人文科学とコンピュータ　52－3
（２００１．１０．１２）

Unicode 漢字文書の処理について

二階堂　善弘（茨城大学人文学部）

中国の古典漢字文献を電子化する場合、常に文字コードの問題があった。このたび Unicode 3.1 の CJK Unified Ideographs Extension A 及び B が、Windows で使用できることになり、多くのアプリケーション上で、約 7 万字の漢字が使用可能となった。これは大変有用なツールである。ただ、幾つかの問題が解決される一方で、新たな問題も発生すると考えられる。

On Text Processing for Unicode 3.1

CJK Unified Ideographs

NIKAIDO Yoshihiro (Ibaraki University)

About text processing of Chinese classics, we can use Unicode 3.1 CJK unified ideographs extension A and extension B on Microsoft Windows. It has nearly 70 thousand characters, very useful for construction of Chinese classics e-text. But it also has many problems.

１．はじめに

コンピュータ上で漢字文書を扱う際には、必ずと言ってよいほど文字コードが問題となった。特に古典文献を電子化する場合、無い漢字をどう表記するかは重要な問題であり、現在でも十全といえる解決策は無いと言ってよい。

しかしながら、パーソナルコンピュータとその OS の発展により、数多くのツールが開発され、古典漢字文献を表記することも、かなりの範囲まで可能となりつつある。ここでは、現時点（2001 年 9 月）におけるこれまでの多漢字ツールの発展を整理しつつ、新しい Unicode 3.1（及び ISO/IEC 10646-2）を使用した漢字文書の利用について考えてみたい(1)。

２．古典漢字文献の利用からみたツール

多漢字を扱うためのツールが充実し始めたのは、ここ数年のことである。

それまでは、JIS X 0208 に収録されている以上の漢字を扱うには、外字を使って処理することが多かった。また、中国大陸系の GB 2312 や台湾系の Big5 を扱うためのツールも少なく、さらに多言語を混在する手段もほとんど無かった。

Unix 系の OS においては、JIS X 0212 がかなり早い段階から使用可能であったが、一般化するには至らなかった。また当時はまだ Linux や FreeBSD は一般的とは言えず、Unix を導入するのは特に文系研究者にとっては敷居が高かった。特に、Shift-JIS が大半のパーソナルコンピュータの OS で使われていることが大きなネックであり、『論語』や『孟子』のような、字数の少ない古典を扱うことすらかなりの困難があった。これを踏まえ、古典電子テキストについては、Shift-JIS による抜けた字ばかりの MS-DOS テキストファイルで作成されることが多く、不便な状況であった。

流れが変わったのは、マイクロソフト社[(2)]の Windows NT4.0（1996 年）や Windows 98（1998 年）などの発売後である。これにより Windows の多くのアプリケーション上で、JIS X 0208 を上回る数の漢字が使用できるようになった。特に Windows NT 系のツールにおいては、ISO/IEC 10646-1 つまり Unicode（UCS-2）の BMP がテキストで扱えるため、漢字数では 20902 字が平易に使用可能となった。

実装面ではマイクロソフト社より SimSun や SimHei などのフォントが提供され、Word や Excel などのアプリケーションの上では、20902 字すべてが問題無く使用できるようになった。またジャストシステム社[(3)]の一太郎や ATOK など、Unicode に対応しているアプリケーション上では容易に約 2 万字の漢字が使用できるようになった。さらに UTF-8 を使うことにより、HTML を記述して Internet Explorer や Netscape Communicator などの使用を前提にすれば、データを公開することも可能となった。

もっとも、約 2 万字と称するものの、UCS-2 に収録される漢字は、JIS X 0208 や JIS X 0212、また GB2312 や CNS-11643 の第 1 面と第 2 面（Big5 にほぼ相当）をソースとしており、中途半端な統合を行っているため、古典作品などを電子化する場合、実質的な使用可能な漢字数というのは、ほぼ Big5 の漢字数、つまり約 1 万 3 千字程度である。また実際、JIS X 0212 と Big5 とは期せずして一致する部分が多い。

