐshì wénzì*) in Chinese, 指事文字 (*shiji moji*) in Japanese, and 지사문자/指事文字 (*jisa munja*) in Korean.

Table 2-28: Simple Ideographs (continued)

Chinese Character	Meaning
中	*center, middle*
一	*one*
二	*two*
三	*three*

Pictographs and simple ideographs can be combined to represent more complex characters, and usually reflect the combined meaning of its individual elements. These are called compound ideographs.* Table 2-29 lists examples of compound ideographs.

Table 2-29: Compound Ideographs

Chinese Character	Components	Meaning
林	木 + 木	*woods*
森	木 + 木 + 木	*forest*
明	日 + 月	*clear, bright*

Phonetic ideographs account for more than 90 percent of all Chinese characters.† They usually have at least two components: one to indicate reading, and the other to denote etymological meaning. Table 2-30 provides examples that all use the same base reading component.

Table 2-30: Phonetic Ideographs with Common Reading Component—Japanese

Chinese Character	Meaning	Reading	Meaning Part	Reading Part
銅	*copper*	dō	金 (*metal*)	同 (dō)
洞	*cave*	dō	氵 (*water*)	同 (dō)
胴	*torso*	dō	肉 (*organ*)	同 (dō)
恫	*threat*	dō	忄 (*heart*)	同 (dō)

Note that each uses the 同 radical (*dō*) for its reading component. Table 2-31 lists several Chinese characters that use the same base meaning component.

Table 2-31: Phonetic Ideographs with Common Meaning Component—Japanese

Chinese Character	Meaning	Reading	Meaning Part	Reading Part
雰	*fog*	fun	雨 (*rain*)	分 (fun)
雲	*cloud*	un	雨 (*rain*)	云 (un)

* Written 会意文字/會意文字 (*huìyì wénzì*) in Chinese, 会意文字 (*kaii moji*) in Japanese, and 회의문자/會意文字 (*hoeyi munja*) in Korean.

† Written 形声文字/形聲文字 (*xíngshēng wénzì*) in Chinese, 形声文字 (*keisei moji*) in Japanese, and 형성문자/形聲文字 (*hyeongseong munja*) in Korean.

Table 2-31: Phonetic Ideographs with Common Meaning Component—Japanese (continued)

Chinese Character	Meaning	Reading	Meaning Part	Reading Part
震	*shake*	shin	雨 (*rain*)	辰 (shin)
霜	*frost*	sō	雨 (*rain*)	相 (sō)

Note that each uses the 雨 ("rain") radical for its meaning component. The 雨 radical is another example of a radical that can stand alone as a single Chinese character.

Chinese characters are subsequently combined with other Chinese characters as words to form more complex ideas or concepts. These are called *compounds* (熟語 *jukugo* in Japanese) or *Chinese character compounds* (漢語 *kango* in Japanese). Table 2-32 lists a few examples. Note that you can decompose words into pieces, each piece being a single Chinese character with its own meaning.

Table 2-32: Chinese Character Compounds

Compound	Meaning	Component Chinese Characters and Their Meanings
日本	*Japan*	日 means *sun*, and 本 means *origin*
短刀	*short sword*	短 means *short*, and 刀 means *sword*
酸素	*oxygen*	酸 means *acid*, and 素 means *element* (the acid element)
曲線	*curve*	曲 means *curved*, and 線 means *line* (curved line)
剣道	*Kendo*	剣 means *sword*, and 道 means *path* (the way of the sword)
自動車	*automobile*	自 means *self*, 動 means *moving*, and 車 means *car*
火山	*volcano*	火 means *fire*, and 山 means *mountain* (fire mountain)

There, that should have given you a sense of how Chinese characters are constructed and how they are combined with other Chinese characters to form compounds. But how did they come to be used in Korea and Japan? These and other questions are answered next.

The History of Chinese Characters

This section provides some brief historical context to explain the development of Chinese characters, and how they came to be used in other cultures, such as Korea, Japan, and Vietnam.

The development of Chinese characters

Chinese characters, believe it or not, share a history similar to that of the Latin alphabet. Both writing systems began thousands of years ago as pictures that encompassed meanings. While the Latin alphabet eventually gave up any semantic association with the characters' shapes, Chinese characters retained (and further

exploited) this feature. Table 2-33 lists several Chinese reference works whose year of publishing span a period of approximately 2,000 years.

Table 2-33: The Number of Chinese Characters During Different Periods

Year (AD)	Number of Chinese Characters	Reference Work
100	9,353	說文解字
227–239	11,520	聲類
480	18,150	廣雅
543	22,726	玉編
751	26,194	唐韻
1066	31,319	類編
1615	33,179	字彙
1716	47,021	康熙字典
1919	44,908	中華大字典
1969	49,888	中文大辭典
1986	56,000	汉语大字典
1994	85,000	中华字海

Note the nearly five-fold increase in the number of hanzi over this 2,000 year period. The majority of the Chinese characters that sprang into existence during this time were phonetic ideographs (see Tables 2-30 and 2-31 on page 56).

Chinese characters in Korea—hanja

One of the earliest cultures to adapt Chinese characters for their own language was Korea. Although Chinese characters—called *hanja*—were extensively used in years past, most Korean writing today is completely in hangul.

While there appears to be an attempt by the Korean Ministry of Education to restore the use of hanja into its society—by requiring that students learn a basic set of 1,800 hanja—it does not appear to be having much of an effect. This set of 1,800 hanja was introduced in 1972, so there may still be time necessary for it to effectively restore the use of hanja in Korea.

The definitive Korean hanja reference is a dictionary entitled 大字源 (대자원 *daejaweon*), first published in 1972.

Chinese characters in Japan—kanji

There is no evidence to suggest that there was *any* writing system in place in Japan prior to the introduction of Chinese script. In fact, it is quite common for writing systems to develop relatively late in the history of languages. A writing

system as complex as that used in Chinese is not really an ideal choice for borrowing, but perhaps this was the only writing system from which the Japanese could choose at the time.

The Japanese borrowed Chinese characters between 222 AD and 1279 AD. During this millennium of borrowing, the Chinese increased their inventory of characters nearly three-fold. Table 2-33 illustrated the number of Chinese characters that were documented in Chinese at different periods. That table clearly indicated that the Chinese, over a period of about 2,000 years, increased their inventory of characters by roughly a factor of five (from 9,353 to 49,888). As you can see, the Japanese were borrowing from the Chinese even while the Chinese were still creating new characters.

The Japanese began borrowing the Chinese script over 1,600 years ago. This massive borrowing took place in three different waves. Several kanji were borrowed repeatedly at different periods, and the reading of each kanji was also borrowed again. This led to different readings for a given kanji depending on which word or words it appeared in, due to dialectal and diachronic differences in China.

The first wave of borrowing took place sometime between 222 and 589 AD by way of Korea, during the Six Dynasties Period in China. Characters borrowed during this period were those used primarily in Buddhist terminology. During this period, the Chinese had between 11,520 and 22,726 hanzi.

The second wave took place between 618 and 907 AD, during the Tang Dynasty in China. Characters borrowed during this period were those used primarily for government and in Confucianism terminology. During this period, the Chinese had between 22,726 and 26,194 hanzi.

The third wave occurred somewhere between 960 and 1279 AD, during the Song Dynasty in China. Characters borrowed during this period were those used in Zen terminology. The Chinese had between 31,319 and 33,179 hanzi by this period.

During all three waves of borrowing, most Chinese characters were borrowed as compounds of two or more kanji, rather than as isolated characters. It is in this context that you find differences in reading of a particular kanji depending on what word it appears in. For example, the kanji 万, meaning "ten thousand," can be found in kanji compounds with either the reading *man* or *ban*, such as 万一 (*man* + *ichi*) and 万歳 (*ban* + *zai*—yes, the actual kanji compound for *banzai*!). This (*m*)*an*/(*b*)*an* alternation would indicate to a trained linguist that these two words were probably borrowed at different periods.

The first two waves of borrowing had the most significant impact on the Japanese lexicon, which accounts for dual On readings for many kanji (lexicon simply refers

to the individual words that constitute a language). The third wave of borrowing had very little effect on the Japanese lexicon.

I suggest the front matter of Jack Halpern's *New Japanese-English Character Dictionary* as additional reference material on the history and development of the Japanese writing system. More specifically, pp 50a through 60a of that reference. The definitive Japanese kanji reference is a 13-volume dictionary entitled 大漢和辭典 (*dai kanwa jiten*), first published in 1955.

Chinese characters in Vietnam—chữ Hán

Vietnam also adopted Chinese characters for their language, but in a unique way. There are two ways to represent Vietnamese using Chinese characters. One way is equivalent to Chinese itself (but with approximated readings when pronounced in Vietnamese), and uses characters called *chữ Hán* (genuine Chinese characters). The other way involves characters that look and feel like Chinese characters, but were created by the Vietnamese. These are called *chữ Nôm* (字喃). These methods of writing Vietnamese are unique in that they are never used together in the same text—you write using either chữ Hán (Chinese) or chữ Nôm (Vietnamese). More details about chữ Nôm are provided at the end of this chapter.

