

## 前言

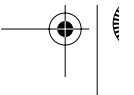
「*Understanding Japanese Information Processing*」業已出版六年了。在這期間產生了很多變化。本書延遲到現在才出版的原因之一是，筆者決定不僅要更新日文部分，而且還要收編很多有關中文和韓文的資訊，以及增加有關越文的資訊。因此，書名也隨之更改。1996 在 Togo 附近的加州大學柏克萊校園，與「*Understanding Japanese Information Processing*」編輯 Peter Mui 的一次長談中，筆者得到鼓勵決定以「中日韓越」為中心重編此書。

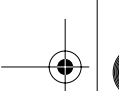
讀完這本很有份量的書，「中日韓越」（中國、日本、韓國、越南，統稱方塊字）將成為讀者知識寶庫中的一部分。然而，在進入正題之前，筆者需要使用一些讀者一定熟悉的用語。否則，可能沒有必要繼續看下去了。

越來越多的人都知道，「國際化」和「地區化」成為在電腦行業中常見用詞，而且是需要把軟體推銷到全球的高科技公司和研究員中之熱門話題。本書專門針對方塊字化，即使軟體可以適合於一個或一個以上的方塊字市場。筆者的目的在於提供方塊字化方面的資訊。

介紹國際化和地區化的書籍通常都介紹字集和編碼的資訊，但本書還提供其它方面的資訊。簡而言之，本書提供書寫體系的簡介，字集的歷史背景和現狀，以及編碼法、轉碼技術、輸入法、鍵盤排列、字體格式、版面、輸出法、配有原始碼的演算法、用於資訊處理的工具、在電子信或全球資訊網上如何處理方塊字文本等方面的詳細資訊。除了許多有關字集的跨平台資訊和討論以外，還提供方塊字文本在一些平台上的編碼法和處理法，以及開發用於方塊字市場的軟體之指南和技巧。

在此說明本書沒有涉及到的方面。本書沒有涉及到如何設計文書處理器，如何為自己的電腦設計字體（但筆者提供一些有關字體工具方面的資訊），如何正確處理方塊字地區的數字、貨幣、日期、時間等格式。本書雖然不是國際化和地區化的完整說明書，但是完全可以作為一本這方面的參考書（好在這方面的參考書慢慢多起來了）。





筆者希望本書成為方塊字資訊處理資訊的權威性來源（「*Understanding Japanese Information Processing*」重點介紹有關處理日文的問題，而且顯然成為這方面資訊的權威性來源）。因此，本書把重點放在如何以跨平台的方式，在電腦中處理方塊字文本。本書對所有的內容都進行很好的編排、分類，而且很容易找出所需要的部分。

本書的目的是消除方塊字資訊處理方面的資訊不足之問題。為此首先這幾年來，筆者一直維護著一份網路文件 *JAPAN.INF*（日文文本電子處理）。世界各地都可以透過 FTP 來得到該文件。該文件也被公認為是，在電腦中處理日文文本方面資訊的權威性來源。「*Understanding Japanese Information Processing*」從 *JAPAN.INF* 中抽選出重要的資訊，進而進行更詳細的說明。1993 年出版了「*Understanding Japanese Information Processing*」以後，*JAPAN.INF* 就做廢了。但這並不是悲劇的接尾，而是變成了一份新的網路文件 *CJK.INF*。筆者為 *CJK.INF* 所做的工作，成為本書的基礎。本書在原有的基礎上增加了很多有關中文、韓文、越文方面的資訊（當然這就需要一個新的書名）。筆者希望本書也像前書那樣得到廣泛好評。

雖然筆者為了充分提供方塊字電算方面的資訊做了不懈的努力，但是本書的很多部分仍然是以日文為主。然而，本書所介紹的知識幾乎都可以應用到所有這些語言上。使用漢字的越文電算還在發展之中，在此只做少許介紹。

## 讀者對象

本書適合對在電腦中方塊字文本的處理方法感興趣之讀者，包括那些想投身於方塊字資訊處理行業的讀者，以及那些已身在該行業但需要更多資料的讀者。本書還適合使用任何種類電腦以及任何種類作業系統（MacOS, MS-DOS, Unix, Windows）的讀者。

