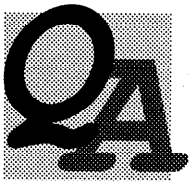
素朴な疑問



漢字コードの種類と相互関係はどうなっているの？†

宮　澤　彰††

1. 内部コードと交換用符号

漢字コードの話というと，JISコードとか，シフトJISコードといった名称が一般的にはなじみ深いものでしょう．しかし，JISコードは交換用符号，シフトJISコードは内部コードで，同列に並べて論じるのは誤解のもとになるということを最初に指摘しておきたいと思います．

ここでいう交換用符号とはJIS（日本工業規格）やISO（国際標準化機構）で決められている符号化文字集合のことです．普通7ビット94文字とか，2×7ビットの94×94文字などの枠に文字を割り当てた表になっています．交換用符号は国際的な登録制度があり，登録されると識別のための終端文字が与えられます．

一方，内部コードはハードウェア，ソフトウェアシステムの中で使われる値と文字との対応です．1つまたは複数の交換用符号をJIS X 0202「情報交換用符号の拡張法」という規格（ISOでは2022）に従って7ビットまたは8ビットの環境に「呼び出す」ことにより内部コードとなります．このX 0202はかなり複雑な規格で，同じ交換用符号でも様々な呼出し方が可能であり，呼出し方が異なれば内部コードとしては違ったものとなります．さらに，JIS X 0202によらない方法で交換用符号を使用しているシステムも多く，これらも多くの異なった内部コードを生み出す原因となっています．

最初に交換用符号について見てみましょう．

2. JIS X 0208 情報交換用漢字符号

「JISコード」といえばこれを指すといっても過言ではないでしょう．94区×94点の枠に記号類，英数字，仮名，ギリシャ文字，ロシア文字と6千3百余字の漢字を割り当てています．以前はC 6226という番号でした．

有名なコードですが，現在の1990年版以前に，1983年版，1978年版があり，やや異なっていることは広くは知られていません．90年版と83年版の異なりは人名用漢字の追加にともなう2字の追加のみですが，78年版から83年版への変更は記号類71文字，漢字4文字の追加のほか，字体の入れ替えをともなう大きなものでした．なお，83年版と90年版は同じ終端文字で識別されますが，78年版の終端文字とは異なっています．また，一般に旧JIS，新JISという言い方をするときは78年版と83年版以降の終端文字や字体について言っています．

字体の入れ替えは81年の常用漢字表の公布に対応するもので，たとえば「桧」と「檜」などのコード値が入れ替えられました．あるいは「鴎」のように字体が換えられたものもあります．この変更は少なくとも字体に注意深い人々にはかなりの混乱を引き起こしました．JISとしては最新版のみが有効なわけですが，市場に出回っているハードウェア，ソフトウェアは必ずしも最新版に従ってはいません．たとえば新しいプリンタで印刷したら字体が変わっていたとか，画面と印字が違うとか，別のパソコンに持っていったら字体が変わってしまったなどの現象が起きています．現在もこの混乱はおさまっていません．このため，自分の使用しているシステムがどの版のJISであるかは注意しておく必要があります．

† What are the *Kanji* Character Sets and their Relations? by Akira MIYAZAWA (R & D Department National Center for Science Information Systems).

†† 学術情報センター研究開発部

3. JIS X 0212 補助漢字集合

X 0208 の補助として用いるために 1990 年に制定されたもので，記号類 21 字，ウムラウトなどの拡張アルファベット 251 字，および X 0208 に含まれない漢字 5801 字を含んでいます．漢字はメーカや大規模ユーザの外字表を集め共通性のあるものが選定されたほか，X 0208 の 78 年版の文字で 83 年版での改訂にともなって字体を換えられたもののうち 28 字が選定されています．

この交換用符号は一般的なコンピュータシステムにはほとんど普及していません．

4. 諸外国の「漢字コード」

漢字を含む交換用符号は中国，韓国，台湾，米国，およびベトナムで規格化されています．中国では GB 2312 という番号で JIS X 0208 と似た構成の中に記号類，英数字，日本語仮名，ギリシャ文字，ロシア文字に漢字を 6 千 7 百余字を含んでいます．このほかに，GB 2312 に対応する繁体字，拡張の文字集合などが別の番号で規格となっていますが，実用システムで普及しているのは GB 2312 のみと言っていいでしょう．