しかし、実際にはこの程度の字数でも、かなりの数の主な中国古典文献の電子化がカバーできるようになっている。ただこのことは意外に知られておらず、古典の電子化というと、いたずらに漢字数を増やせばよいという議論に終始している面があることは、甚だ遺憾である。

３．古典漢字文献の利用からみた多漢字ツール

多漢字という点では、TRON 技術を応用したパーソナルメディア社[4]の超漢字があり、また、エー・アイ・ネット社[5]の今昔文字鏡がある。いずれも大規模な数の漢字を収録する。

超漢字の場合、その最新版である超漢字 3 は、約 17 万字の文字が扱えると称する。特に GT 書体[6]の約 6 万 5 千字の漢字を搭載していることが特色である。超漢字における文字種の数え方は、概念が異なっているために単純に比較はできないが、その収録漢字の多くは大修館書店の『大漢和辞典』に拠っていることから、実質上使用できる漢字数は約 6 万 5 千字程度とみなすことができる。超漢字は、膨大なコード領域を持っており、データの交換が可能という点では優れている。しかし、残念ながらアプリケーションソフトが少なく、他 OS とのデータ変換は困難である。また Web 上の既存データを利用することも不得手であり、例えば Big5 や UTF-8 のデータを利用できない。もちろん、ある特殊な用途、或いは超漢字だけの閉じられた世界であれば、約 6 万 5 千の漢字を使うことに問題はない。

今昔文字鏡は、最新版は単漢字 10 万字版というバージョンが発売され、10 万の漢字を扱えるとする。この 10 万という字数も、むろん多くの異体字を含むものであり、現実に使用できる字数はやや少なくなる。また今昔文字鏡は、ワープロなどのアプリケーション上によって、本来の文字コードの位置を別の字のフォントに置き換えることによって、多漢字環境を実現している。だから、まったくのプレーンテキストでデータを渡した場合は、データの交換が保証されない。また、Web による公開においても、文字の表示が難しい面がある。また例えばブラウザから直接ワープロへ貼り付けるなどの作業も困難である。

今昔文字鏡については、あくまで巨大な外字集として考えるべきであろう。メインの電子テキストについては、JIS なり Unicode なり、他の文字コードを使用し、表示できない文字を外字として利用するという方法が有用であると考える。

Linux も急速に Unicode への対応を進めていることから、多くのアプリケーションで利用が可能になっている。またこれは Linux に限らないが、KDE や Gnome などの Unix 系デスクトップ環境の多言語化の動きが盛んであり、著しい発展を遂げている。ただ、個々のアプリケーションレベルでは、対応のばらつきがあり、設定も面倒なものが多い。Linux ではむしろ中国大陸独自のディストリビューション[7]の方が、漢字処理においては強い。

４．大規模漢字データベースと Unicode

昨今の中国大陸や台湾における中国古典データベースの発展は、驚異的と言ってよい。それは台湾の中央研究院が「漢籍電子文献」[8]を公開し、『史記』『漢書』

から『十三経注疏』に至るまでの膨大なデータを検索可能にしたことに始まった。いまや台湾故宮博物院の「寒泉」(9)や、中国北京の「国学」(10)など、古典を中心としてデータを公開するサイトが続々と増えており、膨大な数の中国古典が使用できる。もっとも、ネット上には現代文学の作品も増えており、こちらも数多くの作品が利用可能である。

これに加え、8億字のテキスト及び画像検索が可能な『四庫全書』がCD-ROMとして既に発売されている。また優れた版本を収録することで知られる『四部叢刊』、また、古代から近代までの漢籍を電子化し、20億字ものデータ量を有する『中国基本古籍庫』などが続き、いまや中国の古典文献については、膨大な資料を活用しての研究が求められている。

これらのデータを活用する場合、やはりUnicode（UCS-2）を使った方が、交換性という点では有利である。例えば、中央研究院のデータを検索するに際し、その表示についてはBig5で行うものの、これをUCS-2を介した形でWordなり一太郎なりにコピー&ペーストすれば、それはUCS-2のUnicodeテキストとして利用しても、JISのテキストとして利用しても、はたまた単なるWord文書として使っても構わない。特にWord文書の場合、中国語など、他の国・地域のWindowsなどでも読み書きが可能である。