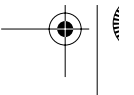
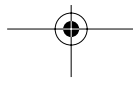
雖然本書著重於方塊字資訊處理，但是對製作多語言軟體、或對國際化和地區化的一般問題感興趣之讀者，可以學到很多在電腦上處理複雜的書寫體系方面之知識。對方塊字文本處理感興趣的讀者就更是如此。遺憾的是，方塊字文本處理方面的資訊還是比較缺乏。

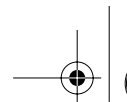
本書假定讀者只有一點、或者根本沒有中日韓越（中文、日文、韓文、越文）及其書寫體系方面的知識。第 2 章〈書寫體系〉介紹中日韓越及其書寫體系方面的知識。若只知道中日韓越中的一種語言，那麼第 2 章將起很有用。

## 本書的慣例

本書經常出現漢字、韓字、假名、平假名、片假名等用語。還出現 ANSI, ASCII, CNS, EUC, GB, ISO, JIS, KS, TCVN 等縮寫。在正文以及字彙集（參考附錄 X〈詞彙集〉）中，對這些用語和縮寫都做了解釋。

本書中的十六進位數值都以 0x 開頭；例如，0x8080。跟在 0x 後面的每兩個十六進位數值都代表一個位元組。例如，0x20 代表一位元組數值，而 0x0020 則代表二位





元組數值。十進位數值則不加任何修飾。還可以利用附錄 B 〈表記轉換表〉來執行表記法之間的轉換。

本書常常使用後置詞 J, K, S, T, V, CJKV 來代表軟體產品的地區版本或方塊字版本。因為軟體開發商對方塊字版本產品的表示方法都不統一，所以筆者決定使用這些後置詞來保持用語的統一。而實際的產品名稱，通常透過後置「日本語版」、或前置「Kanji」、或前置「日本語」來代表日文軟體產品。而中文軟體產品名稱則通常前置「中文」。本書一般不標出軟體的版本數字（這種資訊很快就過時）。本書只標出代表產品發展或開發的重要階段之版本數字。

在本書中，「中國」代表「中華人民共和國」，又稱為大陸。「臺灣」代表「中華民國」。很多情況下，這兩種名詞需要區分開來。

在本書中，人名的拉丁字母音譯，按照歐美的慣例，名在先，姓在後。方塊字人名則是，姓在先，名在後。

在本書中，ISO 10646-1:1993 與 Unicode 沒有區別。只在少數的情況下，意思才有所不同。

「斜體字體」用於代表路徑名稱、檔案名稱、程式名稱、新名詞、新聞群組名稱、Internet 位址（如領域名稱、URL、電子信位址）。

「等寬字體」用於代表指令輸出的說明部分、檔案內容、電子信信息。

「等寬粗體」用於說明指令等使用者鍵入的文本。有時也用於把例子分段。

「等寬斜體字體」用於在程式碼以及例子中，需要根據具體情況而改變的變數。例如，作為變數的電子信位址需要換成實際的電子信位址。

「%」字符用於代表 Unix 指令行中的 Unix shell 提示符號。

腳註用於提供附加的資訊。為了簡明起見，有時使用了一些不精確的說法（特別是介紹書寫體系的第 2 章）；在這種情況下，腳註通常（但不都是）用於說明事情的真相。

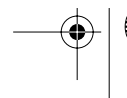
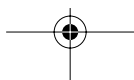
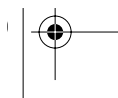
## 本書的結構

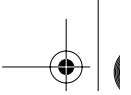
在此簡單介紹各章的內容。不一定要按部就班一章一章地看下去，可在章節或附錄之間跳來跳去著看。本書還提供了索引。

第 1 章〈中日韓越資訊處理概觀〉簡單介紹本書所涉及到的問題，給讀者初步印象。讀者可以從中瞭解到本書對自己的工作或研究的哪方面有用。

第 2 章〈書寫體系〉介紹與方塊字書寫體系有直接關係的資訊。介紹組成方塊字文本的各種字符。這一章主要是為那些不熟悉中文、日文、韓文、越文（或只熟悉其中一種或二種語言）的讀者而預備的。每個人都能從這一章學到一些新的知識。

