コンピュータに選ばれる漢字

石　井　久　雄

■ 規格

年次　現在の規格番号［規格化当初の番号］:名称

1963年　ANSI X 3.4-1986 (R1997): American national Standard Code for Information Interchange
文字を処理する今日の方法の基底となっている。

1973年　ISO/IEC 646 [ISO 646]: 7-bit coded character set for information interchange
ISO/IEC 2022 [ISO 2022]: character code structure and extension techniques
ASCIIを一部入れ換えてヨーロッパ諸言語に拡張し、また、2言語を併存させる。

1975年　JIS X 0202 [JIS C 6228]：文字符号の構造及び拡張法

1976年　JIS X 0201 [JIS C 6220]：7ビット及び8ビットの情報交換用符号化文字集合
カタカナを扱う。SI/SOによる7ビット方式と、拡張した8ビット方式とを規定する。

1978年　JIS X 0208 [JIS C 6226]：7ビット及び8ビットの2バイト情報交換用符号化漢字集合
漢字6355（当初6349）字を含む6869（当初6802）字を規定する。

1980年　中国　GB 2312：信息交換用漢字編碼字母集 基本集
漢字6763字を含む7445字を規定する。後に「輔助集」も規定されている。

1984年　台湾　Big5
漢字13053字を含む13523字を規定する。政府のものではない。

1986年　台湾　CNS 11643：中文標準交換碼
7（当初2）面で47711（当初13735）字を規定する。Big5をすべて含む。

1987年　ISO/IEC 8859 [ISO 8859]: 8-bit single-byte coded graphic character sets
ASCIIを前半に置き、後半にヨーロッパ諸言語の一つを置く。諸言語版がある。

1990年　JIS X 0212：情報交換用漢字符号——補助漢字
漢字5801字を含む6067字を規定する。

1992年　韓国　KS X 1001 [KS C 5601]：情報交換用符号（ハングルおよび漢字）
ハングル2350字・漢字4888字を含む8224字を規定する。

1993年　ISO/IEC 10646 : universal multiple-octed coded character sets
Unicode 2.1との間で内容を統一させる。

1997年　JIS X 0208　［→1978年］：7ビット及び8ビットの2バイト情報交換用符号化漢字集合
シフトJIS、EUC、ISO-2022-JPの符号化方法を、参考としてながら規定する。

2000年　JIS X 0213：7ビット及び8ビットの2バイト情報交換用符号化拡張漢字集合
第三水準漢字・第四水準漢字を規定する。

漢字のみならず文字一般をめぐる環境の変化として、現在のインターネットの展開は、印刷の普及に匹敵するものです。印刷はグーテンベルクの銀河系と呼ばれる世界を拓き、インターネットはそれに終焉を告げさせると言われます。印刷はだれもを文字の読み手とし、インターネットはだれもを文字の書き手とすると言うこともできるでしょう。　→参考　マクルーハン（1962)、ボルツ（1993)、バーカーツ（1994)、レビンソン（1999)

インターネットは情報交換の手段であり、情報は当初は文字で記されるのみでした。今は図像・音声によることもできますが、主としてはやはり文字によっていると考えられます。インターネットにおける文字の問題を、コンピュータにとっての漢字に集約させて述べることとします。インターネットの技術的問題はおおむねコンピュータの問題であり、文字の問題は漢字から提起されたところが大きいように見えます。

→以下全般の参考　伊藤（1996)、加藤（2000)、清水（2001)

人がコンピュータに何を要求してきたか、コンピュータが何を実現してきたか、ということを追うと、要求の拡大と技術の展開とは、相互に背景ないし環境となっていると了解されます。その日本の流れには、国際的な流れが環境となっていることも了解されます。そのことは、以下に改めて指摘することを省きますが、このシンポジウム分科会のテーマにかかわります。

漢字の学習・教育については、最後に一言するに止めます。しかし、学習・教育の言わば後に、漢字にどのように向かうことになるか、その背景を以下の全体を通して述べることになります。

■　コンピュータで漢字を扱うまで

コンピュータは、本質的には、ある点を電気が通っているかいないかを検知する機械であり、それを2進数と見なして数値計算に用いるのが、コンピュータの基本的な使いかたです。いろいろなことがらは、数値に変換することができれば、コンピュータで処理することができます。文字の並びに番号を付けるという着想があって、コンピュータで文字を処理することが可能になりました。