また『四庫全書』のデータなどの場合、始めからUCS-2を使っての電子化がなされており、利用する上では当然UCS-2を使った方がよい。

このように、現在既に構築されている膨大な電子化文献を利用する上では、Unicodeを使わざるを得ないのが現状である。むろん、これはフォントなどが用意されていれば、Windowsのみならず、Mac OSやLinuxでも利用可能となっている。しかし、超漢字や今昔文字鏡を使用しては、部分的な交換しかできない。このあたりの交換性が、Unicodeの優位性として挙げられる。

但し、Unicodeのデータ交換性も、対ローカルコードということであれば、様々な問題がある。このことについても考慮する必要はある。

５．GB18030とUnicode 3.1

最近になって漢字の古典文献を扱う上で、また重要な動きが現れた。それは中国大陸のGB18030の制定、及びUnicode 3.1によるUnicodeの拡張である。

中国大陸のGBの漢字コード(11)は、簡体字中心のGB2312-1980から拡張を続けてきた。GBKと呼ばれるGB13000-1993、そして2000年3月に発布されたGB18030-2000(12)である。

GB13000は、実際には漢字部分に関しては、ほとんどUCS-2をそのまま取り込んでいる。Web上では既に広く使われており、ブラウザにおいてGB2312と表記されていても、実際にはGB13000で多くの漢字が表示されている場合が多い。

GB13000 は GB2312 の上位互換コードであるため、このようなことが可能となっている。

GB18030 については、中国はソフトウェアをこのコードに適合させるよう要求していると言われる。この文字コードは、拡張された Unicode のすべてのコードを収録してさらに拡張したもので、1 バイト・2 バイト・4 バイトの可変長のコードを使用する。すべてで 160 万の膨大な領域を持つというものである。

また、Unicode 3.1 については、Unicode 3.0 が拡張されており、これまでの UCS-2 で使用していた BMP 第 0 面に加え、第 1 面・第 2 面及び第 14 面が使用できる。これがまた ISO/IEC 10646-2 となっている。

BMP の第 0 面には、CJK Unified Ideographs Extension A として、漢字が 6582 字追加され、また第 2 面には、同 Extension B として 42711 字が追加されている〔13〕。これにより、Unicode 3.1 において使用可能な漢字数は、約 7 万字にのぼっている。

これらの漢字のソースとなっているのは、主に CNS-11643 に含まれる漢字であり、また JIS X 0213 の漢字部分である。これらは『康熙字典』や『大漢和辞典』や『漢語大字典』などに収録される漢字の多くをカバーする。古典漢字文献を表示するための漢字数としては、完全とは言えないまでも、かなり十分なものであると言える。

GB18030 も、現在はその膨大な領域のほとんどは定義されていないが、とりあえず UCS-2 に Extension A に含まれる漢字を加え、27484 字が使用できるようになっている。GB18030 への対応については、幾つかの Linux ディストリビューションが既に行っているというが、その詳細については不明な部分が多い。

６．Office XP の Unicode 対応

このような Unicode の拡張については、既に部分的な実装が行われており、それが今後の古典漢字文献の電子化に影響を与えることとなっている。

マイクロソフト社の Office XP では、Windows 2000/XP 上で使用した場合、サロゲートペアによる Unicode 3.1 のエリアが使用できるようになっている。また北大方正〔14〕の提供による Extension A 及び B のフォントを使用すれば、Word 2002 や Excel 2002 などのアプリケーションで、約 7 万字の漢字が使用できる。このフォントは、中国の中文 Office XP にバンドルされている他、英語版の Office Proofing Tools 2002 にも搭載されている。このフォントは、Simsun（Founder Extended）というものである。ただこの新 Simsun フォントは、約 40MB もの容量を持つ。

これらの漢字は、Windows 2000 のメモ帳でも使用でき、テキストファイルとして保存も可能である。Unicode の交換における利便性を考えれば、今後は古典

漢字文献に新 Unicode テキストを使用することも考えられよう。

但し、テキスト処理において、これまでと異なる注意も必要となろう。そもそも、UCS-2 においては、漢字データについて 2 バイト単位で一律に処理することが可能であった。しかし、Extension A はともかく、Extension B については、サロゲートペアによる処理を考える必要が出てきた。