韓国では KSC 5601 という規格で，やはり JIS と同様の構成の中に記号類と，英数字，日本語仮名，ロシア文字，ハングル 2 千 3 百余字と，漢字 4 千 8 百余字が含まれています．韓国でもまた，記号類，拡張アルファベット，ハングル，漢字を含んだ拡張セットが制定されていますが，普及はしていません．

台湾では CNS 11643 という番号の規格があります．これは，94×94 文字の面を 2 面使い，記号類，英数字，ギリシャ文字に 1 万 3 千余字の漢字を含んだものでした．その後，漢字が拡張されて数面に及んでいます．実用システムでは CNS の他，BIG 5 など何種かのコードが用いられているようです．

米国では ANSI Z 39.64 という番号で書誌用の東アジア文字が規格化されており，図書館目録データベースの世界で使用されています．台湾で考えられた CCCII という交換用符号をもとにしたもので，3×7 ビットの構成の中に，日本語仮名，ハングル千 9 百余字および漢字 1 万 3 千数百字を含んでおり，漢字の異体字関係をコードの構造で表現しているところに特徴があります．

ベトナムでは 20 世紀初頭まで字喃（チュノム）と呼ばれる漢字を借応用した文字が使われていました．近年これをコンピュータ処理するための交換用符号が規格化されています．

5. ISO 10646 UCS と JIS X 0221

ISO/IEC 10646 Universal Multiple Octet Coded Character Set（UCS）は 2 バイトまたは 4 バイトの枠に世界中の文字を集めた交換用符号です．ただし，現在のところ 2 バイトの枠内 Basic Multilingual Plane: BMP しか規定されていません．

UCS は前期の ISO 2022（JIS X 0202）の 94 文字枠にはしばられず，このまま内部コードとして使用されることを想定しており，16 ビットまたは 32 ビットの全コード値を使用します．16 ビットでは 65536 文字の枠があることになります．現在の BMP の中にはラテンアルファベット，拡張ラテン（ウムラウトなど），ギリシャ，キリル，アルメニア，ヘブライ，アラビア，漢字等々の世界中の文字および大量の記号類，合計 3 万余字が含まれています．

UCS の決定に至る審議経過はすっきりしたものではありませんでした．当初 ISO とは別個に活動を続けていた米国の UNICODE コンソーシアムの案が最終段階になって採用されるということになったためです．いきさつはともあれ，国際標準と業界のコードという 2 本立ては避けられました．

漢字部分に関しては CJK-JRG（China, Japan, Korea Joint Research Group）というところで各国の交換用符号の文字を統合化して作成した 20902 文字の表が採用されました．前章に述べた日本，中国，韓国，台湾の交換用符号を集め，同字，字体のわずかな違いであるものを横並びにし

Row/Cell Hex code	C G - Hanzi - T		J Kanji	K Hanja
098/208	拐	拐	拐	拐
62D0	0 - 3955 0 - 2553	1 - 4D66 1 - 4570	0 - 327D 0 - 1893	0 - 4E58 0 - 4656

図-1　UCS 漢字部分の一部

て康熙字典順に並べたもので，**図-1**のようになっています．G，T，J，Kの各列は中国のGB，台湾のCNS，日本のJIS，韓国のKSに対応しており，字の下に書いてある数字が交換用符号の種類と16進コードおよび句点番号です．したがって各国の交換用符号との間の変換テーブルになっているわけです．日本の列では交換用符号の種類0および1でJIS X 0208，X 0212の全漢字が含まれています．

JIS X 0221国際符号化文字集合（UCS）はこの国際規格を翻訳してJISとしたもので1995年制定，4月に出版予定です．内容は国際規格とまったく同じですが，付属書で日本用のサブセットが定義されています．UCSに対応したオペレーティングシステムが最近ようやく市場に出始めたところです．

6. 内部コード

日本のコンピュータシステムはそのほとんどがJIS X 0208を使用しているが，最初に述べたように内部コードとしてはいくつかの使用法があります．

内部コードのコード値を表現するために以下の記法を導入しましょう．

21	16進数8ビット値
21 22	8ビット値の列
{21,22}	21と22の集合
{21-23}	={21,22,23}
21{A1,A2}	={21 A1,21 A2}
{21,22}{A1,A2}	
	={21 A1,21 A2,22 A1,22 A2}