第 3 章〈字集規格〉介紹電子和非電子這兩種方塊字字集。電子字集可以再分為國家規格和國際規格的兩種。還對方塊字字集之間進行比較。





第 4 章〈編碼法〉介紹在電腦上如何把第 3 章所介紹的字集加以編碼。編碼雖然複雜，但卻是用電腦來表達和處理自然語言的文本時的重要步驟。還介紹方塊字轉碼軟體，以及如何修復受損的方塊字文本檔案。

第 5 章〈輸入法〉介紹輸入方塊字文本的資訊。首先介紹方塊字輸入的一般事項，然後介紹在電腦上輸入方塊字字符的幾種方法。還介紹輸入方塊字時，所需要的硬體，也就是鍵盤。有各式各樣的鍵盤排列，從英文鍵盤排列（QWERTY）、到有成千上萬的按鍵之漢字字版。

第 6 章〈字體格式〉從方塊字的角度的角度，介紹點陣字體格式以及外框字體格式的資訊。這一章其實與筆者在 Adobe Systems 公司的工作內容密切相關，因此有些部分可能過於詳細。

第 7 章〈版面〉介紹如何在頁面上對方塊字文本進行適當的排版。光有方塊字字體並不夠；還有一些規則，包括，字符可以用於哪裡、不可以用於哪裡，哪些字符放在一起時需要特殊的處理。這一章的末尾介紹具有先進排版功能的軟體。

第 8 章〈輸出法〉介紹如何使用顯示、列印等方法輸出方塊字文本。還介紹列印和顯示最新技術的資訊。

第 9 章〈資訊處理技術〉介紹方塊字轉碼技術以及方塊字文本處理技術的資訊和演算法，並做出詳細說明，適時附上使用 C、Java 等程式設計語言編寫的演算法。這一章的末尾簡要說明，筆者這幾年來一直編寫和維護的三種日文字碼處理工具。這些工具展示了如何在處理日文時應用這些演算法。

第 10 章〈作業系統、文本編輯器、文書處理器〉介紹可以處理方塊字，即可以支援一個或一個以上方塊字地區的作業系統、文本編輯器、文書處理器。

第 11 章〈字典與字典軟體〉介紹在處理方塊字時，起重要作用的普通字典和電子字典。還介紹一些如何有效地利用各種字典索引來查找漢字的技巧。

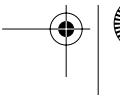
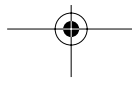
第 12 章〈Internet〉介紹在電子信系統和新聞閱讀器等軟體上，如何妥善處理通過網路的方塊字文本。還介紹有關如何確保所發出的信息，可以完整無缺地傳送到對方的技巧，以及方塊字地區 Internet 網域的資訊。

第 13 章〈全球資訊網〉介紹有關使用各種網頁瀏覽器來顯示方塊字文本方面的資訊，提供用於編寫包含方塊字文本的 HTML（超文本標記語言）以及 XML（可延伸標記語言）文件之指令。還詳細介紹 Adobe Acrobat、PDF（可攜式文件格式）、CGI（通用閘道介面）程式設計。

附錄 A〈轉碼表〉提供在十進位區位、十六進位 ISO-2022、十六進位 EUC、十六進位 Shift-JIS（日文）之間的轉碼表；還包括 Shift-JIS 的使用者定義區域。

附錄 B〈表記轉換表〉使用二進位、八進位、十進位、十六進位的四種常用表記法列出所有 256 個八位元位元組數值。

附錄 C〈廠商字集規格〉是有關廠商方塊字字集的參考資料，提供給對其感興趣的讀者。



附錄 D 〈廠商編碼法〉是有關附錄 C 廠商方塊字字集編碼的參考資料，提供給對其感興趣的讀者。

附錄 E 〈GB 2312-80 字碼表〉按十進位區位碼的順序，列出 GB 2312-80 所定義的字符（包含了 GB 6345.1-86 所指出的更正和增補）。

