解　説

S 5000 漢字情報処理システムにおける文字コード*

菊　池　利　徳**

1. まえがき

情報処理における漢字の必要性が強調されて久しくなる．しかし，特定の使用目的や限られた範囲では試作研究の段階を脱しているものの，一般的に情報処理における情報媒体として漢字を取り扱うとなると，いくつかの大きな障害につき当たる．文字数の多いことや各文字が複雑であることによる装置への負荷は当初から予想されたものであり，これらハードウェアの技術の制約が漢字情報処理の可否を左右して来たことも事実であった．一方，長い漢字の歴史の中から発生した諸問題，日本人の漢字に対する意識，情報処理における漢字の役割など漢字情報処理の背景ともいえる部分にもいくつかの解決されない問題を残したまま現在に至っているのも事実である．具体的には，文字種の選択やその配列に代表される文字コード（日本語情報処理で扱う全ての文字，記号に付加されるコード）がひとつの大きな問題となっている．

本稿では，S5000シリーズ漢字情報処理システムにおいて標準となる文字コード（TK コードと呼称しているので以下この名称を使用する）の決定までのいきさつを述べるとともに，標準文字コード化のむずかしさ，および TK コードの不備を補う方法としての JOB FONT MASTER 方法について述べることにする．

2. 文字コード

漢字は英数字や仮名のようには文字セットを限定しがたい．基本文字数としては当用漢字，人名用漢字，補正漢字の合計 1,970 字をまずあげることができる．しかし，これら 2,000 文字弱の文字数では日本語処理を満足させることは到底できない．基本文字数が 2,000 字では不足だとすると，特にこれといった基準がなくなってしまう．漢字情報処理の基本をなす文字コードの標準化が遅れ***，漢字情報処理の一般化が進むに従い漢字コードの互換性や漢字のコンピュータ処理に影響を及ぼす結果となっている．標準文字コードは存在しないが，漢字情報処理システムを提供しているメーカ側からはすでに数種の文字コードが発表されている．残念ながら，新聞用 CO-59 コード以外は1メーカにつき1文字コード体系という混乱ぶりである．この多種文字コード現象はコード化に対する検討不足または協力不足にも原因があるが，漢字情報処理装置そのものも影響をおよぼしているという事実を見逃すことはできない．冒頭でも述べたように，これは情報媒体として漢字を実用化するためにはハードウェア技術に負わねばならなかった当初の事情を顧りみれば無理からぬことかもしれない．しかし，従来の情報処理分野における標準化の経緯と漢字情報処理の将来を併せて考えた場合，現在の状態が長く続けば続くほど，漢字という特殊性ゆえに混乱の度を増すのではないかと思われる．

現在でも標準文字コードが存在しないのであるから，まして我々が漢字情報処理システムの商用化に着手した当時（昭和44年）はその気運さえもなかった．システムの商用化に際し，文字コードの重要性を以前の経験から得ることができた．当時公開されていた文字コードとしては新聞界で使用される CO-59 コードがあった．しかし，CO-59 コードはコンピュータ処理の立場から見た場合熟慮されたコード体系とはいえないものであった．その理由は（現在でもその傾向にあるが．）漢字入力装置上の文字配置を従来の活版部門における文選（活字を選び文章を組立てる．）の形態を取り入れ，しかも配置された位置にもとづいて機械的にコードを発生させる方式をとっていることである．漢字出力装置も，入力装置から機械的に発生したコードで直接出力する方式であり，その間にはコンピ

---

*The character code for S 5000 series KANJI Processing Systems, by Toshinori Kikuchi. (Systems Dept., Showa Information Systems Co., LTD)

** 昭和情報機器(株)システム部

*** 昭和46年10月，「標準コード用漢字試案」が情報処理学会規格委員会から提出されている．

ュータ介入の必要がほとんどなかった．すなわち文字コードが物理的な位置を示すものである．現在メーカ側から提供される文字コードもこの方式を基本的に踏襲している．漢字が情報媒体であるためにはコンピュータ処理などの加工に適したコード体系を持つ必要があり，TK コード化に際してはこの点に十分な配慮をしたつもりである．もちろんコード決定の要因は他にもいくつかあり，次にあげる項目に要約することができよう．