EUC（Extended Unix Code）は日本語環境ではJIS X 0201「情報交換用符号」のローマ文字用図形文字集合94文字および片仮名用図形文字集合94文字，JIS X 0208，拡張用の94×94文字集合が使用できます．{21-7E}にローマ文字用集合，{A1-FE}{A1-FE}にX 0208，8E{A1-FE}に片仮名用集合，8F{A1-FE}{A1-FE}に拡張用文字集合を呼び出すものです．この内部コードは制御文字SS2（8E），SS3（8F）の後ろで最上位ビット1を使用していることを除けばX 0202の拡張方法に従っています．なお，EUCではプログラム内部での処理用に2バイト等長の処理コードも規定されています．

シフトJISは日本のパーソナルコンピュータのほとんど，およびワークステーションの一部で使われており最も普及している内部コードです．X 0201およびX 0208が使用でき，{21-7E}にローマ文字用集合，{A1-DF}に片仮名用集合，{81-9F，E0-EF}{40-7E，80-FC}にX 0208という割当てです．これは，X 0201の未使用部分を2バイト文字の第1バイトとして使用し，47×188（=94×94）の2バイトコード値の集合を作って，X 0208を大小順を保存しながらその中にいれたものです．シフトJISの欠点としては，制御文字部分を図形文字として使用している点と，各種のシステム間で外字領域の互換性のないことがあげられます．

メインフレーム用のOSでは各社異なる内部コードを使用しています．例として富士通のJEFコードを取り上げると，{40-FF}にEBCDICまたはEBCDIK，{41-7F}{A1-FE}に富士通拡張文字，{80-A0}{A1-FE}にユーザ定義外字，{A1-FE}{A1-FE}にX 0208を割り当てています．1バイト2バイトの切り替えは制御文字を用いてロッキングシフトで行っています．ロッキングシフトと呼ぶのは，1バイトの中で2バイト開始の制御文字が現れると以降の文字を2バイトとして扱い，1バイト開始の制御文字が現れるまで続くという方式です．これに対し，シフトJISやEUCでは1バイト2バイトの切り替えのための制御文字を必要としません．

IBMのメインフレーム用OSは{41-FE}{41-FE}に独自の文字集合を割り当てており，JIS X 0208との対応はコード変換で行っています．他社のメインフレーム用OSはロッキングシフト方式，使用する交換用符号などJEFと類似していますが，割当てスペース，制御文字のコード値は各々異なっています．

以上のようなところが主な内部コードですが，通信などにからんでJISと呼ばれる内部コードがあります．JISのX 0201とX 0208とを使用し，エスケープシーケンス1B 24 42でX 0208へ，1B 28 4AでX 0201へロッキングシフトする方式です．42と4Aは各々X 0208とX 0201（ローマ文字）を識別する終端文字です．呼び出し場所は{21-7E}{21-7E}および{21-7E}で，X 0201のローマ文字，片仮名の切り替えはSI

(0F), SO (0E) で行います．システムによっては，これをJIS 7と称し，片仮名を{A1-FE}に呼び出してSI，SOを使用しない方式をJIS 8と称しています．

実は上記のエスケープシーケンスは現在のJIS X 0202では旧式となっていますが，83年版のX 0208にはこのように規定してあったものです．新しいJISに従えば1B 24 28 42 0Fとなるのですが，これ以外の指示呼び出し方法も可能です．また，78年版のX 0208に対しては終端文字が40となります．これらの理由でいくつかの変種がありますが，このいわゆるJISコードは端末との通信，メールのやりとりなどに広く使われている内部コードです．

参 考 文 献

交換用符号については各規格を参照のこと．また，内部コードについては各OSのマニュアルを参照のこと．

(平成7年4月3日受付)

宮澤 彰 (正会員)

1949年生．1975年東京大学大学院理学系研究科修士課程（科学史および科学基礎論）修了．理学修士．同年国文学研究資料館助手．1983年東京大学文献情報センター助教授．1986年学術情報センター助教授．現在同センター教授．データベース作成の基礎論という観点から，文字コード論，計算言語学，情報図書館学などに興味を持つ．東洋音楽学会会員．