Both chữ Hán and chữ Nôm were replaced by *Quốc ngữ* in 1920. Today, chữ Hán and chữ Nôm are still being used—not for the purpose of common communication, but rather for specialized, religious, or historical purposes.

Chinese Character Simplification

Over time, frequently used and complex Chinese characters tend to simplify. Such simplifications have been different depending on the locale using them. For example, Chinese characters in their traditional form are still being used in Taiwan. The same holds true for Korea. Also, Chinese characters in an even more simplified form than found in Japanese are being used in China and Singapore, although there are some exceptions to this rule. A large number of Chinese characters are used in an almost identical form in all CJKV locales. Table 2-34 illustrates several Chinese characters in both traditional and simplified form.

Table 2-34: Traditional and Simplified Chinese Characters

Traditional	Simplified (Japan)	Simplified (China)
廣	広	广
兒	児	儿
兩	両	两
氣	気	气

Table 2-34: Traditional and Simplified Chinese Characters (continued)

Traditional	Simplified (Japan)	Simplified (China)
豐	豊	丰
邊	辺	边
國	国	国
學	学	学
點	点	点
黑	黒	黑
佛	仏	佛
骨	骨	骨

Both the simplified and traditional forms of Chinese characters sometimes coexist within the same character set standard, and some of the pairs from Table 2-34 are such examples—most of them are part of the basic Japanese character set standard, specifically JIS X 0208:1997. You can also see that some simplifications are more extreme than others.

Such simplifications in Japan have led to variants of many characters, and in some character sets both the simplified and traditional forms are included (the examples given above are such cases). As an extreme example, let's examine the JIS X 0208:1997 kanji 剣 (Row-Cell 23-85), whose five variant kanji are also encoded within the same character set standard. These variants are listed in Table 2-35 (Row-Cell values are given).

Table 2-35: Chinese Character Variants in the Same Character Set

Chinese Character	Character Code
劍	49-88
劒	49-89
劔	49-90
劍	49-91
釼	78-63

Non-Chinese Chinese Characters

What is a non-Chinese Chinese character? Simple. Characters that look, feel, and behave like Chinese characters, but were not borrowed from China. The following sections describe this interesting and remarkable phenomenon as it has manifested in Japan, Korea, and Vietnam. Examples are also provided.

Japanese-Made Chinese Characters—Kokuji

The Japanese have created their own Chinese characters known as *kokuji* (国字 *kokuji*), literally meaning "national characters," or, more descriptively, "Japanese-made Chinese characters." Kokuji behave like true Chinese characters, following the same rules of structure, specifically that they are composed of radicals, radical-like elements, and strokes, and can be combined with one or more Chinese characters to form compounds or words. These Chinese characters were created out of a need for characters not borrowed from China.* Most kokuji are used to represent the names of indigenous Japanese plants and fish. They are also used quite frequently in Japanese place and personal names.

Approximately 200 kokuji have been identified in the basic Japanese character set standard, specifically JIS X 0208:1997. There are even more in the supplemental set, specifically JIS X 0212-1990. Table 2-36 provides a few examples of kokuji (JIS X 0208:1997 Row-Cell values are provided).

Table 2-36: Examples of Kokuji

Kokuji		Readings	Meanings
鰯	16-83	iwashi	*sardine*
粂	23-09	kume	Used in personal names
込	25-94	komu	(*to*) *move inward*
榊	26-71	sakaki	A species of tree called *sakaki*
働	38-15	hataraku, dō[a]	(*to*) *work*
峠	38-29	tōge	*mountain pass*
畑	40-10	hata, hatake	*dry field*
枠	47-40	waku	*frame*
凩	49-62	kogarashi	*cold, wintry wind*

[a] Considered an On reading.

Additional kokuji were created when the Japanese isolated themselves from the rest of the world for approximately 250 years: from the mid-1600s to the late 1800s. Without direct influence from China, the Japanese resorted to creating their own Chinese characters as necessary. There is at least one kokuji that was subsequently borrowed by China, specifically 腺 (33-03; read *sen*, meaning "gland"). In Chinese it is read *xiàn* (GB 2312-80 47-57).

* In fact, some kokuji were even borrowed back by the Chinese as genuine Chinese characters, as you will soon learn about.

Seven kokuji have made their way into the standard set of 1,945 kanji called Jōyō Kanji, and four are in Jinmei-yō Kanji (Chapter 3 provides a full treatment of these and other related character sets). Those in Jōyō Kanji are 込 (25-94), 働 (38-15), 峠 (38-29), 畑 (40-10), 塀 (42-29), 匁 (44-72), and 枠 (47-40). Those in Jinmei-yō Kanji are 笹 (26-91), 凪 (38-68), 柾 (43-79), and 麿 (43-91). Nozomu Ohara (大原望 *ōhara nozomu*) has compiled a list of kokuji, which includes those that are included in the JIS X 0208:1997 and JIS X 0212-1990 character set standards, plus links to other kokuji-related web sites.*

Korean-Made Chinese Characters—Gugja

Like the Japanese, the Koreans have had the opportunity to create their own Chinese characters. These are known as *gugja* (국자/國字 *gugja*) in Korean. While you'd expect to find gugja only in Korean character set standards, there are approximately 100 gugja included in a Chinese character set standard designated GB 12052-89 (you'll understand why after reading about this character set standard in Chapter 3 starting on page 117).

Gugja—unlike kokuji in Japanese—have many tell-tale signs of their status as non-Chinese Chinese characters. Table 2-37 lists elements of gugja that are used to indicate a final consonant.

Table 2-37: Reading Elements of Gugja

Gugja Element	Reading
乙	L
ㄱ	G
叱	D
ㅇ	NG

Many other gugja look and feel like genuine Chinese characters. It is only after you explore their etymology that you may discover their true Korean origins.

The basic Korean character set standard for use on computers, KS X 1001:1992, includes many gugja. The supplemental Korean character set standard, KS X 1002:1991, includes even more gugja. Table 2-38 provides some examples of gugja, along with their readings and meanings (KS X 1001:1992 Row-Cell values are provided).

* *http://member.nifty.ne.jp/TAB01645/ohara/index.htm*

Table 2-38: Examples of Gugja

Gugja	Reading	Meaning
乭 42-65	갈 gal	Used in personal names
畓 51-44	답 dab	*paddy, wet field*
乭 52-44	돌 dol	Used in personal and place names
朱 56-37	말 mal	Used in place names
鐥 64-54	선 seon	Used in place names
篔 72-04	오 o	Used in place names
岾 79-32	점 jeom	*mountain pass*[a]

[a] Compare with the (Japanese) kokuji 峠 in Table 2-36 on page 62—I find it fascinating that both Japan and Korea created their own Chinese character meaning "mountain pass."

Only one gugja, 畓 (답 *dab*), is known to be included in Korea's standard set of 1,800 hanja called Sangyong Hanja—this gugja is not in the middle school subset of 900 hanja, though.

Vietnamese-Made Chinese Characters—Chữ Nôm

Unlike Japanese and Korean, in which non-Chinese Chinese characters are used together with genuine Chinese characters—a sort of mixing of scripts—Vietnamese has three distinct ways to express its language through writing:

- Latin script (called *Quốc ngữ*)
- Chinese characters (called *chữ Hán*)
- Vietnamese-made Chinese characters (called *chữ Nôm*)

Writing Vietnamese using chữ Hán is considered equivalent to writing in Chinese, not Vietnamese. Using Quốc ngữ or chữ Nôm is considered writing in Vietnamese, not Chinese. For some chữ Nôm characters, there is a corresponding chữ Hán character with the same meaning. Table 2-39 provides a handful of chữ Nôm characters, along with their chữ Hán equivalents (TCVN 5773:1993 and TCVN 6056:1995 Row-Cell codes are provided for chữ Nôm and chữ Hán, respectively).

Table 2-39: Chữ Nôm and Chữ Hán Examples

Chữ Nôm	Reading	Chữ Hán	Reading	Meaning
𠀧 21-47	ba	三 42-06	tam	*three*
𣍐 29-55	giữa	中 42-21	trung	*center, middle*
𡨸 34-02	chữ	字 50-30	tự	*character*
𣋚 35-77	trăm	百 64-02	bá	*hundred*

Because there are far fewer chữ Nôm characters than chữ Hán characters, there are times when chữ Hán characters are used in chữ Nôm context (that is, with chữ Nôm characters). Table 2-40 lists two types of chữ Hán characters: those that have different readings depending on context (chữ Nôm versus chữ Hán), and those that have identical readings regardless of context. TCVN 6056:1995 Row-Cell codes are provided for reference purposes.