1. 漢字情報処理に必要な文字種
   A．基本漢字
   B．基本漢字以外の漢字
   C．漢字以外の文字，記号
2. 漢字の正字，異体字の区分および一般的な取り扱い
3. 漢字の collating sequence
4. コンピュータ処理と文字コード
5. 漢字入出力装置と文字コード

文字コード化に当たり最初に問題となるのは，4万とも5万*とも言われる漢字から情報処理で必要とする漢字を選択することである．幸いにして，次にあげる貴重な調査研究資料があり，使用頻度調査や文字種調査の参考資料とすることができた．

現代雑誌の用語用字[1]	3,328 字
新聞文章実態調査[2]	2,613 字
姓氏の文字[3]	2,482 字

これらの資料のうち，雑誌や新聞など日常の日本語情報処理は3,000字前後の漢字で十分であることがわかる．しかし，印刷関係では常用漢字と言われるもので4,000字[4]，標準漢字で8,400字[5]が用意されている．前記調査資料でも明確なように，平易な用語用字に心がけている新聞関係でも約6,000字が漢字表に登録されている[6]が，それでも漢字表外の漢字が必要になるということである．

漢字を扱う上で，問題点の1つにあげられるのが地名および姓名である．前記調査資料のうち姓氏の文字(一部には漢字以外も含む)が2,482字と比較的少ないのは，主に印刷発行物からの調査のため異体字を正字(後述)にまとめる集計結果となったことによる．情報としては当然，姓と名の両方が一体となって使用されるから，姓名に使用される漢字の文字種調査が必要となった．しかし，当時それらを調査した資料と目されるものはなかったものの，ある信託銀行で使用している人名漢字表を入手することができた．この漢字表は手書き文字やふりがなの確認のために使用され，常に更新されていたものである．内容は辞典に記載されていない文字を含め，手書の原薄からそのまま転記され，調査集計した結果では，姓名で合計5,371字（厳密調査を要す）が登録されている．この中には手書き原簿からの転記であるため全ての文字が有効とは考えられない．多少の誤差があるにせよ，姓名だけでも正確（この漢字表に登録されている文字そのものの正確度は高いと思われる）に表現するためには約5,300字の漢字を必要とする．

一方，地名に使用される漢字を集計した結果では全国10,992市町村[7]名で2,279字が使用されている[8]．

さらに，文字種決定の要素となるのが異体字や辞書に記載のない文字の取り扱い方である．漢字には正字，本字，異体字の区別があり，これらは長い漢字の歴史の中で発生した問題である．これらの解釈は必ずしも一定していないが，“現代字体字典[9]”の解釈を基本として次のように定義する．

正字：国などの権威によって認定された標準，基準となる字体を正字とする．たとえば“浜”は従来“濱”の俗字として用いたが当用漢字で正字と認められた．

本字：なりたちから考えて基本体と認められる字体を本字とする．“浜”に対して“濱”は本字である．

異体字：基本となる字体から字形が逸脱し，多くの人に通用した結果，定着した字体を異体字とする．異体字の中には，本字がくずれた形で通用している**俗字**．文字の全体または一部を省略した**略字**．誤った字形のまま部分的には通用しているが使用が望ましくない**誤字**．異なったしくみを持つが，なりたちから考えて同音同義の字である**別体字**などがある．

このうちで特に問題となるのは，当用漢字として新しく認められた正字とその本字の関係，および正字と異体字との関係である．特に問題となるのは姓名に関するものである．前者の正字と本字との関係では，当用漢字表が告示**される以前は姓名の多くは本字（当時は正字であった）が使用され，戸籍などは本字で登録されている．現在でも，公式書類の中には戸籍上の姓名漢字を必要とするものもあり，通常でも意識的に本字を使用している人もある．このうちでも正字と本字が大幅に異なる場合，たとえば万（萬）や竜（龍）などは比較的使用頻度が高いことから独立した文字と