附錄 F 〈GB/T 12345-90 字碼表〉按十進位區位碼的順序，列出 GB/T 12345-90 所定義的字符。

附錄 G 〈CNS 11643-1992 字碼表〉按十進位區位碼的順序，列出 CNS 11643-1992 所有七個字面的字符。還包括第 15 字面等、屬於 CNS 11643-1986 的字碼表。這是一個很長的附錄，列出超過五萬個漢字！

附錄 H 〈Big5 字碼表〉按十六進位 Big5 碼的順序，列出 Big5 所定義的字符。

附錄 I 〈香港 GCCS 字碼表〉按十六進位 Big5 碼的順序，列出香港政府頒布的 3,049 個漢字；還包括香港司法部門所定義的 145 個漢字。

附錄 J 〈JIS X 0208:1997 字碼表〉按十進位區位碼的順序，列出 JIS X 0208:1997 所定義的字符。

附錄 K 〈JIS X 0212-1990 字碼表〉按十進位區位碼的順序，列出 JIS X 0212-1990 所定義的字符；還包括將來可能加入該規格的四個片假名（但看來不太可能）。

附錄 L 〈KS X 1001:1992 字碼表〉按十進位區位碼的順序，列出 KS X 1001:1992 所定義的字符。

附錄 M 〈KS X 1002:1991 漢字表〉按十進位區位碼的順序，列出 KS X 1002:1991 所定義的漢字部分。

附錄 N 〈韓字語音表〉列出 KS X 1002:1992 字集所定義的所有 2,350 個現代韓字之語音索引。

附錄 O 〈TCVN 6056:1995 字碼表〉按十進位區位碼的順序，列出 TCVN 6056:1995 所定義的字符。

附錄 P 〈字碼表索引〉提供漢字的語音索引、部首索引、筆畫數索引；可以和本書的其它附錄並用。

附錄 Q 〈字符表和對應表〉列出本書所使用的字符表和對應表。

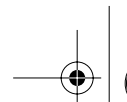
附錄 R 〈漢字表〉列出（第 3 章〈字集規格〉所介紹的）非電子字集的漢字。

附錄 S 〈一位元組字碼表〉按十六進位字碼的順序，列出 ASCII, EBCDIC, EBCDIK, ISO 8859-1:1998, CJKV-Roman, 半寬片假名, 半寬韓字母。

附錄 T 〈軟體和文件來源〉提供本書所涉及的軟體和文件之地址以及聯繫資訊。

附錄 U 〈通信論壇〉提供讀者可能感興趣的通信論壇方面之資訊。

附錄 V 〈專業組織〉提供與方塊字資訊處理有關的組織之資訊。



附錄 W 〈Perl 程式範例〉提供第 9 章所介紹的演算法之 Perl 程式，以及一些其它有用的資訊。

附錄 X 〈詞彙集〉定義本書（以及其它的書）所涉及的概念和用語。

〈參考書目〉列出很多參考書籍。筆者在寫本書時使用了其中的一些書籍。

## 謝辭

寫這麼厚的一本書，需要與世界各地的很多人進行交往。在此無法把這些年來幫過筆者數以百計的人都一一列出。

有時有些人求筆者在特定的問題上幫忙（我想要是人家知道你的電子信位址，經常就是這樣。筆者為了能收到大量的電子信，在本書的很多地方都寫上筆者的電子信位址 *lunde@oreilly.com*）。有時筆者可能不知道特定問題的答案，但這些問題通常促使筆者去尋找答案；答案總是存在的。

到 1998 年筆者在 Adobe Systems 公司已經工作了七年。這是美好的七年。在那裡，筆者每天都可以面對有關中日韓越的問題。Adobe Systems 公司的先進字體技術及其對客戶的承諾，都給筆者留下很深的第一印象，也是吸引筆者在那裡工作的原因。公司還允許筆者在辦公室裡，擺設三呎高鮮橘色的 Godzilla 模型，來紀念和致敬這個大恐龍（它在 1995 年的電影 *Godzilla versus Destroyah*\* 中悲壯死去）。講到致敬，無論是在製作本書的哪個方面，都應該對 Adobe Systems 的出版技術表示致意。