Table 2-40: Chữ Hán Characters Used in Chữ Nôm Context

	Character		Chữ Nôm Reading	Chữ Hán Reading	Meaning
Unique	主	42-26	chúa	chủ	*main, primary*
	印	45-85	in	ấn	*printing*
	急	53-14	cấp	kíp	*fast, rapid*
	所	54-35	thửa	sở	*place, location*
Identical	文	56-16	văn	văn	*sentence*
	武	59-22	vũ	vũ	*weapon*
	爭	62-44	tranh	tranh	*war*
	香	76-23	hương	hương	*fragrant*

Chữ Nôm was the accepted method for writing Vietnamese since the 10th century AD. It was not until the 1920s when chữ Nôm was replaced by Quốc ngữ (see a description starting on page 38 of this chapter).

3

Character Set Standards

In this chapter:

A rock-solid understanding of and a deep appreciation for CJKV character set standards—what character classes they contain, how many characters they enumerate, how they evolved, and so on—form the foundation on which the remainder of this book is based. Without such a basic understanding, it would be pointless to discuss issues such as encoding methods, input methods, and font formats. This chapter represents what I consider to be the core of this book.

CJKV character sets can be classified into two basic types, depending on their intended purpose and reason for establishment:

- Non-coded (also known as "non-electronic") Character Sets—NCSs
- Coded (also known as "electronic") Character Sets—CCSs

"Non-coded" refers to a character set established without regard to how it would be processed on computer systems, if at all. "Coded" refers to being electronically encoded, that is, such character sets were specifically designed for processing on computer systems. You will soon realize that the characters enumerated in non-coded character sets generally constitute a subset of the characters contained in coded character sets, and sometimes affect their development.

After reading this chapter, you will have a firm understanding about which character classes constitute a particular character set, and information fundamental to dealing with CJKV-related issues.

If you are especially interested in a particular CJKV character set covered in this chapter, I encourage you to obtain the corresponding character set standard document. While this chapter provides some insights not found in the original documents, it does not (and, quite frankly, could not) duplicate all the information that those documents contain.

NOTE Some character sets discussed in this chapter are not yet established—they are in draft form, which means that their designations *may* change. Such character sets are indicated by a trailing "X" in the portion of their designation used to specify the year of establishment. Affected standards include China's GB/T 13131-9X and GB/T 13132-9X.

Non-Coded Character Set Standards

Long before there were any coded character set standards in the CJKV locales (or even before the concept of a coded character set standard existed!), several non-coded standards were defined for pedagogical purposes. These are considered to be the first attempts to limit the number of Chinese characters in common use.

The non-coded character sets described in this book include only Chinese characters. Everyone is expected to learn hiragana and katakana (in Japan) or hangul (in Korea). Only for Chinese characters, which number in the tens of thousands, is there a need to define a set (and thus, limit the number) of characters that are taught in school.

Chapter 2, *Writing Systems*, provided a brief description of Chinese characters. If you skipped that chapter and are unfamiliar with Chinese characters, I suggest going back to read it.

Hanzi in China

The educational system in China requires that students master 3,500 hanzi during their first years of instruction. These hanzi form a subset from a standardized list of 7,000 hanzi defined in 现代汉语通用字表 (*xiàndài hànyǔ tōngyòngzì biǎo*), published on March 25, 1988. We can call this large list Tōngyòng Hànzì. Two other hanzi lists further define this 3,500-hanzi subset. The first list, 现代汉语常用字表 (*xiàndài hànyǔ chángyòngzì biǎo*), defines the 2,500 hanzi that are taught during primary school. The second list, 现代汉语次常用字表 (*xiàndài hànyǔ cìchángyòngzì biǎo*), defines an additional 1,000 hanzi that are taught during middle school. We can call these character sets Chángyòng Hànzì and Cìchángyòng Hànzì. These hanzi lists are commonly abbreviated as 常用字 (*chángyòngzì*) and 次常用字 (*cìchángyòngzì*), respectively, and were published

on January 26, 1988. Appendix R provides a complete listing of the 3,500 hanzi defined in 现代汉语常用字表 and 现代汉语次常用字表. The dictionary entitled 汉字写法规范字典 (*hànzì xiěfǎ guīfàn zìdiǎn*) is useful in that it includes both sets of hanzi, and differentiates them through the use of annotations.

In addition, the Chinese government published a document, entitled *Simplified Character Table* (简化字总表 *jiǎnhuàzì zǒngbiǎo*), that enumerates 2,249 simplified hanzi (and illustrates the traditional forms from which they were derived—some simplified hanzi were derived from more than one traditional hanzi). This document is divided into three tables, the contents of which are listed in Table 3-1.

Table 3-1: Simplified Character Table Contents

Table	Characters	Description
1	350	Independently simplified hanzi
2	146	Simplified components used in other hanzi[a]
3	1,753	Hanzi simplified by using simplified components from "Table 2" of the *Simplified Character Table*

[a] Among these, 132 are also used as hanzi themselves.

There has been more than one version of this document, the most recent being published in 1986. It is important to note that its development has not been static—some minor corrections and adjustments have been made over the years, one of which is known to have caused an error in a coded character set, specifically in GB/T 12345-90 (described later in this chapter, starting on page 83). The propagation of errors from one character to another—whether coded, non-coded, or both—is something that *can* occur.

Note that there are many hanzi used in China that do not require further simplification—only those that were deemed frequently used *and* complex were simplified.

Hanzi in Taiwan

The basic set of hanzi in Taiwan is listed in a table called 常用國字標準字體表 (*chángyòng guózì biāozhǔn zìtǐ biǎo*), which enumerates 4,808 hanzi. An additional set of 6,341 hanzi is defined in 次常用國字標準字體表 (*cìchángyòng guózì biāozhǔn zìtǐ biǎo*), 18,480 rare hanzi are defined in 罕用字體表 (*hǎnyòng zìtǐ biǎo*), and 18,609 hanzi variants are defined in 異體國字字表 (*yìtǐ guózì zìbiǎo*). All of these hanzi lists were established by Taiwan's Ministry of Education (教育部 *jiàoyùbù*).

Table 3-2 lists these standards, along with their dates of establishment. These lists, when added together, create a set of 48,238 hanzi.

Table 3-2: Hanzi Lists in Taiwan

Standard	Nickname	Date of Establishment	Number of Hanzi
常用國字標準字體表	甲表 (*jiǎbiǎo*)	September 2, 1982	4,808
次常用國字標準字體表	乙表 (*yǐbiǎo*)	December 20, 1982	6,341
罕用字體表	丙表 (*bǐngbiǎo*)	October 10, 1983	18,480
異體國字字表	*none*	March 29, 1984	18,609

These hanzi lists will become useful when discussing the CNS 11643-1992 and CCCII coded character set standards from Taiwan later in this chapter, starting on pages 93 and 98, respectively. Appendix R provides a complete listing of the hanzi that make up the first two lists, 常用國字標準字體表 and 次常用國字標準字體表.

Compared to other CJKV locales, Taiwan has established non-coded character sets with the most characters.

Kanji in Japan

Non-coded Japanese character sets include Gakushū Kanji (preceded by Kyōiku Kanji)—the 1,006 kanji formally taught during the first six grades in Japanese schools; Jōyō Kanji (preceded by Tōyō Kanji)—the 1,945 kanji designated by the Japanese government as the ones to be used in public documents such as newspapers; and Jinmei-yō Kanji—the 285 kanji sanctioned by the Japanese government for use in writing personal names.* The growth and development of these character sets are listed in Table 3-3 (note that some were renamed).

Table 3-3: Evolving Kanji Lists in Japan

Year	Kyōiku Kanji	Tōyō Kanji	Jinmei-yō Kanji
1946		1,850[a]	
1948	881		
1951			92
1976			120
1977	996 (Gakushū Kanji)		
1981		1,945 (Jōyō Kanji)	166
1990			284
1992	1,006[b]		
1997			285

[a] The corresponding glyph table (当用漢字字体表 *tōyō kanji jitai hyō*) was published in 1949, and likewise, the corresponding reading table (当用漢字音訓表 *tōyō kanji onkun hyō*) was published in 1948.
[b] Established in 1989, but not fully implemented until 1992.

* The 285th kanji added to this list is 琉 (JIS X 0208:1997 46-16).

There is some overlap among these character sets. Gakushū Kanji is a subset of Jōyō Kanji (likewise, Kyōiku Kanji was a subset of Tōyō Kanji).

Table 3-4 shows how you write the names of these character sets in native Japanese orthography, and indicates their meaning.