コンピュータで最初に規格となった文字は、ラテン＝アルファベットです。それも、アクセント記号などをもたないアメリカのものでした。タイプライタのキーを見合わせて、大文字・小文字各26個、数字10個、句読記号・括弧など32個、合計94個がまず扱うべき対象である。加えて、文字ではないが、タイプライタが具えるスペース・タブレタ・キャリッジリターンその他の機能を盛り込む。など、すべて2^7＝128個に番号を付けた、その7ビットのうちが文字の世界でした。　→　ASCII

ヨーロッパ諸国は、その7ビットの体系に多少の変更を加えたり、8ビットの体系に拡張したりして、文字を扱ってきました。8ビットの体系というのは、7ビットの体系に、同様の7ビットの体系を加えたものです。

→　ISO/IEC 646, ISO/IEC 2022, ISO/IEC 8859

日本のいわゆる半角カタカナも、8ビットの体系の拡張部分で扱われます。ラテン＝アルファベットを7ビットの第1の体系に据え、カタカナを第2の体系として添えたものが、規格となりました。ひらがなは形が曲線的であり、それに対して、カタカナは直線的であって、コンピュータで扱いやすいものです。電報をカタカナのみで記してきたといった実績もありました。　→　JIS X0201

日本では、しかし、ひらがなと漢字とでものを書き記すのが普通です。それをコンピュータで扱いたいという要求が当然に生まれ、16ビットの体系が開発されました。理屈からすると2^{16}＝65536個の文字を盛り込むことができますが、それまでの体系に配慮して94^2＝8836個を最大限とすることになります。現実には、漢字約6350個、ほかにひらがな・カタカナも含み、ラテン＝アルファベットの再規定もあるなどして、合計6800個余りの文字・記号を規定しています。　→　JIS X 0208

コンピュータで処理するために漢字に規格を設けることは、日本がまず行い、直ちに中国・台湾・韓国が続きました。同じく漢字と言いはしても、それぞれで、書き記す言語も、他に用いる文字も違い、漢字についての考え方も違います。起源を同じくする一字一字を取り上げても、形も読みも意味も違います。したがって、漢字のセットの内容は異ならざるをえません。しかし、コンピュータのための規定の方式は、おおむね同様です。

→参考　安岡・安岡（1999)

以上が、コンピュータで文字を扱うことを着想してから漢字を取り込むまでの、非常に大まかな流れです。国などによる公的な規格のみを追いました。その前に、コンピュータメーカなどの私的な方式があり、例えば、IBM

は独自に体系EBCDICを作り上げ、台湾では私的な独自の方式BIG 5が普及している、といったことがあります。これには立ち入りません。

■ コンピュータでの漢字の規定のしかた

コンピュータで文字を扱うには、文字を例えばアルファベット順に並べて、それに番号を振ります。文字とはその番号であることになります。今日最も基本的なものは、上に触れたラテン＝アルファベット（以下「Lab」と記します）のものであって、その一字一字が、0から127までのうちの数値のどこかに対応しています。

漢字は、数が多く、Labの100倍を優に越えるでしょう。番号を振りさえすれば、Labと同様に多彩に利用することができます。番号を振ることは簡単であるように見えますが、すでに存在して使用されているLabとの関係があり、面倒です。

どのような番号を与えることができるかが、取り扱うことができる漢字の量を決定します。その検討の結果が、上に述べた、最大量を$94^2=8836$個とするものとなりました。Lab・数字・記号の2個を組み合わせて示し、漢字の番号はそのLabの番号から計算します。例えば、漢字「大」はLabなどの組み合わせとしては「Bg」であり、その番号66および103から$66\times2^8+103=16999$という番号を得、同様に「学」は「3X」であって、番号が$51\times2^8+88=13144$です。

漢字がLabなど2個の組み合わせであることは、「大学」という漢字の並びも、「Bg3X」というLabなどの並びも、コンピュータにとっては区別がつかないということです。それではコンピュータで処理しようとしてもできませんので、方法が二つ考えられました。