---

* 康熙字典：47,035字，大漢和辞典（大修館）：50,291字，このうち死語となった漢字が多数ある．

** 昭和21年内閣告示第35号

して扱った．一見して異差の明確でない文字は文字種選択の対象から除いた．後者の正字と異体字の関係では原則として全て独立した文字として扱った．この中には使用頻度と一般性を優先させたために“淵渕淵”のように同一レベルで扱った文字や“桧”(檜)“篭”(籠)のように正字より優先する文字もある．“卆”(卒)“曽”(曾)などの略字も独立した文字として扱うが“議”の“義”の部分を“ギ”と略す部類のもの（訛字と言い，誤字とは異なる）は文字種選択の対象から除いた．このような方法で8,844の漢字を選択した．

コード化でのもう1つの問題は選択された漢字の配列で，コード体系の可否を左右する．漢字を配列する手段として次の方法が考えられる．

1. 部首画数順配列
2. 総画数順配列
3. 音訓五十音順配列
4. 使用頻度順配列
5. 熟語配列

このうち1～3は漢和辞典や国語辞典などに使用される配列方法で比較的慣れた配列といえる．1の方法は“当”が田の部であったり，“相”が目の部というように，漢字の意味を基本として配列される康熙字典以来の方法と“当”はツの部に，“相”は木の部に配列する漢字の形態を基本とする方法があるにせよ，全ての漢字を配列することは可能である．2の方法も数え方を統一することによって可能であるが，漢字を検索するのには適していない．3の方法は2とは逆に漢字の検索には適した配列であるが，限られた読みの個所に集中すること，漢字を読む能力を必要とすることに欠点がある．通常の漢字知識では 2,500～3,000 字が限度であり，8,000 字以上もの漢字をこの方法で配列するには問題がある．4，5の方法は限られた文字種であれば可能であるが，全ての漢字については意味がない配列方法である．以上のように，1つの方法で配列するならば1の部首画数順配列以外に考えられない．しかし，1つの方法で配列する必要性がないことと，3や5の特徴も見逃すことはできない．

最後に，漢字情報機器と文字コードの関係について考えてみたい．漢字情報処理の場合，扱う対象である漢字の文字種が多く各文字が複雑であることから，入出力の手段も特徴あるいくつかの手法がとられている．そのため，ハードウェアの機能が確認された後にコードが決定されるケースが多く，現在数種類ある文字コードの大部分が機器に都合のよいコード体系をとる結果となり，標準コードの決定を遅らせる原因の一つとなっている．

TK コードの決定に際しても，部分的ではあるにせよ例外ではなかった．入力から出力までコード変換なしで処理できることを設計概念としたため，入力装置上の文字配列と文字コードを一致させる必要があった．その結果，次の条件を満たす入力装置および文字コードの検討がなされた．

1. 漢字の検索を容易にする配列であること．
2. 基本漢字として 2,500～3,000 字を収容すること．
3. 英数字，仮名および高使用頻度漢字は操作しやすい位置に配置すること．
4. 利用者側で特殊文字を使用する場合を考えて多少の予備を設けること．
5. 熟練した人でも文字の位置を記憶できる範囲は3,000 字前後と言われているところから鍵盤の総文字数を最高3,000前後とすること．

以上の条件を検討した結果次の結論を導き出すことができた．

A．2,4,5 から鍵盤の総文字数を3,072字分とする
B．2とAから漢字総数を2,700字とする（英数記号および仮名で約300字必要である）
C．3とBから高使用頻度漢字約2,000字とそれ以外を鍵盤上で別位置とする（文字コードがブロック化されることを意味する）
D．1とCから高使用頻度漢字を音訓五十音順配列とし，それ以外の漢字は部首画数順配列とする