Table 3-4: The Meanings of Non-Coded Japanese Character Set Standards

Character Set	In Japanese	Meaning	Content
Kyōiku Kanji	教育漢字	*Instructional kanji*	881
Gakushū Kanji	学習漢字	*Educational kanji*	1,006
Tōyō Kanji	当用漢字	*Common use kanji*	1,850
Jōyō Kanji	常用漢字	*Everyday use kanji*	1,945
Jinmei-yō Kanji	人名用漢字	*Personal name use kanji*	285

While Table 3-3 appears to show that the Gakushū Kanji list gained only ten kanji between 1977 and 1992, the list also experienced some internal shifts. Gakushū Kanji and Kyōiku Kanji can be decomposed into six sets, each corresponding to the grade of school during which they are formally taught. Table 3-5 indicates the six grade levels on the left, along with the number of kanji taught during each one—this is done for Kyōiku Kanji and both versions of Gakushū Kanji.

Table 3-5: The Development of Gakushū Kanji

Grade	1958 (881 Kanji)[a]	1977 (996 Kanji)	1992 (1,006 Kanji)
1	46	76	80
2	105	145	160
3	187	195	200
4	205	195	200
5	194	195	185
6	144	190	181

[a] Kyōiku Kanji was not divided into the six grade levels until 1958.

The general trend shown by Table 3-5 is that more kanji (although not significantly more) are now taught in the earlier grades.

Appendix R provides complete listings of the Jōyō Kanji, Gakushū Kanji, and Jinmei-yō Kanji character sets.

Hanja in Korea

Korea has defined a list of hanja called *Sangyong Hanja* (상용 한자/常用漢字 *sangyong hanja*), and enumerates the 1,800 hanja that students are expected to learn during their school years. The first 900 of these hanja are expected to be

learned by students during middle school—the remaining 900 are expected to be learned through high school. These hanja lists were established on August 16, 1972.

When coding these 1,800 hanja electronically, the list expands to 1,953 hanja due to the duplicate hanja in the KS X 1001:1992 character set (covered later in this chapter). Likewise, the list of 900 middle-school hanja expands to 978 hanja for the same reason. Appendix R provides a complete printout of the 1,800 Sangyong Hanja and the 900 hanja taught during middle school (expanded to 1,953 and 978 hanja, respectively, to accommodate the KS X 1001:1992 character set standard—later in this chapter, starting on page 111, you'll know and appreciate why this expansion is necessary).

The Korean Supreme Court (대법원/大法院 *daebeobweon*) also defined, at various periods, lists of hanja that are considered acceptable for use in writing Korean names—these lists are called Inmyeong-yong Hanja (인명용 한자/人名用漢字 *inmyeongyong hanja*). The latest list enumerates 2,964 hanja, and was established in July of 1994. Previous versions of this list were established in January and March of 1991.

Coded Character Set Standards

Proliferation of computer systems necessitated the creation of coded character set standards. Initially, each vendor (such as IBM, Fujitsu, Hitachi, and so on) established their own corporate standard for their products alone. However, the first multiple-byte national coded character set standard among the CJKV locales was established by the Japanese Standards Association (JSA) on January 1, 1978, and was designated JIS C 6226-1978. Without a doubt, the birth of this character set standard sent waves throughout the CJKV locales.

Other CJKV locales, such as Korea and China, being inspired by the success of JIS C 6226-1978, followed soon after by imitating the Japanese solution, and in some cases copied more than merely the encoding method or arrangement of characters. For example, in the case of Taiwan's Big Five character set, it has been claimed that the Taiwanese borrowed many Chinese character forms from Japan's JIS C 6226-1978.

Character Set Standards Overview

The character set standards described in this section constitute those maintained by a government or a government-sanctioned organization within a given country, and are considered the standard character sets for the locale. In addition, some character set standards form the foundation from which other character set stan-

dards are derived, such as international or vendor character set standards. (Vendor character set standards are covered in Appendix C, *Vendor Character Set Standards*.)

Tables 3-6 through 3-11 summarize the national character sets described in this chapter, along with the number and classes of characters enumerated by each. I have decided to use separate tables for each locale because one large table would have been far too overwhelming.

Table 3-6: Chinese Character Set Standards—China

Character Set	Level 1	Level 2	Extra Hanzi	Symbols	Control Codes
GB 1988-89[a]				94	34
GB 2312-80	3,755	3,008		682	
GB 6345.1-86	3,755	3,008		814	
GB 8565.2-88	3,755	3,008	636	751	
ISO-IR-165:1992	3,755	3,008	775	905	
GB/T 12345-90	3,755	3,008	103	843	
GB 7589-87	7,237				
GB/T 13131-9X	7,237				
GB 7590-87	7,039				
GB/T 13132-9X	7,039				
GBK	3,755	3,008	14,240	883	

[a] Also known as GB-Roman.

Table 3-7: Chinese Character Set Standards—Taiwan

Character Set	Level 1	Level 2	Extra Hanzi	Symbols	Control Codes
Big Five	5,401	7,652		441	
Big Five Plus	5,401	7,652	7,619	913	
CNS 5205-1989[a]				94	34
CNS 11643-1986	5,401	7,650	13,488[b]	684	
CNS 11643-1992	5,401	7,650	34,976	684	
CCCII[c]	75,684				

[a] Also known as CNS-Roman.
[b] Planes 14 and 15.
[c] The "Level 1" figure represents the total number of characters

Table 3-8: Japanese Character Set Standards

Character Set	Level 1	Level 2	Extra Kanji	Symbols	Control Codes
JIS X 0201-1997[a]				157[b]	34
JIS C 6226-1978	2,965	3,384		453	

Table 3-8: Japanese Character Set Standards (continued)

Character Set	Level 1	Level 2	Extra Kanji	Symbols	Control Codes
JIS X 0208-1983	2,965	3,384	4	524	
JIS X 0208-1990	2,965	3,384	6	524	
JIS X 0208:1997	2,965	3,384	6	524	
JIS X 0212-1990	5,801			266	
JIS X 0213:2000	1,249[c]	2,436[d]		659	

[a] Part of this standard includes JIS-Roman.
[b] This figure includes 94 JIS-Roman characters plus 63 half-width katakana characters.
[c] JIS Level 3.
[d] JIS Level 4.

Table 3-9: Korean Character Set Standards

Character Set	Country	Hangul	Hanja	Symbols	Control Codes
KS X 1003:1993[a]	South Korea			94	34
KS X 1001:1992	South Korea	2,350	4,888	986	
KS X 1002:1991	South Korea	3,605[b]	2,856	1,188	
KPS 9566-97	North Korea	2,679	4,653	927	
GB 12052-89	China	5,203[c]	94	682	

[a] Also known as KS-Roman.
[b] These 3,605 hangul are split into two levels, enumerating 1,930 and 1,675 characters, respectively. The second set of hangul (1,675 characters) are considered to be ancient hangul.
[c] These 5,203 hangul are split into three levels, enumerating 2,068, 1,356, and 1,779 characters each.

Table 3-10: Vietnamese Character Set Standards

Character Set	Chinese Characters	Symbols	Control Codes
TCVN 5712:1993[a]		233[b]	34
TCVN 5773:1993	2,357		
TCVN 6056:1995	3,311		

[a] Also known as TCVN-Roman.
[b] This figure includes 94 ASCII characters plus 139 additional (mostly accented) characters, 5 of which are combining marks.

Table 3-11: Other National Character Set Standards

Character Set	Country	Total Characters	Control Codes
ASCII	USA	94	34
ANSI Z39.64-1989	USA	15,686	
Hong Kong GCCS	Hong Kong	3,049	

What a list of standards, eh? If you read this chapter carefully, Tables 3-6 through 3-11 will no longer seem overwhelming. They are also useful for general reference, so be sure to dog-ear these pages.

The national standards that are based on ISO 10646-1:1993, specifically GB 13000.1-93, JIS X 0221-1995, and KS X 1005-1:1995, are covered in the section entitled "International Character Set Standards," beginning on page 120.

The terms Level 1 and Level 2 have not yet been described. They simply refer to the two such groups of Chinese characters usually defined within each CJKV character set standard. Level 1 typically contains frequently-used Chinese characters, whereas Level 2 contains less–frequently-used Chinese characters. Some character sets, such as JIS X 0212-1990 and CNS 11643-1992, contain only a single block of Chinese characters, or else consist of multiple planes.

ASCII

Most readers of this book are familiar with the ASCII encoding, so it is a good place to begin our discussion of coded character set standards, and will serve as a common point of reference.

The ASCII character set is covered in this book because it is quite often mixed with CJKV characters within text. Note, however, that the ASCII character set standard is not specific to any CJKV locale.

ASCII stands for *American Standard Code for Information Interchange*. The ASCII character set standard is described in the standard designated ANSI X3.4-1986;* it is the US version and at the same time the International Reference Version (IRV) of ISO 646:1991,† which defines the framework for similar national standards.

The ASCII character set is composed of 128 characters, 94 of which are considered printable. There are also 34 other characters, which include a space character and many control characters (such as tab, escape, shift-in, and so on), which are defined in ISO 6429:1992, *Information Technology—Control Functions for Coded Character Sets*. The control codes are technically not part of ASCII or ISO 646:1991. Table 3-12 lists the 94 printable ASCII characters.