一つは、ここから漢字が始まる、ここで漢字が終わる、ということを示す信号を入れることです。

いま一つの方法は、漢字を示すLabなどにあらかじめ演算を施し、Labなどを示さないものにして、その組み合わせで漢字とするものです。いわゆるシフトJISは、JISで定められた番号をこの方式で変換します。「大学」の「Bg」「3X」では、「B」「3」のように先のものがLabなどにならないようにして、$145\times2^8+229$および$138\times2^8+119$となります。ここの最後の119はLab「w」の番号ですが、直前の138がLabなどでないので、それと一体になって漢字を示しているとします。

シフトJISは、MicroSoftがMS-DOSのために開発して、特にパーソナルコンピュータで普及し、Apple Macintoshの漢字TALKも採用しました。JISでも追認されました。　→ JIS X 0208

さて、コンピュータが広く使用されるようになると、文字について、制約された範囲にあるもののみでは不足が言われるようになり、大量のものが要求されるようになります。国際化の流れもあって、漢字の処理も、世界の多様な文字を処理する体系の一環として構想されるようになります。

大量の文字を取り込むためには、割り振る番号の範囲を広げなければなりません。7ビットないし8ビットの体系を十分に使い、しかもその組み合わせを2個に止めずに幾つかにします。現在構想されている最大のものは国際的な規格であり、$2^7\times2^8\times2^8\times2^8=2,147,483,648$（2ギガ）個でしょう。ただし、実際に使用されているのは、まだ$1\times1\times2^8\times2^8=65536$個のうちでしかなく、そのなかの漢字は、中国・台湾・日本・韓国のものを統合した20902個です。　→ ISO 10646/Unicode 2.1

日本では、公的なものとは別に、漢字を中心として数万の文字を収めるプロジェクトが進められています。文字は、番号で指定するほか、画像として文章中に貼り込むことができるようになっているものもあります。

→参考　それぞれのホームページ

▶超漢字3　　150万字が可能であるが、現在は、JIS、KS、GB、CNS、Unicode、GT書体、大漢和辞典などから、17万字以上を収める。TRONに実装し、TRONアプリケーションで用いる。

▶今昔文字鏡　JIS、Unicode、大漢和辞典、甲骨文字などから、9万字を収める。Windows、MacOS、Unixで用いることができる。

▶e漢字　　JIS、康熙字典、大漢和辞典などに収録されている漢字のフォント。Windows、MacOSなどで用いることができる。

▶GT書体　　漢字66733個のフォント。Windows、MacOS、TRONで用いることができる。

なお、コンピュータ上の漢字の体系として有名なものに、EUCがあります。UNIXないしLINUXに載り、

番号の付けかたが違いますが、内容はJISに準じます。番号が違うために、例えばWindowsに電子メールを送っても、読むことができないといったことが起こります。番号は容易に置き換えることができるので、置き換えれば読むことができます。

■ コンピュータのために選ばれてきた漢字

日本が世界で初めて漢字を公的規格としようとしたとき、取り込むことができる漢字は、かななどを加えて94^2＝8836個に制限されていました。当然のことながら、漢字の選定を行わなければなりません。以下、日本における漢字の選定を追います。

今から20年以上前の当時、コンピュータはまだパーソナルなものではありませんでした。コンピュータで漢字を用いるのは、官公庁や保険会社などが人名・地名をデータベース化するためです。

漢字のセットとしては、当時、1850字を定めた当用漢字表が最も有力でした。しかし、当用漢字表は、人名・地名を考慮していません。そこで、当用漢字表を含めて種種の漢字表37点が集められました。その28点以上に掲載されているもの約2000字をまず選定して、これを事務処理に不可欠のものとし、さらに人名・地名などの漢字を追加して、合計2965字が決定します。JIS第一水準の漢字の誕生です。

また、37点の漢字表のうち、次の4点に重複する漢字を抽出し、第一水準にあるものを除く3386字を第二水準として決定します。

情報処理学会標準漢字コード委員会標準コード用漢字表試案	収録漢字数6086
国土行政区画総覧使用漢字	3251
日本生命収容人名漢字表	3044
行政情報処理用標準漢字選定のための漢字使用頻度および対応分析結果	2817