ここで出された結論は基本原則であり，細部にわたる検討が加えられた．その結果，次のような構成をとることとなった．

1. 鍵盤内第1グループ
   ひらがな，カタカナ，英大小文字，数字，記号など漢字以外の文字記号
2. 鍵盤内第2グループ
   全ての漢字処理に共通して使用頻度の高い漢字60字（昭和，年月日など）を熟語配列
3. 鍵盤内第3グループ
   当用，人名用漢字を中心に使用頻度が高く音訓索引の容易な漢字2,016字を音または訓の五十音配列
4. 鍵盤内第4グループ
   市区郡名，姓名などに使用され，第3グループにつぐ高使用頻度漢字624字を部首画数順配列

表-1　TK コード構成表

区　分	鍵盤内字	外 字 1	外 字 2	合　計
漢　字	2,700	4,608	1,536	8,844
仮　名	168	84(84)		252(84)
数　字	20(10)	20(10)		40(20)
英　字	52	52(52)		104(52)
記　号	49(4)	319(28)	512	880(32)
予　備	73	37		110
計	3,072(14)	5,120(174)	2,048	10,240(188)

注　( ) 内はハーフ・サイズ文字数示す.

以上のように，鍵盤内漢字 2,700 字を使用頻度と漢字の性格から 3 グループに分割し，各々を文字種に合った配列を行い，機能的な文字コード体系とした．

記号を含めて鍵盤外文字は出力装置との関係において文字コードが決定された．S5000 システムは処理量の拡大にともなって文字イメージ記憶容量が拡張できるようにモジュール設計がなされている．そのため，文字種および文字コードはシステムに合わせたモジュール構造をもったコード体系とした．1 モジュールは 512 字で構成され，漢字 12 モジュール 6,144 字，記号 2 モジュール 831 字を鍵盤外字としてコード化した．漢字各モジュールは部首画数順配列とし，どのモジュールに該当するかは使用頻度調査，各レベルの辞書，漢字表などから決定した．記号第 1 モジュールは主に学術専門書や文書情報に使用される記号を収容し，第 2 モジュールは主に印刷関係で使用される特殊記号を収容した．文字コードでは，記号第 1，第 2 モジュールは漢字第 9 モジュールと第 10 モジュールの間に位置している．**表-1** は TK コードの構成を示した表である．

4. JOB FONT MASTER 方式

TK コードは S5000 シリーズ漢字情報処理システムにとって最も標準と考える文字コード体系であり，必ずしも他のシステムに適応できるコードとは言えないし，S5000 システムにおいても個々の作業単位に見ると，必ずしも望ましいコード体系ではない場合がある．その理由の 1 つには，処理目的によって使用文字種や使用頻度が異なり，しかも同一作業内においても変動することがあげられる．この傾向は漢字に限ったことではなく，特殊記号や科学記号なども同じことが言える．

TK コードを含めて標準コードと言われるものは，あくまで一般的または平均的な利用を目的として作られている．S5000 システムのように 32×32 ドットと 1 文字を表現するのに 1,024 ビットも使用する方式では尨大な記憶容量を必要とする．新しい記憶素子の開発や記憶素子が廉価になったとは言っても，8 千とか 1 万文字を常備することは経済的，使用効率の点で望ましくない．特に文字発生装置として磁気ディスクを使用した低速出力装置は別として，高速度装置ともなると大きな障害となる．しかも，1 万文字を固定して常備しても日本語処理は完全だとする保障がないところに漢字情報処理のむずかしさがある．

前項の使用文字種調査の通り，使用目的を限定すれば，漢字だけで約 2,500～3,500 程度，その他の文字や記号を入れても約 3,000～4,000 字分の記憶容量で処理が可能である．当然，文字コードや文字イメージ記憶装置を JOB 単位に設計することはできないが，文字コードと文字イメージ記憶装置を限られた JOB を処理する間だけ専用化できればよいわけである．それを可能にするのが **JOB FONT MASTER** 方式である．通常この方式を **JFM 方式**と称しているので以下この名称を使うことにする．