Table 3-12: The ASCII Character Set

Lowercase Latin	abcdefghijklmnopqrstuvwxyz
Uppercase Latin	ABCDEFGHIJKLMNOPQRSTUVWXYZ
Numerals	0123456789
Symbols	!"#$%&'()*+,-./:;<=>?@[\]^_`{|}~

* ANSI is short for *American National Standards Institute*; an earlier version of this standard was designated ANSI X3.4-1977.

† ISO is short for *International Organization for Standardization*; an earlier version of this standard was designated ISO 646:1983.

Most of these printable characters are also used in EBCDIC, an encoding method covered in the next chapter. The binary nature of computers allows these 128 characters to be represented using seven bits, but because computers evolved through processing information in eight-bit segments (a typical *byte*), these 128 ASCII characters are usually represented by eight-bit units in which the eighth bit (also known as the highest-order bit) is usually set to zero. Other character sets often incorporate the characters of ASCII.

ASCII Variations

There are, as of this writing, ten variations of the ASCII character sets, all approved by and published through ISO. These character sets contain the ASCII character set as their common base, plus additional characters. Extended ASCII character sets are used to represent other writing systems, such as Arabic, Hebrew, and Cyrillic. There is also an extensive collection of additional Latin characters. These characters are usually additional symbols and accented versions of Latin characters.

Eight-bit representations can theoretically handle 128 more characters than seven-bit representations—the reality is that they handle only up to 94 or 96 additional characters. The documents ISO 8859 Parts 1 through 10 (*Information Processing—8-Bit Single-Byte Coded Graphic Character Sets*) describe character sets that can be encoded in the additional 128 positions when an eight-bit representation is used.* Table 3-13 lists the contents of each of the ten parts of ISO 8859, indicating what languages are supported by each.

Table 3-13: The Ten Parts of ISO 8859

Part	Year	Contents	Languages
1	1998	Latin alphabet No. 1	Danish, Dutch, English, Faeroese, Finnish, French, German, Icelandic, Irish, Italian, Norwegian, Portuguese, Spanish, Swedish
2	1987	Latin alphabet No. 2	Albanian, Czech, English, German, Hungarian, Polish, Rumanian, Serbo-Croatian, Slovak, Slovene
3	1988	Latin alphabet No. 3	Afrikaans, Catalan, Dutch, English, Esperanto, German, Italian, Maltese, Spanish, Turkish
4	1998	Latin alphabet No. 4	Danish, English, Estonian, Finnish, German, Greenlandic, Lappish, Latvian, Lithuanian, Swedish, Norwegian
5	1988	Latin/Cyrillic alphabet	Bulgarian, Byelorussian, English, Macedonian, Russian, Serbo-Croatian, Ukrainian

* *http://czyborra.com/charsets/iso8859.html*

Table 3-13: The Ten Parts of ISO 8859 (continued)

Part	Year	Contents	Languages
6	1987	Latin/Arabic alphabet	Arabic
7	1987	Latin/Greek alphabet	Greek
8	1988	Latin/Hebrew alphabet	Hebrew
9	1989	Latin alphabet No. 5	Danish, Dutch, English, Finnish, French, German, Irish, Italian, Norwegian, Portuguese, Spanish, Swedish, Turkish
10	1998	Latin alphabet No. 6	Danish, English, Estonian, Finnish, German, Greenlandic, Lappish, Latvian, Lithuanian, Swedish, Norwegian

Table 3-14 lists the 95 additional non-ASCII characters from ISO 8859-1:1998 (also known as ISO Latin-1 or ISO-8859-1). Appendix S, *Single-Byte Code Tables*, provides a complete ISO 8859-1:1998 code table.

Table 3-14: ISO 8859-1:1998 Character Samples

Lowercase Latin	àáâãäåæçèéêëìíîïðñòóôõöøùúûüýþÿ
Uppercase Latin	ÀÁÂÃÄÅÆÇÈÉÊËÌÍÎÏÐÑÒÓÔÕÖØÙÚÛÜÝÞß
Symbols	¡¢£¤¥§¨©ª«¬-®¯°±²³´µ¶•¸¹º»¼½¾¿×÷

These characters, as you can probably guess, are not that useful when working with CJKV text. This table simply illustrates the types of characters available in the ISO 8859 series. Note again that these additional ASCII character sets require a full eight bits per character for encoding because they contain far more than 128 characters.

CJKV-Roman

The Chinese, Japanese, Koreans, and Vietnamese have developed their own variants of the ASCII character set, known as GB-Roman (from GB 1988-89), CNS-Roman (from CNS 5205-1989), JIS-Roman (from JIS X 0201-1997), KS-Roman (from KS X 1003:1993), and TCVN-Roman (from TCVN 5712:1993), respectively. Or, CJKV-Roman, collectively. These character sets, like ASCII, consist of 94 printable characters, but there are some minor differences.* The characters that differ are indicated in Table 3-15 (full-width forms are used for illustrative purposes).

Because the difference between ASCII and the CJKV-Roman character sets is minor, they are usually treated as the same throughout this book. You will also find that most terminal software supports only one of these character sets. This

* But one, specifically TCVN 5712:1993, contains more than these 94 characters.

Table 3-15: Special CJKV-Roman Characters

Code	ASCII[a]	GB-Roman	CNS-Roman[b]	JIS-Roman	KS-Roman
0x24	$ (dollar)	¥ (yuan)	$	$	$
0x5C	\ (backslash)	\	\	¥ (yen)	₩ (won)
0x7E	~ (tilde)[c]	‾ (overline)	‾	‾	‾

[a] TCVN-Roman is identical to ASCII as far as these three characters are concerned
[b] CNS-Roman is ambiguous with regard to glyphs. The glyphs shown in this column were made consistent with the other CJKV-Roman character sets.
[c] The vertical positioning of the tilde may vary depending on the implementation.

means that terminals which support only JIS-Roman display the ASCII backslash as the JIS-Roman yen symbol. For systems that require the backslash, such as MS-DOS for indicating directory hierarchy, the yen symbol is used instead. Stranger yet, Perl programs displayed on a terminal that supports GB-Roman would have variables prefixed with yuan symbols (instead of the customary dollar sign). You will also find that most CJKV software supports CJKV-Roman instead of ASCII. It is possible that the computer supports both ASCII and CJKV-Roman. Changing the display from CJKV-Roman to ASCII (and vice versa), though, may be as simple as changing the display font. You will learn in the next chapter that this is because ASCII and CJKV-Roman almost always occupy the same encoding space, which can actually lead to code conversion problems when dealing with Unicode.

It is important to realize that character set standards do *not* prescribe the widths of characters—it is simply customary to associate characters with specific widths, usually half- and full-width.

The document designated GB 1988-89, *Information Processing—7-Bit Coded Character Set for Information Interchange* (信息处理—信息交换用七位编码字符集 *xìnxī chǔlǐ xìnxī jiāohuàn yòng qīwèi biānmǎ zìfújí*), established on July 1, 1990, contains the definition of the GB-Roman character set.* This manual is virtually identical to ISO 646:1991 except that it is written in Chinese.

The document designated CNS 5205-1989, *Information Processing—7-Bit Coded Character Set for Information Interchange* (資訊處理及交換用七數元碼字元集 *zīxùn chǔlǐ jí jiāohuàn yòng qīshùyuán mǎzìyuánjí*), contains the definition of the CNS-Roman character set.† This manual is virtually identical to ISO 646:1991 except that it is written in Chinese.

The document designated JIS X 0201-1997, *7-Bit and 8-Bit Coded Character Sets for Information Interchange* (7ビット及び8ビットの情報交換用符号化文字集合 *nana-bitto oyobi hachi-bitto no jōhō kōkan yō fugōka moji shūgō*), established on January 20, 1997, provides the definition for the JIS-Roman character set.‡ Like GB

* The original version of this standard was designated GB 1988-80.

† Earlier versions of this standard were dated 1980, 1981, and 1983.

‡ JIS X 0201-1997 was formerly designated JIS X 0201-1976 (which itself was reaffirmed in 1984 and in 1989).

1988-89, this manual is virtually identical to ISO 646:1991 except that it is written in Japanese, and defines the extensions for half-width katakana.

The document designated KS X 1003:1993, *Code for Information Interchange* (정보 교환용 부호 (로마 문자) *jeongbo gyohwanyong buho (roma munja)*), established on January 6, 1993, contains the definition of the KS-Roman character set.* Like GB 1988-89, this manual is identical to ISO 646:1991 except that it is written in Korean.