現在は、この第一水準・第二水準の別を意識することはないのではないかと思いますが、選定当初は、コンピュータの能力を配慮したようです。　→ JIS X 0208　→参考　芝野（1997）

こうしていわゆるJIS漢字が規定されるとともに、5年後に見直しをすることも決められました。1983年の見直しでは、主として字形の整理が行われました。近ごろよく言われた「鴎」などの問題が発生することになります。小さくない混乱が社会的に生じた改定でしたので、簡略な字形を採用したこの改定のものを新JISと言い、当初のものを旧JISと言うようになりました。

次の見直しは、1990年に行われました。しかし、ここでは、改定よりも、それに伴って規定された補助漢字が重要です。コンピュータは、ワードプロセッサも含めてパーソナルに普及し、広汎な分野で使われるようになっています。それとともに漢字の不足を訴える声が大きくなり、規格化されることになります。

このいわゆる補助漢字は、記号やアルファベット266個、漢字5801個から成りますが、ほとんど実装されることがありませんでした。シフトJISで番号を付けるのに、余裕がなかった。必ずしも要求されない漢字が多かった。そのような事情があったためのようです。　→ JIS X 0212

漢字の追加は、次いで2000年に規定されます。50種の資料に基づく検討が行われ、すでに規定されている第一水準・第二水準に加えるべく、第三水準漢字1249個、第四水準漢字2436個、記号・ローマ数字など659個が選定されました。第一水準以下、漢字は1万を超えることになります。

補助漢字と第三水準漢字・第四水準漢字とは、2748個が重複します。第三水準漢字・第四水準漢字の番号には、シフトJISのことも配慮されました。補助漢字は、廃案になったことになるでしょう。

時間が前後しますが、第一・第二水準漢字の見直しが1997年にも行われています。ここでは、シフトJISの番号づけが追認されます。しかし、シフトJISを開発したMicroSoftは、日本の漢字も文字の国際的なセットのうちで処理するように、方針を移しつつありました。Appleも同様です。つまり、第三・第四水準漢字も含めて、シフトJISへの配慮は無駄になりつつあります。

文字の国際的なセットというのは、情報交換が国際的になるとともに、国際的に求められるようになってきたものです。大きく二つあり、一つは、JISのような各国の代表が検討していて、つまりISOのものです。いま一つは、コンピュータに関連するメーカ、IBMやMicoroSoftなどアメリカが中心ですが、それが集まって検討していて、番号づけの名をUnicodeと言います。この二つは、ともに1980年代に始まりましたが、求めるところ

は同じですから、1993年には内容の統一を図ります。その後、番号づけのしかたに違いがありますが、文字のセットも番号も同じになるようになっています。　　→参考　ユニコード漢字情報辞典編集委員会（2000）

さて、この国際的な文字セットは、先に触れたように、膨大な量の文字をこなすことができます。差し当たって規定されている最小のものでも、漢字が2万個です。日本のものばかりではないにせよ、JIS第四水準までの倍になります。そのようなことが何を意味するか。各国の漢字の選定の結果が基礎になっているとはいえ、漢字の選定について観念を改めなければならないことになると思います。

なお、いまの国際的な文字セットには、日本の漢字からは、JIS第一・第二水準と補助漢字とが入っています。第三・第四水準は、タイミングの関係で入りませんでした。補助漢字が、国際的に蘇ったことになります。

■　コンピュータによって選ばれてゆく漢字

コンピュータで文字を扱うことができるようになり、私たちにとって、文字がどのようなものとなり、文字を書き記すことがどのような行為となったか、日本語を書き記すことをめぐって顧みてみましょう。

文芸作家が漢字の不足を訴えていること、電子メールで文字化けが起こることは、コンピュータにかかわる文字の問題として、よく知られている代表でしょう。コンピュータでどのように文字を扱いやすくするか、それは、技術的な問題であるよりは、文化的・社会的な課題であるということです。そのことを承知したうえで、そのことには、ここでは触れないこととします。

コンピュータでものを書き記す作業の始めには、キーボードでローマ字を叩くことがあります。結果としては見えないが、実はローマ字が普及しているという事態、それは、コンピュータが普及しないうちは、予想もつかなかったことでしょう。しかも、そのローマ字綴りは、かつて提案された方式のどれとも違ったまま、多分、現在の日本で最もよく行われる方式となりました。