JFM 方式の基本概念は “JOB 単位に専用の文字コードと専用の文字イメージ記憶装置を持つこと” である．物理的に多種類の文字イメージ記憶装置を持つのではなく，非固定記憶 (RAM) とし JOB 単位に内容を書き換えて使用する意味である．**FONT MATTER** とは文字イメージと文字コードで構成される文字に関する MASTER file で，S5000 システム全体で原則として 1 種類しか存在しない．JOB 単位の専用コードを **JOB CODE** と言う．**JOB FONT MASTER** は FONT MASTER から JOB で使用する文字について，TK コードの代りに JOB CODE を持った一種の FONT MASTER file である．

図-1 (次頁参照) は TK コードで作られた MASTER file (漢字処理以外の部分は問題としない) から **JOB MASTER** file と JOB FONT MASTER file (JFM file) を作成するフローを示した図である．JOB MASTER file は TK コードを JOB CODE に変換された MASTER file である．MASTER file M は漢字処理の目的で新しく作られたか，または JFM 方式を採用する以前に漢字処理に使用されていたもので，文字コードの部分は TK コードである．

処理は MASTER file (M file) の TK コードの部分が取り出され，固定コード (英数字，カナなどは基本コードとして固定した方が便利である) 以外の文字

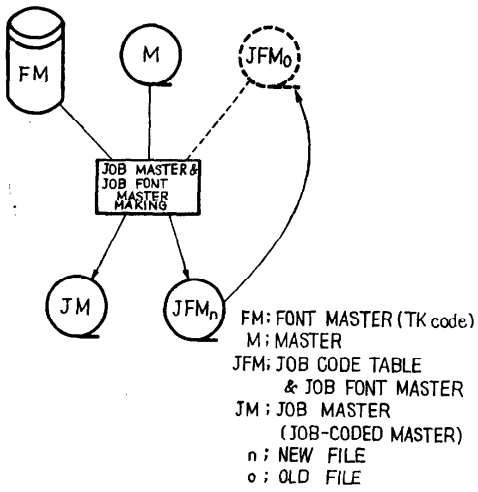

図-1 JOB MASTER, JOB FONT MASTER の作成方法

コードについて JOB CODE 化される．JOB CODE 化処理は JOB CODE table を参照し，すでに登録されている TK コードであれば，table に TK コードと対応して記入されている JOB CODE に変換して出力する．table に登録されていない TK コードであれば，TK コードと新しい JOB CODE を登録する．JOB CODE の決め方は物理的な文字イメージ記憶容量内のアドレス（アドレスを指定すれば文字イメージを読み出すことができる）のうち，固定コードに該当するアドレスを除いたものを順に指定する．すなわち，JOB CODE table は TK コードと JOB CODE の対応表で，JOB CODE は TK コードの出現順に連続した物理的アドレスとして決定される．

この処理によって，MASTER file の TK コード部は全て JOB CODE に置き換えられて JOB MASTER file (JM file) として出力される．全ての DATA 処理が終了した時点では JOB CODE table が作成されており，その内容は JOB FONT MASTER file (JFM$_n$ file) の先頭に出力される．JFM file には JOB CODE table に続いて，JOB CODE として登録されたものだけを FONT MASTER file (FM file) から取り出し，TK コードに代えて JOB CODE を付加し JOB FONT MASTER が出力される．同一 JOB で複数の MASTER file が存在する場合，JFM$_n$ file は JFM$_o$ file となり JFM file は更新される．このようにして作成された JM file は TK コード部が JOB CODE に変更されただけで M file と同じと見ることができる．