また、これまでのデータとの整合性を考える上で、膨大な異体字をどう処理するかが大きな問題となってきた。多くの漢字を増やした結果、ある漢字間でのデータの交換が単純にいかなくなってきたのである。

また現在では、どうやって漢字を入力するかも問題となる。約 7 万字もの膨大な漢字について、部首や発音によって検索できるツールや IME は、まだ有力なものが無い。

ただブラウザについては、Internet Explorer 6 が対応しており、データの公開は可能であると思われる。しかしこの場合は、クライアントのそれぞれに新 SimSun（SurSun）或いはそれに類するフォントがインストールされていることが前提となる。

７.『三国志平話』の場合

それでは実際に、中国の古典漢字文献をデータ化する場合に、この拡張されたエリアが有用であるかどうかみてみたい。

約 6 万字のデータ量を持つ『三国志平話』は、筆者が公開する漢字テキストデータである〔15〕。このうち、大半の漢字は、UCS-2（BMP）の範囲内であった。そこで表示できない文字は、僅かに 5 字であった〔16〕。このうち、拡張された Unicode の Extension A 及び Extension B に含まれていた漢字は、4 字であった。

そもそも『三国志平話』で使用されている漢字がかなり僻字であることを考えると、拡張された漢字の有用性は高いと言えるのではないかと考えられる。なお、含まれなかった漢字 1 字については、『漢語大字典』にも未収録の漢字であり、他の漢字ツールにも通常含まれていない。

８．おわりに

一般に広く使われている Word や Internet Explorer を使って、約 7 万の漢字が使用可能となったことは、漢字文献の処理に大きな影響を与えると考えられる。今後は、もっと多くの古典漢字文献によってその有効性を検証する必要があろう。また電子化された漢字テキスト処理において、これまでとは異なったアプローチが必要になる可能性が出てきた。これらについては今後の課題としたい。

注

1. Unicode については、Unicode Consortium（http://www.unicode.org/）のサイトを参照。
2. マイクロソフト（http://www.microsoft.com/ms.htm）
3. ジャストシステム（http://www.justsystem.co.jp/）
4. パーソナルメディア（http://www.personal-media.co.jp/welcome.html）
5. 文字鏡ネット（http://www.mojikyo.org/）
6. 東京大学多国語処理研究会（http://www.l.u-tokyo.ac.jp/GT/）
7. 中国のディストリビューションについては、紅旗 Linux など、多くの種類が存在する。幾つかのものについては、筆者のサイトにおいて紹介している（http://nika01.hum.ibaraki.ac.jp/~nikaido/）。
8. http://www.sinica.edu.tw/ftms-bin/ftmsw3
9. http://210.69.170.100/s25/index.htm
10. http://www.guoxue.com/
11. GB には漢字コード以外にも多くの規格がある。
12. GB18030 については、漢字文献情報処理研究会の解説ページを参照のこと（http://jaet.gr.jp/gb18030/index.html）。
13. 追加された Extension A の領域は、U3400-4DB5 であり、Extension B の領域は、U20000-U2A6D6 となっている。
14. http://www.founder.com.cn/fontweb/main1.htm
15. http://nika01.hum.ibaraki.ac.jp/~nikaido/heiwa.html
16. これについては、拙論「全相平話二種データベース構築の問題点」（『全相平話二種データベースの構築』平成 11・12 年度日本学術振興会科学研究費補助金奨励研究 A 報告書 2001.3, pp.3-6）を参照。

＜参考文献＞

＊注記はしなかったが、以下の資料については随所で参考にさせていただいた。

・安岡孝一・安岡素子著『文字コードの世界』（東京電機大学出版局・1999 年）
・小林龍生・安岡孝一・戸村哲・三上喜貴編『bit 別冊インターネット時代の文字コード』（共立出版・2001 年）
・トニー・グラハム著、乾和志・海老塚徹訳、関口正裕監修『Unicode 標準入門』（翔泳社・2001 年）
・川幡太一「新 ISO/IEC 10646 と Unicode の漢字を検証する」（『漢字文献情報処理』第 2 号・漢字文献情報処理研究会・2001 年）