The document designated TCVN 5712:1993, *Công Nghệ Thông Tin—Bộ Mã Chuẩn 8-Bit Kí Tự Việt Dùng Trong Trao Đổi Thông Tin* (*Information Technology—Vietnamese 8-Bit Standard Coded Character Set for Information Interchange*), established on May 12, 1993, contains the definition of the TCVN-Roman character set. TCVN-Roman contains the basic 94 ASCII characters plus up to 139 additional characters, most of which are adorned with diacritic marks (and represent all possible Quốc ngữ characters). Five of these 139 additional characters are combining marks that indicate tone.

Chinese Character Set Standards—China

As you learned earlier, Japan was first to develop and implement a multiple-byte national character set. The other major CJKV locales—China, Taiwan, and Korea—soon followed by developing their own. This section describes the character set standards established by China, or more specifically, the People's Republic of China or PRC (中华人民共和国 *zhōnghuá rénmín gònghé guó*).

All Chinese character set standards begin with the designator GB, which stands for "Guo Biao" (国标 *guóbiāo*), which is short for "Guojia Biaozhun" (国家标准 *guójiā biāozhǔn*), and means "National Standard." Some GB standards have a "/T" tacked onto the "GB" to form "GB/T." The "T" here stands for "Tui" (推 *tuī*), which is short for "Tuijian" (推荐 *tuījiàn*), and means "recommended" (as opposed to "forced" or "mandatory"). The "T" does *not* stand for "traditional" (as in "traditional hanzi").

GB 2312-80

This character set standard, established on May 1, 1981 by the People's Republic of China (PRC), enumerates 7,445 characters. Its official name is *Code of Chinese Graphic Character Set for Information Interchange Primary Set* (信息交换用汉字编

* This standard was previously designated KS C 5636-1993. The original version, KS C 5636-1989, was established on April 22, 1989.

码字符集—基本集 *xìnxī jiāohuàn yòng hànzì qīwèi biānmǎ zìfújí—jīběnjí*). Table 3-16 lists how characters are allocated to each row.

Table 3-16: The GB 2312-80 Character Set

Row	Characters	Content
1	94	Miscellaneous symbols
2	72	Numerals 1–20 with period, parenthesized numerals 1–20, encircled numerals 1–10, parenthesized hanzi numerals 1–20, uppercase Roman numerals 1–12
3	94	Full-width GB 1988-89 (GB-Roman; equivalent to ASCII)
4	83	Hiragana
5	86	Katakana
6	48	Upper- and lowercase Greek alphabet
7	66	Upper- and lowercase Cyrillic alphabet
8	63	26 full-width pinyin characters, 37 zhuyin (bopomofo) characters
9	76	Line-drawing elements
10–15	0	Unassigned
16–55	3,755	Level 1 hanzi (last is 55-89)
56–87	3,008	Level 2 hanzi (last is 87-94)
88–94	0	Unassigned

Level 1 hanzi (第一级汉字 *dìyījí hànzì*) are arranged by reading. Level 2 hanzi (第二级汉字 *dì'èrjí hànzì*) are arranged by radical, then total number of strokes. To give you a feel for the GB 2312-80 character set, Table 3-17 briefly illustrates the types of characters in GB 2312-80.

Table 3-17: GB 2312-80 Character Samples

Character Class	Sample Characters		
Miscellaneous symbols	、 。 · ˉ ˇ ¨ 〃 々 —	…	□■△▲※→←↑↓〓
Annotated numerals	1. 2. 3. 4. 5. 6. 7. 8. 9. 10.	…	Ⅲ Ⅳ Ⅴ Ⅵ Ⅶ Ⅷ Ⅸ Ⅹ Ⅺ Ⅻ
Full-width GB-Roman	! " # ¥ % & ′ ( ) *	…	u v w x y z { | } ¯
Hiragana	ぁ あ ぃ い ぅ う ぇ え ぉ お	…	り る れ ろ ゎ わ る ゑ を ん
Katakana	ァ ア ィ イ ゥ ウ ェ エ ォ オ	…	ロ ヮ ワ ヰ ヱ ヲ ン ヴ ヵ ヶ
Greek characters	Α Β Γ Δ Ε Ζ Η Θ Ι Κ	…	ο π ρ σ τ υ φ χ ψ ω
Cyrillic	А Б В Г Д Е Ё Ж З И	…	ц ч ш щ ъ ы ь э ю я
Full-width pinyin	ā á ǎ à ē é ě è ī í	…	ǔ ù ü ê ɑ ḿ ń ň ǹ ɡ
Zhuyin (bopomofo)	ㄅ ㄆ ㄇ ㄈ ㄉ ㄊ ㄋ ㄌ ㄍ ㄎ	…	ㄠ ㄡ ㄢ ㄣ ㄤ ㄥ ㄦ ㄧ ㄨ ㄩ
Line-drawing elements	― — | ┃ ┄ ┅ ┆ ┇ ┈ ┉	…	┴ ┸ ┵ ┹ ┼ ╀ ┽ ╁ ┾ ╂
Level 1 hanzi	啊 阿 埃 挨 哎 唉 哀 皑 癌 蔼	…	尊 遵 昨 左 佐 柞 做 作 坐 座
Level 2 hanzi	亍 丌 兀 丐 廿 卅 丕 亘 丞 鬲	…	黲 黯 齘 鼬 齾 齴 齵 龘 龑 虌

Encoding methods for GB 2312-80 (and its extensions, described shortly) include ISO-2022-CN, ISO-2022-CN-EXT, EUC-CN, and GBK.

CJKV font developers should be aware that early printings of the GB 2312-80 manual had the code points of two uppercase Cyrillic characters (in row 7) swapped. Table 3-18 illustrates the incorrect and correct order of these characters in GB 2312-80. Note the different ordering of the two uppercase Cyrillic characters Ф (Row-Cell 07-22) and Х (Row-Cell 07-23), both of which have been underlined.

Table 3-18: Uppercase Cyrillic Character Ordering in GB 2312-80

Incorrect	А Б В Г Д Е Ё Ж З И Й К Л М Н О П Р С Т У Х Ф Ц Ч Ш Щ Ъ Ы Ь Э Ю Я
Correct	А Б В Г Д Е Ё Ж З И Й К Л М Н О П Р С Т У Ф Х Ц Ч Ш Щ Ъ Ы Ь Э Ю Я

I have encountered at least one Chinese type foundry whose font data propagates the character-ordering error illustrated in Table 3-18.

There are three common extensions to GB 2312-80, one of which was used to issue two corrections. Table 3-19 illustrates the number of characters in GB 2312-80 and its three extensions.

Table 3-19: GB 2312-80 and Its Three Extensions

Character Set	Characters	Characters Added	Number of Corrections
GB 2312-80	7,445		
GB 6345.1-86	7,577	132	2
GB 8565.2-88	8,150	705	
ISO-IR-165:1992	8,443	998	

These extensions to the GB 2312-80 character set standard are described in the following sections.

GB 6345.1-86—corrections and extensions to GB 2312-80

Corrections for and additions to GB 2312-80 have been issued through a separate character set standard designated GB 6345.1-86, established on December 1, 1986. This standard is entitled *32×32 Dot Matrix Font Set of Chinese Ideograms for Information Interchange* (信息交换用汉字32×32点阵字模集 *xìnxī jiāohuàn yòng hànzì 32×32 diǎnzhèn zìmújí*), and resulted in 132 additional characters for a new total of 7,577 characters (6,763 hanzi plus 814 non-hanzi). Table 3-20 highlights the additional characters for GB 2312-80 specified by GB 6345.1-86.

While Table 3-20 clearly shows what characters were added to GB 2312-80, it does not list the corrections. Table 3-21 shows the two corrections to GB 2312-80 mandated by GB 6345.1-86.

Table 3-20: The GB 6345.1-86 Character Set

Row	Characters	Content
1	94	Miscellaneous symbols
2	72	Numerals 1–20 with period, parenthesized numerals 1–20, encircled numerals 1–10, parenthesized hanzi numerals 1–20, uppercase Roman numerals 1–12
3	94	Full-width GB 1988-89 (GB-Roman; equivalent to ASCII)
4	83	Hiragana
5	86	Katakana
6	48	Upper- and lowercase Greek alphabet
7	66	Upper- and lowercase Cyrillic alphabet
8	69	32 full-width pinyin characters, 37 zhuyin (bopomofo) characters
9	76	Line-drawing elements
10	94	Half-width GB 1988-89 (GB-Roman; equivalent to ASCII)
11	32	Half-width pinyin characters
12–15	0	Unassigned
16–55	3,755	Level 1 hanzi (last is 55-89)
56–87	3,008	Level 2 hanzi (last is 87-94)
88–94	0	Unassigned

Table 3-21: GB 6345.1-86 Corrections

Row-Cell	GB 2312-80	GB 6345.1-86
03-71	g	g
79-81	鍾	锺

The GB 2312-80 character form for Row-Cell 79-81 happens to be the same as in GB/T 12345-90, that is, the traditional form, and at the same code point. GB/T 12345-90 is described shortly. This error is still found in recent publications that list all GB 2312-80 hanzi, so evidently information about this correction is not yet widely known.