かつて提案されたローマ字の綴りは、語ないし音を意識していました。コンピュータのためのローマ字綴りは、仮名を強く意識しています。仮名を記すための綴りです。

それでいながら、書き記す作業は、ローマ字ないし仮名を意識しているか。何かを書き記すときに、手で直接に紙などに向かっていたときには、文字を書き記しているつもりであったのではないでしょうか。しかし、コンピュータに向かっているときには、ローマ字・仮名の向こうに、語句ないし文を書き記しているつもりであるのではないでしょうか。特に漢字を記しているときの、文字を書いているといういわば抵抗の実感が、なくなっているように思います。

コンピュータに向かったときには、漢字は、キーボードを叩いてから変換し確定するものであり、書くものであるよりは選ぶものです。コンピュータに向かわなくとも、例えば大学入試センター試験でも、漢字の書き取りはすでになく、示されたいくつかのうちから選ぶ問題になっています。見えても読めない漢字があるのと同様に、読めても書けない漢字があるのが普通になるかもしれません。

漢字あるいは仮名は、縦書きに適した形になっているかと思います。コンピュータによって、日本でも横書きに慣らされることになりました。そのようなことまで考えるならば、漢字・仮名は、ますます、人の手でなく、コンピュータに書き記させるべきであるということになります。

それでは、漢字について、手で書くことが無用になるかというと、そうでもなさそうです。コンピュータに向かっていても、漢字をコンピュータに教えなければならないことがあります。画面に実際に書くこともあります。コンピュータがもっている漢字のうちから検索するために、筆順によることもあります。

さて、漢字の問題に絞って、二つのことを最後に述べます。

一つは、漢字の形にかかわることです。

漢字に番号をつけることについて、考え直してよいかもしれません。JIS第一水準や国際的な文字セットは、文字に番号を付けようとしているのであって、字形に番号を付けようとはしていないと理解されます。例えば、「国」「國」「圀」「口」は、一つのものとして、一つの番号を付けられてよい。ということです。日本では、特にJIS第二水準から、この思想がなくなりました。

漢字のみの問題ではなく、ちかごろ、「さいたま市」の書き記しかたについて、「さ」の第2筆から第3筆へのところが、続くのか切れるのか、問題とされました。「い、た、ま」いずれも、同様のことがあるにもかかわら

ず、問題とされなかったところに、考えるべきところがありますが、ともかく、そのようなことは、番号付けに際しては配慮しないでよいということになります。

文字の形の違いは、必要ならば、別の段階で区別すればよい。番号付けに段階を設けることは、二つの字形を一つにするか二つにするか、考えるためにも必要かつ便利であろうと思います。例えば、いずれ行書・草書のようなものを取り込むことがあるかもしれませんが、楷書ほどに簡単に片付けられないと容易に想像することができます。

楷書で細かいことを言っていても、問題はそう簡単ではないと、揶揄するつもりはありませんが、形は奇妙なところをもっていて、なお考えが及びそうにないところがあります。逆説的な例を挙げるならば、欧文タイプライタでは、数字‘0，1’はラテン文字‘O（オー大文字）、l（エル小文字）’で記されていました。

漢字について最後に述べる問題のいま一つは、漢字を使用するのに、常用漢字表に制約されるという感覚がなくなってきているようであることについてです。

ある漢字を使うかどうか、規制するものは、コンピュータがその漢字をもっているかどうかにかかわるところが大きくなっています。そのことを考えるならば、コンピュータで、どのような漢字を使いやすくしておくか、他をそうしないか、ということは、言語計画の一環として、重要な意義をもつことになります。

国際的な文字セットでは、漢字にJISのような水準がありません。かりにあったとしても、またハードウエアの制約があったとしても、ソフトウエア技術は、多分、それを乗り超えます。すなわち、どのような漢字を扱いやすくするかという問題は、ローマ字から仮名・漢字への変換を担うソフトウエアにかかわっていると言ってよいかと思います。ただし、ソフトウエアは、辞書の編成などを使用者に委ねますから、最終的には問題は使用者個個人にかかってくるとも思います。