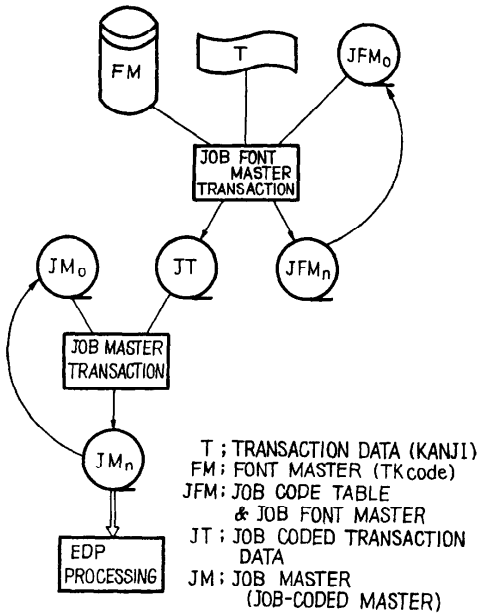

図-2 JOB MASTER, JOB FONT MASTER の TRANSACTION 方法

図-2 は図-1 で作成された JM file と JFM file の transaction 処理方法を示した図である．MASTER file である以上 transaction 処理が発生し，併せて JFM file の transaction 処理も必要になる．transaction 処理は JFM file に関するものと JM file に関する処理とに分けられる．

JFM file の transaction 処理では，TK コードから成る TRANSACTION DATA file (T file) だけについて JOB CODE table の更新を行う．更新の方法は図-1 で述べた方法と同様であり，transaction data の中に更新すべき未登録の TK コードが含まれる場合は JFM$_n$ file が出力される．T file は図-1 の M file 同様，JOB CODE table を参照し JOB CODE に変換された JT file となる．続いて，JT file をもとに JM file の transaction 処理が行われる．JM file, JT file とも JOB CODE 化されており，この時点ではコード上の問題は解決されているので JFM 方式とは直接関係のない通常の transaction 処理を行えばよい．

以上述べたように，JFM 方式と言っても図-1 の JOB MASTER と JOB FONT MASTER を作成するプログラムと図-2 の JOB FONT MASTER 更新

のプログラムを用意するだけである．しかも，図-1の処理は最初の1度だけで以後は図-2の transaction 処理が発生するだけである．この処理は T file の量にもよるが通常は非較的少量と考えられるので，JFM 方式を採用したために必要とする処理時間は問題にならない．JFM 方式を採用して効果があるのは，漢字処理を目的とした MASTER file を持ち，使用される文字コードが予想されないようなものである．

4. むすび

漢字情報処理に対する関心が高まるに従い，以前のようにただ漢字が扱えればよいという概念から，文字品位とか処理システム相互間の互換性などを問題視する傾向にある．互換性すなわち文字コードに関して，標準文字コードを1種類にまとめることは困難であろうが，せめて基本文字を標準化することを早急に行う必要がある．本稿で述べた TK コードの決定に当っては，細部にわたるまでの絶対の根拠があったわけではない．部分的には主観的な判断をよぎなくされ，学会誌に発表するまでの科学性を持っているとは考えていないが，漢字情報処理にたずさわる1人として，あえて筆をとった次第である．最後に，現在工業技術院を中心として，51年を目標に国で使用する漢字字種コードが検討されており大いに期待するものである．

参考文献

1) 国立国語研究所：現代雑誌九十種の用語用字，(1963)．
2) 朝日新聞調査研究室：新聞文章実態調査，(1949)．
3) 野村広：5万5千の姓氏に使われた文字の調査，(1969)．
4) 全日本漢字配列協議会：常用漢字目録，(1968)．
5) 日本活字工業：標準活字目録，(1966)．
6) 共同通信社：漢テレハンドブック，(1969)．
毎日新聞社印刷局：活字表，(1958)．
7) 日本リーダースダイジェスト社：地名コード総覧，(1970)．
8) 昭和情報機器：漢字の字種に関する研究，(1973)．
9) 日本書道研究所：現代字体字典，(1970)．

(昭和50年3月14日受付)
(昭和50年8月12日再受付)