GB 8565.2-88—another extension to GB 2312-80

The GB 8565.2-88 standard, established on July 1, 1988, defines additions to the GB 2312-80 character set. This standard is entitled *Information Processing—Coded Character Sets for Text Communication—Part 2: Graphic Characters* (信息处理—文本通信用编码字符集—第二部分—图形字符集 *xìnxī chǔlǐ—wénběn tōngxìn yòng biānmǎ zìfújí—dì'èr bùfen—túxíng zìfújí*). These additions, however, are independent from those specified by GB 6345.1-86. The number of additional characters totals 705, bringing the total number of characters to 8,150 (7,399 hanzi plus 751 non-hanzi).

Table 3-22 provides a listing of characters in GB 8565.2-88, and those above and beyond GB 2312-80 are highlighted.

Table 3-22: The GB 8565.2-88 Character Set

Row	Characters	Content
1	94	Miscellaneous symbols
2	72	Numerals 1–20 with period, parenthesized numerals 1–20, encircled numerals 1–10, parenthesized hanzi numerals 1–20, uppercase Roman numerals 1–12
3	94	Full-width GB 1988-89 (GB-Roman; equivalent to ASCII)
4	83	Hiragana
5	86	Katakana
6	48	Upper- and lowercase Greek alphabet
7	66	Upper- and lowercase Cyrillic alphabet
8	63	26 full-width pinyin characters, 37 zhuyin (bopomofo) characters
9	76	Line-drawing elements
10–12	0	Unassigned
13	50	Hanzi from GB 7589-87 (last is 13-50)
14	92	Hanzi from GB 7590-87 (last is 14-92)
15	93	69 non-hanzi plus 24 hanzi (last is 15-93)
16–55	3,755	Level 1 hanzi (last is 55-89)
56–87	3,008	Level 2 hanzi (last is 87-94)
88–89	0	Unassigned
90–94	470	Hanzi from GB 7589-87 (last is 94-94)

Note how GB 8565.2-88 does not include the additions specified by GB 6345.1-86. But, it does include its corrections as shown in Table 3-21 on page 81.

ISO-IR-165:1992—yet another extension to GB 2312-80

ISO-IR-165:1992, also known as the CCITT (Consultative Committee on International Telephone and Telegraph) Chinese Set, enumerates 8,443 characters.* It is based on the GB 2312-80 character set, and includes all modifications and additions specified in GB 6345.1-86 and GB 8565.2-88. That is, 7,445 characters from GB 2312-80, 132 added due to GB 6345.1-86, 705 added due to GB 8565.2-88, plus 161 added by ISO-IR-165:1992.

Table 3-23 provides a listing of characters in ISO-IR-165:1992, and those rows that have content above and beyond GB 2312-80 are highlighted.

* ISO-IR-165:1992 is short for *ISO International Registry #165*, established on July 13, 1992.

Table 3-23: The ISO-IR-165:1992 Character Set

Row	Characters	Content
1	94	Miscellaneous symbols
2	72	Numerals 1–20 with period, parenthesized numerals 1–20, encircled numerals 1–10, parenthesized hanzi numerals 1–20, uppercase Roman numerals 1–12
3	94	Full-width GB 1988-89 (GB-Roman; equivalent to ASCII)
4	83	Hiragana
5	86	Katakana
6	70	48 upper- and lowercase Greek alphabet, 22 background (shading) characters
7	66	Upper- and lowercase Cyrillic alphabet
8	69	32 full-width pinyin characters, 37 zhuyin (bopomofo) characters
9	76	Line-drawing elements
10	94	Half-width GB 1988-89 (GB-Roman; equivalent to ASCII)
11	32	Half-width pinyin characters
12	94	94 hanzi (last is 12-94)
13	94	50 hanzi from GB 7589-87 plus 44 hanzi (last is 13-94)
14	92	Hanzi from GB 7590-87 (last is 14-92)
15	94	69 non-hanzi plus 25 hanzi (last is 15-94)
16–55	3,755	Level 1 hanzi (last is 55-89)
56–87	3,008	Level 2 hanzi (last is 87-94)
88–89	0	Unassigned
90–94	470	Hanzi from GB 7589-87 (last is 94-94)

ISO-IR-165:1992 is, as you can see, a superset of GB 2312-80 and all previous extensions thereof.

GB/T 12345-90—the traditional analog of GB 2312-80

This character set standard, established on December 1, 1990 by the People's Republic of China, enumerates 7,709 characters (6,866 hanzi plus 843 non-hanzi). Its official name is *Code of Chinese Ideogram Set for Information Interchange Supplementary Set* (信息交换用汉字编码字符集—辅助集 *xìnxī jiāohuàn yòng hànzì biānmǎ zìfújí—fǔzhùjí*). Table 3-24 lists how characters are allocated to each row. Note the similarities to GB 2312-80, and that the GB 6345.1-86 additions are included.

As was the case with GB 2312-80, Level 1 hanzi are arranged by reading, and Level 2 hanzi are arranged by radical and total number of strokes. The 103 additional hanzi are arranged by the order in which their counterparts from Level 1

Table 3-24: The GB/T 12345-90 Character Set

Row	Characters	Content
1	94	Miscellaneous symbols
2	72	Numerals 1–20 with period, parenthesized numerals 1–20, encircled numerals 1–10, parenthesized hanzi numerals 1–20, uppercase Roman numerals 1–12
3	94	Full-width GB 1988-89 (GB-Roman; equivalent to ASCII)
4	83	Hiragana
5	86	Katakana
6	77	48 upper- and lowercase Greek alphabet, 29 vertical-use characters
7	66	Upper- and lowercase Cyrillic alphabet
8	69	32 full-width pinyin characters, 37 zhuyin (bopomofo) characters
9	76	Line-drawing elements
10	94	Half-width GB 1988-89 (GB-Roman; equivalent to ASCII)
11	32	Half-width pinyin characters
12–15	0	Unassigned
16–55	3,755	Level 1 hanzi (last is 55-89)
56–87	3,008	Level 2 hanzi (last is 87-94)
88–89	103	Additional hanzi (last is 89-09)
90–94	0	Unassigned

and 2 hanzi appear. Table 3-25 briefly illustrates the types of characters in GB/T 12345-90.

Table 3-25: GB/T 12345-90 Character Samples

Character Class	Sample Characters		
Miscellaneous symbols	、。・ˉˇ¨〃々—	…	□■△▲※→←↑↓〓
Annotated numerals	1. 2. 3. 4. 5. 6. 7. 8. 9. 10.	…	Ⅲ Ⅳ Ⅴ Ⅵ Ⅶ Ⅷ Ⅸ Ⅹ Ⅺ Ⅻ
Full-width GB-Roman	! " # ¥ % & ′ ( ) *	…	u v w x y z { | } ¯
Hiragana	ぁあぃいぅうぇえぉお	…	りるれろゎわるゑをん
Katakana	ァアィイゥウェエォオ	…	ロヮワヰヱヲンヴヵヶ
Greek characters	Α Β Γ Δ Ε Ζ Η Θ Ι Κ	…	ο π ρ σ τ υ φ χ ψ ω
Vertical-use characters	’ 。 、 ： ； ！ ？ ︵ ︶ ﹇	…	﹈ ﹉ ﹊ ﹋ ﹌ ﹍ ｜ ⋮ ｜ ｝
Cyrillic characters	А Б В Г Д Е Ё Ж З И	…	ц ч ш щ ъ ы ь э ю я
Full-width pinyin	ā á ǎ à ē é ě è ī í	…	ǔ ǜ ü ê ɑ ḿ ń ň ǹ ɡ
Zhuyin (bopomofo)	ㄅ ㄆ ㄇ ㄈ ㄉ ㄊ ㄋ ㄌ ㄍ ㄎ	…	ㄠ ㄡ ㄢ ㄣ ㄤ ㄥ ㄦ ㄧ ㄨ ㄩ
Line-drawing elements	——| | ┄┄ ┆ ┆ ┈┈	…	┴ ┬ ┤ ├ ┼ ┼ ┼ ┼ ┼ ┼
Half-width GB-Roman	! " # ¥ % & ′ ( ) *	…	u v w x y z { | } ¯
Half-width pinyin	ā á ǎ à ēé ě è ī í	…	ǔǜ ü ê ɑ ḿ ń ň ǹ ɡ

Table 3-25: GB/T 12345-90 Character Samples (continued)

Character Class	Sample Characters
Level 1 hanzi	啊阿埃挨哎唉哀皑癌藹 … 尊遵昨左佐柞做作坐座
Level 2 hanzi	亍丌兀丐廿卅丕亘丞鬲 … 黲黯黝黜黳黶黷黹黻黼
Additional hanzi	襬闆錶彆葛纔厂冲丑齣 … 髒症隻只緻製种硃筑准

Compare Level 1 and 2 hanzi from Table 3-25 with that for GB 2312-80 in Table 3-17 on page 79, and note how the same hanzi are used, but that a handful are in the traditional form. In fact, there are 2,180 traditional hanzi forms in GB/T 12345-90 when compared to GB 2312-80, most of which are replacements for simplified hanzi.