漢字を野放図に使って文章を書き記してよいということ、例えば行政文書や新聞が一部の人のものになってよいというようなことは、今や考えることができません。そのためにも、漢字の使用についてのガイドラインは必要でしょう。当用漢字表がもっていた思想のようなものが、コンピュータで漢字を扱うときに強烈に意識されなければ、人がコンピュータに敗北することになります。

■ **参考文献**　本文で参照指示をしなかったものを含む。

伊藤　英俊（1996）　漢字文化とコンピュータ。中央公論社、中公PC新書9。

加藤　弘一（2000）　電脳社会の日本語。文芸春秋、文春新書064。

漢字文献情報処理研究会（1998）　電脳中国学――インターネットで広がる漢字の世界。二階堂善弘ほか著、好文出版。

――――（2000）　電脳国文学――インターネットで広がる古典の世界。瀬間正之・谷本玲大・大内英範・岡田百合子著、好文出版。

月刊しにか（1990）　特集　いま漢字の規格化を問う。大修館書店、月刊しにか 1.7。

――――（1992）　特集　古典とコンピュータ。大修館書店、月刊しにか 3.2。

――――（1993）　特集　漢字コードの国際標準化。大修館書店、月刊しにか 4.2。

芝野　耕司（1997）　JIS漢字字典。日本規格協会。

清水　哲郎（2001）　図解でわかる文字コードのすべて――異体字・難漢字からハングル・梵字まで。日本実業出版社。

人文学と情報処理（1996）　文字コード　現状と未来。勉誠社、人文学と情報処理 10。

――――（1999）　特集　歴史学系データベースと文字コード。勉誠出版、人文学と情報処理 25。

――――（2000）　特集　文字コード論から文字論へ。勉誠出版、人文学と情報処理 26。

バーカーツ　Birkert, Sven (1994)　The Gutenberg elegies.

＝（1995）　グーテンベルクへの挽歌――エレクトロニクス時代における読書の運命。船木裕訳、青土社。

平凡社　（1998）　電脳文化と漢字のゆくえ――岐路に立つ日本語。平凡社。

ボルツ Bolz, Norbert（1993）　Am Ende der Gutenberg-Galaxis : die neuen Kommunikationsverhältnisse.

＝（1999）　グーテンベルク銀河系の終焉――新しいコミュニケーションのすがた。識名章喜・足立典

子訳、法政大学出版局、叢書ウニベルシタス 657。
マクルーハン　McLuhan, Marshall (1962)　The Gutenberg galaxy : the making of typographic man.
＝（1986）　グーテンベルクの銀河系――活字人間の形成。森常治訳、みすず書房。
三上　吉彦・池田　巧・山口　真也（1993）　電脳外国語大学――パソコンで世界の言葉に挑戦！。技術評論社。
美崎　薫（2000）　超漢字超解説――BTRON 仕様革命的 OS の全貌。工作舎。
文字鏡研究会（1999）　パソコン悠悠漢字術――今昔文字鏡徹底活用。紀伊国屋書店。
安岡　孝一・安岡　素子（1999）　文字コードの世界。東京電機大学出版局。
ユニコード漢字情報辞典編集委員会（2000）　ユニコード漢字情報辞典。田中裕一・谷村英治・古谷先男・松岡栄志著、三省堂。
ランディ　Lunde, Ken (1999)　CJKV information processing. O'Reilly & Associates.
レビンソン　Levinson, Paul (1999)　Digital McLuhan : a guide to the information millennium.
＝（2000）　デジタル・マクルーハン――情報の千年紀へ。服部桂訳、NTT 出版。

石井　久雄（いしい　ひさお）
1950年生。東北大学文学部卒業。東北大学大学院文学研究科博士課程後期課程中退。同志社大学教授（文学部文化学科）。日本語史専攻。主要業績に、「国語学と日本語研究・言語研究」（第 2 回国立国語研究所国際シンポジウム報告　新しい言語理論と日本語 pp. 7 -20。1997年）、「昔はどう言ったかと，知りたいとき」（国立国語研究所報告104　研究報告集13　pp. 31-76。1992年）、「本文批判」（国立国語研究所報告94　研究報告集 9　pp. 1-25。1988年）。