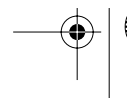
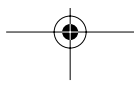
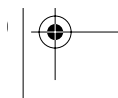
在此感謝所有讀過筆者前作的人、在工作上寬容筆者的遲鈍和不討人喜歡的個性的人、指出筆者的錯誤的人、與筆者交換電子信的人、或幫助筆者變得更好的人。不用列出姓名，這些人都應該不言而喻。

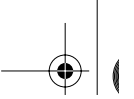
在此應該特別感謝 Tim O'Reilly（O'Reilly & Associates 的總裁和創立者）和 Peter Mui 對筆者第一本書「*Understanding Japanese Information Processing*」的信任。正是 Peter 激勵筆者把這一本書擴展到針對所有中日韓越地區（真抱歉，花了這麼長時間久才完工；這真是一個很辛苦的工作）。感謝 Edie Freedman 贊成筆者在封面上使用河豚的想法。† Mike Sierra 在本書的佈局上幫了很多忙，也促使筆者學了很多 Adobe FrameMaker 的獨特功能。Chris Reilley 把本書很多不高明的插圖都修改成藝術品。而且這是第二次了！編輯 Gigi Estabrook 不停地催著筆者把本書寫完。文章編輯 Ellie Fountain Maden 發現了本書不少錯誤和問題。

在漫長著書過程的各個階段中，以下各位對本書進行了審閱：Joe Becker, Jim Breen, Robert Bringhurst, Woohyong Choi (최우형), James Davis, L. Peter Deutsch, James Đỗ (杜伯福), Terry Dowling, Martin Dürst, Jeff Engelman, Gus Fernandez, Jeffrey

\* 日文為ゴジラ対デストロイア。Godzilla的原創者田中幸在1997年四月過世，享年86。美國拍攝的第一部Godzilla電影是在1998年放映的。如同本書，關鍵在於份量和尺寸。

† Michael Slinn 曾指出使用巴別爾魚 (Babel Fish) 作本書的封面很適合；據 Douglas Adams 的 *The Hitch Hiker's Guide to the Galaxy* 所說，只要在耳朵裡面塞一條巴別爾魚，與大腦連接，頓時，就能明白所有的語言。本書還是使用河豚作封面的原因或許是在 Dover Pictorial Archive 中沒有 19 世紀巴別爾魚的雕刻 ...





Friedl, David Gourley, Jerry Hall, Jack Halpern (春遍雀來), Ken'ichi Handa (半田劍一), 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 (鄭楮璋, 小松章), 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), Haifeng Zhu (朱海峰)。非常感謝他們所提供的各種資訊和建議。然而，筆者對本書的錯誤、疏忽等負所有的責任。

最後，還要感謝父母 Vernon Delano Lunde 和 Jeanne Mae Lunde 多年來對筆者支援。也感謝兒子 Edward Dharmputra Lunde、繼子 Ryuho Kudo、美麗的女兒 Ruby Mae Lunde、深愛的太太和伴侶 Hitomi Kudo。

## 錯誤、疏忽、更正

一本介紹高度技術資訊的書，難免出現錯誤。可以肯定，這些錯誤將在本書的新印刷或新版時更正。在過度期間，在以下的 URL 上登載更正：

<ftp://ftp.oreilly.com/pub/examples/nutsbell/cjkv/errata/> (英文網址)

<http://www.oreilly.com.tw/chinese/other/cjkv.html> (中文網址)

若讀者發現錯誤或疏忽之出，請寫信寄到以下地址：

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)

也可發電子信。想要加入 O'Reilly 通訊論壇，或需要出版目錄，請發電子信到：

[info@oreilly.com](mailto:info@oreilly.com)

有關本書的技術問題或評論，請發電子信到：

[bookquestions@oreilly.com](mailto:bookquestions@oreilly.com) (請以英文書寫)

[bookquestions@oreilly.com.tw](mailto:bookquestions@oreilly.com.tw) (請以中文書寫)

因為本書提供了數以百計的 URL，所以筆者還提供以下的網頁，先按章或附錄後按頁數排列，以便可以迅速取得（這也是保持最新版的一種方法）：

<http://www.praxagora.com/lunde/cjkv-urls.html>