The 2,180 hanzi that are used to transform GB 2312-80 into GB/T 12345-90 can be classified into the two classes, as indicated in Table 3-26.

Table 3-26: GB/T 12345-90 Characters Not in GB 2312-80

Characters	Class
2,118	Traditional hanzi replacements—rows 16 through 87
62	Additional hanzi—scattered throughout rows 88 and 89

In addition to the above replacements and additions, 41 hanzi from GB 2312-80 rows 16 through 87 are scattered throughout GB/T 12345-90 rows 88 and 89, and four pairs of hanzi between Level 1 and Level 2 hanzi were swapped. Appendix Q, *Character Lists and Mapping Tables*, provides more details about the four pairs of swapped hanzi and the mappings for hanzi in rows 88 and 89—it also includes a long and complete listing of the 2,118 traditional hanzi replacements, which is something that even the GB/T 12345-90 does not provide.

Like other character set standards, GB/T 12345-90 is not without errors. Chinese type foundries should take note that the GB/T 12345-90 manual has at least two (but, unfortunately, generally not known) printing errors, as indicated in Table 3-27.

Table 3-27: GB/T 12345-90 Corrections

Original	Corrected	Row-Cell	Original in Unicode	Original in GBK
隸	隸	33-05	96B7	EB5F
鳧	鳧	57-76	9CE7	F844

In addition, there is often some misunderstanding of the scope and content of the GB/T 12345-90 character set standard. Some printouts of the GB/T 12345-90 character set use slightly different glyphs from the official standard. One specific instance of GB/T 12345-90 provided to The Unicode Consortium used 22 different

glyphs, each of which has a *different* Unicode code point. This causes lots of confusion. Table 3-28 lists these characters, along with their (incorrect) Unicode mappings and GBK cross-references. For all 22 of these characters, their glyphs in GB/T 12345-90 are intended to be identical to those in GB 2312-80.

Table 3-28: Incorrect Mappings Between GB/T 12345-90 and Unicode

Correct	GB 2312-80 and GB/T 12345-90	Incorrect	Unicode	GBK
叠	21-94	疊	758A	AF42
换	27-27	換	63DB	9351
唤	27-29	喚	559A	86BE
痪	27-30	瘓	7613	AF88
焕	27-32	煥	7165	9FA8
涣	27-33	渙	6E19	9C6F
晋	29-90	晉	6649	9578
静	30-18	靜	975C	EC6F
净	30-27	淨	51C8	83F4
栖	38-60	棲	68F2	97AB
弃	38-90	棄	68C4	9789
潜	39-17	潛	6F5B	9D93
挣	53-85	掙	6399	92EA
睁	53-86	睜	775C	B1A0
狰	53-88	猙	7319	AA62
争	53-89	爭	722D	A08E
伫	56-89	佇	4F47	81D0
陧	58-77	隉	9689	EA9F
奂	59-28	奐	5950	8A4A
峥	65-31	崢	5D22	8D98
戬	74-15	戩	6229	91EC
筝	83-61	箏	7B8F	B97E

In summary, GB/T 12345-90 is the traditional analog of GB 2312-80. Because of this relationship, we can say that the scope of GB/T 12345-90 is to include *all* traditional forms of hanzi in GB 2312-80. This brings us to one last error that is in GB/T 12345-90. There is one hanzi in GB/T 12345-90, 囉 (88-51), which actually should not be included because its corresponding simplified form, 啰, is not in GB 2312-80! This hanzi is in both GB 7589-87 (22-51) and GB 8565.2-88 (15-93). The reason why the hanzi 囉 was included in GB/T 12345-90 is due to an error in the 1956 draft version of 简化字总表 (*jiǎnhuàzì zǒngbiǎo*; later corrected in the 1964 version) whereby the two hanzi 羅 and 囉 were mistakenly labeled as traditional

forms of the simplified hanzi 罗 (34-62 in GB2312-80)—see Table 3-1 on page 68. Only the hanzi 羅 is the true traditional form of the simplified hanzi 罗.

In the next section, you learn that there are two more Chinese character set standards, GB 7589-87 and GB 7590-87, and that both of them, like GB 2312-80, have traditional analogs. Their traditional analogs, GB/T 13131-9X and GB/T 13132-9X, have not yet been published.

Related Chinese character set standards

There are many other character set standards developed by China, each of which is commonly referred to as a GB standard. All of these GB standards share several common characteristics:

- For every GB standard that includes simplified hanzi, there is a corresponding GB standard that replaces simplified forms by their government–sanctioned traditional forms—GB 2312-80 and GB/T 12345-90, which you read about earlier, represent one such pair of character set standards
- Every GB standard is also referred to by a numeric designation, with the most basic character set being zero (that is, "GB0" for GB 2312-80)

Table 3-29 lists the relevant GB character set standards in a way that indicates their relationship with one another, along with their assigned numeric designation. Note how simplified character sets are indicated by even-numbered designations, and traditional character sets by odd.

Table 3-29: GB Character Set Standards—Simplified and Traditional

Simplified Character Set	Hanzi	Traditional Character Set	Additional Hanzi
GB 2312-80 (GB0)	6,763	GB/T 12345-90 (GB1)	103[a]
GB 7589-87 (GB2)	7,237	GB/T 13131-9X (GB3)	
GB 7590-87 (GB4)	7,039	GB/T 13132-9X (GB5)	

[a] These 103 additional hanzi occupy all of row 88 (94 hanzi) and the first part of row 89 (9 hanzi).

An oddball character set standard in this regard is GB 8565.2-88—it is sometimes referred to as GB8.

The hanzi in GB 7589-87 and 7590-87 (this also applies, of course, to their traditional analogs, specifically GB/T 13131-9X and GB/T 13132-9X) are ordered by radical, then total number of strokes, and begin allocating characters at row 16. GB 7589-87 was established on December 1, 1987, and is entitled *Code of Chinese Ideograms Set for Information Interchange—the Second Supplementary Set* (信息交换用汉字编码字符集—第二辅助集 *xìnxī jiāohuàn yòng hànzì biānmǎ zìfújí—dì'èr fǔzhùjí*). GB 7590-87 was established on the same date, and is entitled *Code of Chinese Ideograms Set for Information Interchange—the Fourth Supplementary*

Set (信息交换用汉字编码字符集—第四辅助集 *xìnxī jiāohuàn yòng hànzì biānmǎ zìfújí—dìsì fǔzhùjí*). Tables 3-30 and 3-31 list the character allocation for GB 7589-87 and GB 7590-87, respectively.

Table 3-30: The GB 7589-87 Character Set

Row	Characters	Content
0–15	0	Unassigned
16–92	7,237	Hanzi (last is 92-93)

Table 3-31: The GB 7590-87 Character Set

Row	Characters	Content
0–15	0	Unassigned
16–90	7,039	Hanzi (last is 90-83)

It is interesting to note that all the hanzi specified in GB 7589-87 and GB 7590-87 are handwritten. Needless to say, fonts that support these character set standards are scarce.

Note that not all hanzi in the simplified character set standards are replaced by a corresponding traditional hanzi. In the case of the GB 2312-80 and GB/T 12345-90 pair, 2,180 additional hanzi are needed to transform GB 2312-80 into GB/T 12345-90. The majority are simple one-to-one replacements, but some are hanzi that swap code points or split into two or more separate hanzi (some simplified hanzi were derived from two or more traditional hanzi).

Appendixes E, *GB 2312-80 Table*, and F, *GB/T 12345-90 Table*, provide complete GB 2312-80 and GB/T 12345-90 code tables, respectively. An inadequate supply of fonts precluded the inclusion of code tables for the other GB character set standards. This volume also includes a reading index for Level 1 hanzi and a radical index for Level 2 hanzi. But note that the GB 2312-80 standard itself, as a printed manual, includes many useful indexes.

GBK—extended GB 2312-80

Another well-known GB character set is aligned to ISO 10646-1:1993 (Unicode Version 1.1), and is designated GB 13000.1-93. It is, for all practical purposes, the Chinese translation of ISO 10646-1:1993. What is interesting about GB 13000.1-93 is the Chinese-specific subset known as GBK.

GBK, known as the *Chinese Internal Code Specification* (汉字内码扩展规范 *hànzì nèimǎ kuòzhǎn guīfàn*), is simply an extension to GB 2312-80 that accommodates the remaining Chinese characters in ISO 10646-1:1993 (GB 13000.1-93). From a character-allocation point of view, GBK is composed as follows: