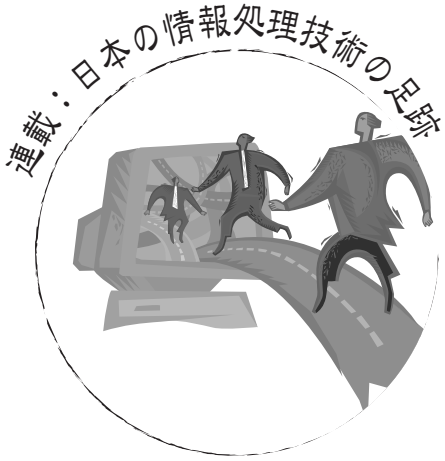


漢字・日本語処理技術の発展：漢字コードの標準化

芝野 耕司
東京外国語大学
shibano@aa.tufs.ac.jp

我が国における漢字コードの標準化

我が国における文字コード，符号化文字集合（Coded Character Set）の標準化は，もうほとんど「歴史」というに相応しい「とき」を経てきた．最初の文字コードの標準化についての「情報処理」での報告は，和田弘氏が1960年の「情報処理」Vol.1, No.2に電気試験所で1958年に発足したコード会の報告[1]に始まる．半世紀を経て，まさに，「歴史」，それも「科学史」の対象とする「とき」に至ったのであろう．

文字コードの検討から半世紀，最初のJIS漢字コード（JIS C 6226, 現JIS X 0208）の制定から四半世紀を経てはいるが，漢字コードの問題は今でも生きている問題である．1つには，日本語を用いるほとんどの人が，日々，漢字コードを用いていることによる．もう1つには，漢字コードの問題は国語と呼ばれる日本語の問題との強い関連から人々の関心を引きつけずにはおかないからである．

ここでは，「科学史」の問題としてではなく，我々が用いている文字コードの標準化がどのように行われ，今日，日常的に広く用いられているJIS漢字コードが最初どのようにして開発されたかについて開発思想を中心に述べよう．なお，JIS X 0208の技術内容に直接関連するものについては，JIS X 0208:1997の解説[2]に，Unicodeとして知られるJIS X 0221[3]，JIS X 0208:1997[4]およびJIS X 0213:2000[5]の開発・改正意図については，それぞれ参考文献を参照されたい．

情報処理学会漢字コード委員会

和田弘氏は，1971年の論文[6]の中で，「わが国ではカナ文字のためのextension法を必要とするので，早くからこの（ISO/TC97/SC 2, 現ISO/IEC JTC 1/SC 2）会議で発言しておりましたし，近くは漢字のために1 pageに代わって1冊とでもいうべき拡張法を提案して承認を得ております」と報告している．

この報告にもあるとおり，我が国における文字コードの標準化は，国際規格との整合性を基本とし，最初のISOコードであるISO R646の1967年の制定を受け，1年半後の1969年にカタカナを含む1バイトコードであるJIS C 6220（現JIS X 0201）を制定した．漢字コードについても，早くも1969年12月に情報処理学会規格委員会に，国立国語研究所所長をはじめ，文部省で戦後の国語政策の中心を担われた林大氏を主査とする漢字コード委員会を設けた[7]．すなわち，「国語」の担当者に漢字コードの検討を依頼した．

和田弘規格委員会委員長から漢字コード委員会に示された基本方針[7]を原表記に忠実に要約すると次のようになる．

(1) code extension procedureに準拠し，
(2) 2個の7単位符号で1つの文字あるいは記号を表し，
(3) 7単位符号の2～7列のうち7/15（DELETE）を除いた95 × 95 = 9,025の文字を表し，
(4) JIS C 6220（現JIS X 0201）で規定するローマ文字用符号およびカナ文字用符号を含むが，
(5) 現行の符号との関係は原則として考えない

と指示している．

また，漢字用制御符号については，今後の検討とし，escape sequenceについては仮に与えるとしている．

(6) 漢字コード用文字記号，
(7) Sub setおよびSub setとfull setとの関係，
(8) key boardとcollating sequence（並び順）

については，検討を依頼している．

この漢字のコード化の問題に対して，林大氏は，「体・躰・體を同一字の3体と見るか，3字各別のものと見るかが問題となる．……委員会では3主の字形それぞれに，すなわち3字として別々のコードを与えるという考えに傾いている．しかしやはり，……1点1画の差，……単なる筆法の差などについては，これらをそれぞれ別のものとして取り扱うのは，あまりにもむだな感じがする」

と記している．

また，林大氏は，「8,000 字を選ぶか，4,000 字，3,000 字を選ぶかでは，選び方がちがってくる」とし，「8,000 字の場合には，制限の色彩は非常に薄くなる．過去の文献を引用復刻するための文字が多くなろうが，8,000 字という範囲には，1,000 字なり 2,000 字なりの実用的な漢字セットは，どのような目的のものでも，みな含まれることになろう」と書く．

そして，林大氏は，漢字の配列順序，すなわち collating sequence に大きな紙幅を割く．

この第 1 期漢字コード委員会は，1971 年 10 月に 6,100 字からなる標準コード用漢字表（試案）を公表する[8]．この第 1 期委員会の活動報告は，それぞれ情報処理 1971 年 12 月号，1972 年 7 月号に掲載された．

この試案の選定方針は，(1) 日下部重太郎「現代国語思潮続編」（昭 8）（日下部表，6,478 字）の付録表にある漢字で B1 ～ 8 のいずれかに現れるものを採り，(2) 日下部表にあって B1 ～ 8 に現れないもの，日下部表になくて B1 ～ 8 および C1 ～ 4 のいずれかに現れるものの中で，委員 2 人以上の意見の合致したものを加え，(3) 全体として 6,000 字前後となることを目安とした．

また，異体字については，(4) 当用漢字で，新旧字体の著しく相違するものは，両体を採用し，(5) それ以外の異体字は，(1) および (2) の手続きによって拾われる範囲で採用した，とする．

試案で参照した B1 ～ 8 および C1 ～ 4 は，次の通り．なお，B1 ～ 8 については，後述の林大氏の保存の委員会資料に基づき，それぞれの漢字数を付すこととする．

B1 大西雅夫「日本基本漢字」（昭 16）3,000 字，

B2 新村出「広辞苑」第 2 版付録「通用漢字一覧」（昭 44）2,935 字，

B3 朝日新聞社「統一基準漢字書体表」（昭 32）4,000 字，

B4 全日本漢字配列協議会「常用漢字目録」（昭 43）4,000 字，

B5 日本活字鋳造株式会社「標準漢字目録」5,000 字，

B6 国会図書館用 NDL70 用コード表（昭 45）3,965 字，

B7 共同通信社「漢テレハンドブック」5,694 字，

B8 日経 FAM・M 型タイプ文字表 4,753 字．

C1 講談社国語辞典付録「漢字音訓総覧」（昭 41），

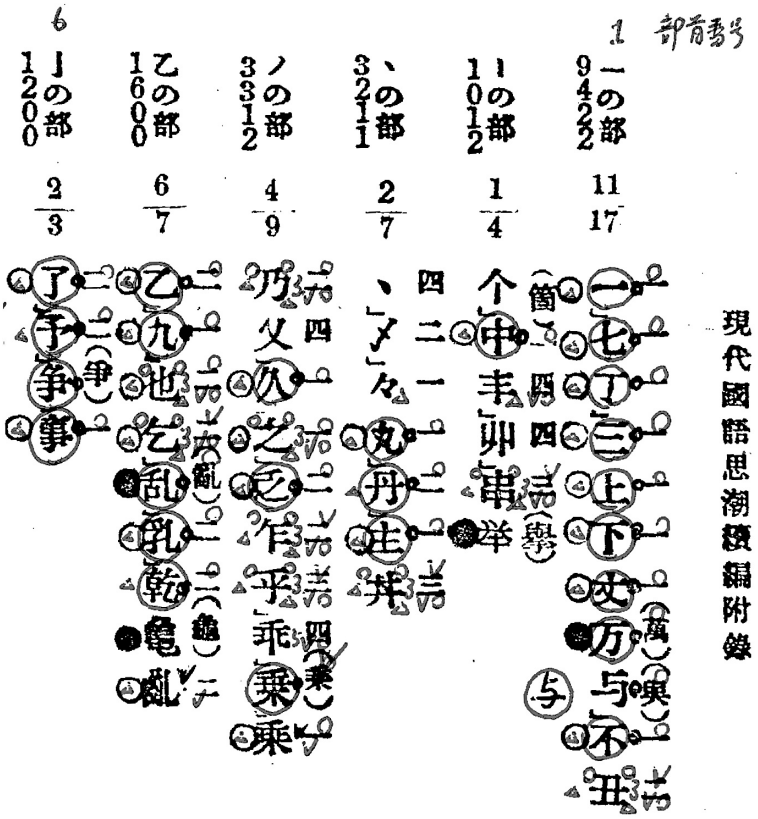


図 -1　漢字コード委員会資料

C2 野村広氏「4 万 5 千の姓氏に使われている文字の調査」（昭 44），

C3 国県郡市町村大字名および中学校名に用いられた漢字，

C4 国語研両調査に共通に現れた表外字．

このように当初，記号類を除き 8,000 字水準を想定して漢字の選定作業を行った委員会であるが，実際のデータから得られる情報から，大漢和辞典の漢字番号順に並べた（康熙字典順の）6,100 字の試案ができることとなった．

日下部表，B1 ～ 8 および C1 ～ 4 の参照文字表は，国語分野での常用字調査を基本とし，新聞，活版印刷，図書館，通信社など各分野での漢字利用を考慮に入れたほか，人名，地名にも配慮した漢字表となっていた．

この情報処理学会試案のもととなった資料を林大氏が保存されていたので，これを図 -1 に示す．このもとになっている漢字表が調査対象で最も文字数の多い日下部表であり，B1 ～ 8 の漢字表にどの漢字が現れているかをそれぞれの文字の周りに赤字で書き入れたものであ

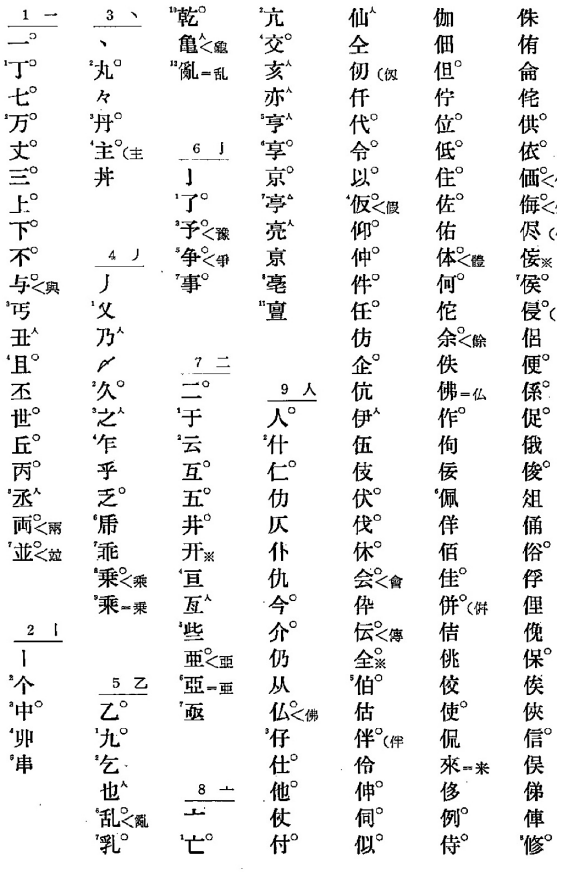


図-2　情報処理学会試案

る. 正にこうした検討から情報処理学会試案が生まれた.

図-2 に，情報処理学会標準コード用漢字表（試案）の最初のページを示す．漢字の横の○は当用漢字を，「人」は人名漢字を，△は当用漢字補正案で加える漢字を示す．また，くで対応する旧字体を，＝で対応する新字体を，（で別掲しない異体を示している．この異体関係の印を詳細に検討することによって，漢字コード委員会がどのような規準で異体字のコード化を行ったかを知ることができる.

このようにして漢字コードの漢字表部分については，成案を得た．しかし，和田弘氏によれば，当時の工業技術院は時期尚早であるとして，JIS化はすぐには実現されなかった.

第 1 期の委員会を受けて，1972 年度から和田弘氏を主査とする第 2 期委員会を設け，検討を継続する．この委員会の活動は，情報処理 1973 年 7 月号および 1974 年 6 月号に報告されている.

工業技術院はなぜ標準化を見送ったのであろうか，また，第 2 期委員会は何を検討したのであろうか.

ここからは，推測になるが，試案は当時実用化が可能と思われた漢字数を大幅に超えた漢字表を提案した．当時一部で用いられていた漢字タイプライタや漢字テレタイプなどのキーボードでは，たかだか 2,000 字程度を収容していたにすぎない．また，6,100 字の試案は，文字数が多いだけではなく，ほとんどの人には読みすら分からない漢字を多く含むことになる.

和田弘規格委員会委員長が開発した現在のJIS X 0201のカタカナ集合では，多くの利用可能な符号位置を残している．符号位置が空いているからといって，和田委員会は文字を与えることはしなかった.

一方，林委員会は，8,000 字を 6,000 字としたとはいっても，符号表のほとんどを埋めた．このことは，和田弘委員長の意図を超えたものではなかったのか？

カナ漢字変換などの技術開発とコンピュータの高性能化と低価格化とのおかげで携帯電話ですらカナ漢字入力が苦痛ではなくなった現在では想像もできないかもしれないが，こうした技術進歩以前の 1970 年初頭の状況では，工業標準として使える技術とするには，まだ検討が必要であった.

40cm角の 2,000 字の漢字タイプライタを 6,000 字に対応させるには，横幅として 120cm必要となる．また，デスクの上に置くタイプライタではなく，デスクそのものを入力装置としたとして，果たして読みも分からず見たこともない漢字を入力することがオフィスで可能となるのか．これが第 2 期委員会に与えられた課題である.

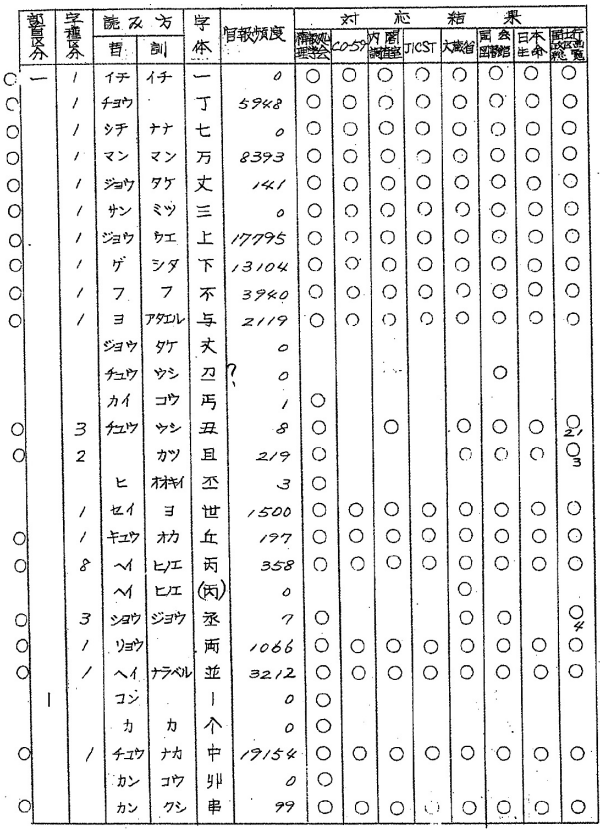
この課題のもと第 2 期委員会は漢字のパターン記述を中心とし，2,000 字ではなく，6,000 字の漢字を実用的に利用可能とする研究開発課題に挑んだ.

この情報処理学会の試案は，すぐにはJISにはならなかったが，その国際性への配慮，国内，特に，国語への配慮および学問的検討水準の高さからJISにつながる検討の基盤を与えたことは特筆に値しよう.

行政管理庁基本漢字

工業技術院が時期尚早と見送った漢字コードであるが，ニーズは待ってくれなかった．情報処理学会が試案を公表した直後に行政管理庁での検討が始まる.

行政管理庁は，行政情報処理の効率化のため，漢字コードを必要とし，少なくとも 1973 年度には，行政情報処理のための基本漢字の検討を始めた.



部首及外	字種区分	読み方 音	読み方 訓	字体	官報頻度	対応結果 情報処理学会	CCCII	内閣調査室	JICST	大蔵省	国会図書館	日本生命	国土行政区画総覧
	一	1	イチ	イチ	一	0	○	○	○	○	○	○	○
		1	チョウ		丁	5948	○	○	○	○	○	○	○	○
		1	シチ	ナナ	七	0	○	○	○	○	○	○	○	○
		1	マン	マン	万	8393	○	○	○	○	○	○	○	○
		1	ジョウ	タケ	丈	1441	○	○	○	○	○	○	○	○
		1	サン	ミツ	三	0	○	○	○	○	○	○	○	○
		1	ジョウ	カエ	上	17795	○	○	○	○	○	○	○	○
		1	ゲ	シタ	下	13104	○	○	○	○	○	○	○	○
		1	フ	フ	不	3240	○	○	○	○	○	○	○	○
		1	ヨ	アタエル	与	2119	○	○	○	○	○	○	○	○
			ジョウ	タケ	丈	0								
			チュウ	ウシ	丑	0						○		
			カイ	コウ	丐	1	○							
		3	チュウ	ウシ	丑	8	○		○		○	○	○	○2
		2		カツ	且	219	○				○	○	○	○3
			ヒ	オオイ	丕	3	○							
		1	セイ	ヨ	世	1500	○	○	○	○	○	○	○	○
		1	キュウ	オカ	丘	197	○	○	○	○	○	○	○	○
		8	ヘイ	ヒノエ	丙	358	○	○	○	○	○	○	○	○
			ヘイ	ヒノエ	(丙)	0					○			
		3	ショウ	ジョウ	丞	7	○				○	○		○4
		1	リョウ		両	1066	○	○	○	○	○	○	○	○
		1	ヘイ	ナラベル	並	3212	○	○	○	○	○	○	○	○
	丨		コシ		丨	0	○							
			カ	カ	个	0	○							
		1	チュウ	ナカ	中	179154	○	○	○	○	○	○	○	○
			カン	コウ	丱	0	○							
			カン	クシ	串	99	○	○	○	○	○	○	○	○

図-3　対応分析結果

この検討では，情報処理学会試案を基本としたほか，1972年の官報の偶数日の使用頻度と7つの漢字表および調査結果を加え，対応分析[9]を行い，2,817字の行政情報処理用基本漢字[10]を選定した．

対応分析結果を図-3に示す．

この行政管理庁の調査研究は，偶数日に限定されてはいるが，官報での漢字使用頻度を1年分にわたって調査したほか，行政で必要とする地名データを国土行政区画総覧から使用漢字一覧としてまとめた．また，日本生命収容の人名漢字をその調査に含めた．

官報頻度の調査は，漢字コード選定のために，漢字表を参照するだけではなく，実際の漢字利用を調査した1次資料を独自に調査したことは高く評価できる．国語学分野では，国立国語研究所の100万字調査が有名であるが，日本の近代活版のもととなったWiliam Gambleも100万字調査を行っており，戦後も大蔵省印刷局や毎日，朝日などの新聞社による活字頻度調査がある．こうした実際の頻度調査によって，工業としての活版印刷の基盤が形成され，工業規格としての漢字コードの基礎資料にもなった．国語学などの「学」ではなく，産業の基礎として，政府調達仕様の基盤としてのJISには，このような調査を持つことは必要不可欠であった．

行政管理庁の調査研究のもう1つの特徴は，地名・人名への配慮である．地名に関しては，最も充実した行政地名の集成である国土行政区画総覧で使用されている漢字を独自に調査し，国土行政区画総覧使用漢字として，その対応分析に納めた．また，人名に関しては，漢字情報処理を先行して行っていた生命保険会社の中から最大の生命保険会社である日本生命の収容漢字表を採用した．後述するようにIBM社の文字コードであるEBCDIC漢字コードも，ある保険会社の人名漢字をもとにしている．この点を考慮すると，NTTの電話帳がデータベース化される1980年代以前に得られる最良の人名漢字データをJISの基礎に据えることができたことになる．

この行政管理庁の動きに触発され，JIS原案委員会が設けられることになる．

JIS C 6226[13]（現JIS X 0208）の制定

1974年に森口繁一氏を委員長とし，3年計画でJIS漢字コードの開発を目指した漢字符号標準化調査研究委員会が日本情報処理開発センターに設けられることになる．

この調査研究では，対応分析結果に加えて，新たに12の漢字表を追加し，20の漢字表の漢字を新字源第63版をもとに整理した[11]．図-4に漢字整理番号表を示す．最終的には，37の漢字表の漢字を整理し，漢字集合の検討を行った．

この委員会によって78JISと呼ばれる最初のJIS漢字コードが開発された[12]．

このJIS原案委員会での検討でも，林大氏を中心に西村恕彦氏らが中心となって検討が行われた．

この検討では，これまでの漢字集合だけの検討ではなく，非漢字集合を含め，文字集合を確定させ，符号化文字集合を完成させたことをまず特筆すべきであろう．また，漢字集合に関しては，“現代日本社会で通用する漢字全体の集合約1万弱”とし，“基本度の高いと評価されている漢字集合約3,000”を第1水準，さまざまな専門分野で用いられる漢字を第2水準と，漢字集合を2つの水準に分け，符号化する漢字の総数を94 × 94 = 8,836中の7,000 ± αと想定し，漢字集合の確定を目指した．

詳細は，JIS X 0208の解説[2]を参照されたいが，頻度調査とクラスタ解析によって，第1，第2水準を決定したとしているが，実質的には，情報処理学会試案，行政

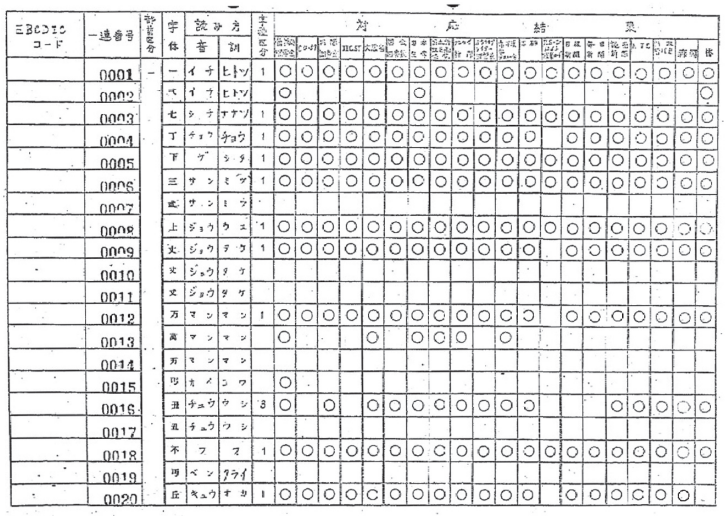


図 -4　漢字整理番号表

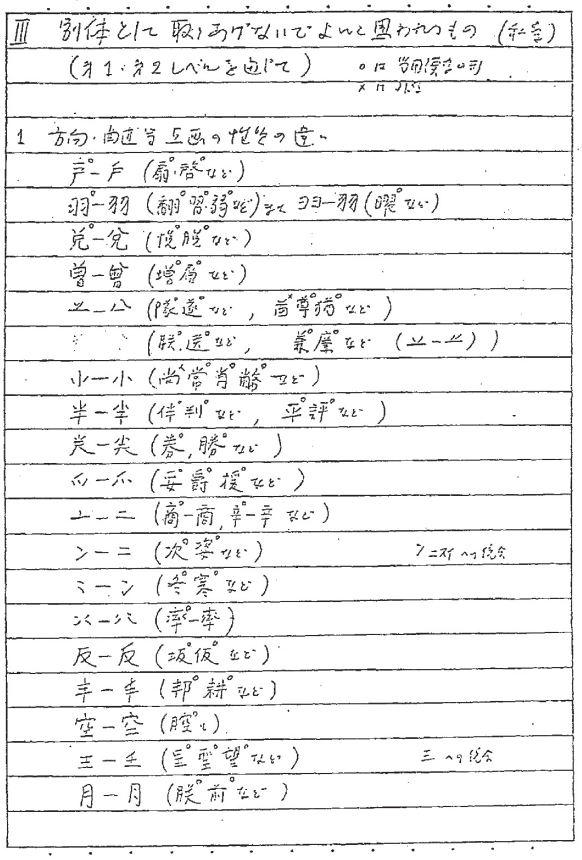


図 -5　異体字の取り扱い資料

基本漢字，日本生命人名漢字，国土行政区画総覧使用漢字の 4 つの漢字表から漢字を選定し，第 1 水準には当用漢字および当用漢字補正案を基本に，地名コードで用いられている地名漢字，人名漢字および一般の漢字表での頻度の多い漢字を収容した．この 2 つの水準分けは，規格には規定していない第 1 水準だけの実装を実質上可能とし，このことによって，低価格のワープロやパソコンでの漢字情報処理を可能とした．

並び順については，基本漢字に当たる第 1 水準については代表音訓順とし，第 2 水準については，新字源順とした．このことによって，ほとんどの場合，漢字の並び順は，ほぼ五十音順となり，第二水準漢字の検索に関しては，大漢和順ではなく，新字源順としたことによって，異体字が隣接することになって，検索の便を大きく改善した．同時に，漢字の部分字体をもとにしたパターン記述をもとに独自の林大方式ともいうべき「部首画数順」の索引を付けることによって，6,355 字という大規模な漢字表を実用に耐えるようにした．

また，異体字の取り扱いについて，詳細な検討を行ったことも特筆に値する．**図-5** に林大氏により“異体字の取り扱いについての資料”の一部を掲げる．残念ながら 78JIS の解説には，この異体字の取り扱いに関しては，簡単に触れられただけで，詳細で厳密な検討の実際は，広く知られることがなかった．

符号化に関しては，初年度の報告書では，JIS X 0201 に規定するローマ字集合およびカナ集合を最初の 2 つの 94 文字のページに収容することも検討したようであるが，日本語としての並び順を優先し，JIS X 0201 との関係は明確にはならなかった．このことが全角・半角問題を生んだと思われる．符号化に関して，78JIS では ISO/IEC 4873 に規定される ISO/IEC 2022 の標準的な使い方ではなく，エスケープシーケンスを規定するだけに留めたことから符号化に関する混乱した状態を十分には収拾することはできなかった．

この 78JIS は，それ以前の専任のオペレータが特殊漢字処理システムを用いて入力する時代が終わり，パソコンとカナ漢字変換の登場によって，誰でもが日本語を日常的に利用する時代の始まりにちょうど間に合うことによって，日本語情報処理の普及を加速し，混乱を押さえることには大きく貢献することができたことも歴史的事実である．

EBCDIC漢字コード，シフトJIS, EUCと漢字コードの混乱

当時，世界のコンピュータのシェア60%以上を占める巨人IBMは，文字コードでもJIS/ISO/ASCIIに対して，EBCDICという独自の文字コードの世界を築いていた．日本語の世界においても，このEBCDICに独自にカナ拡張および漢字拡張を行い，ある生命保険会社の頻度情報をもとに，漢字集合とその並び順を決めたEBCDIC漢字コードを開発した．主要国産各社も英数字に関しては，IBMのEBCDICに合わせるとともに，アーキテクチャもIBMに合わせ，Mシリーズコンピュータの開発を行っていた．カナおよび漢字に関しても，日本IBMから日立などの国産各社に対して，EBCDIC漢字コードの統一が打診されたが，この統一は実現されなかった．結果として，国内ではEBCDICとはいっても，各社独自の漢字拡張が行われることになる．

ISO/IEC 4873で規定する8ビット符号の構造と規則など基本的なISO/IEC 2022の用い方は，エスケープシーケンスを多用するのではない．ISO/IEC 2022の適合性の箇条でも，実質的には，最小のエスケープシーケンスを推奨している．すると，78JISの規格票に規定するエスケープシーケンスを最小化したより効率的で自然な符号化が出現するのは，当然である．この代表的なEUCと呼ばれる文字コードである．

ISO/IEC 10646（Unicode）では，当初7ビットの通信路で通信可能なUTF-7を規定していたが，このUTF-7では，ASCIIがそのままでは利用できない．このため，Unixを含め，多くの処理系での国際化のコストが大きくなるという問題が生じた．このため，ISOでは，UTF-7の代わりに当初FSS-UTF（File System Safe UTF）と呼ばれたUTF-8をUTF-7に代わって，規定した．

日本でも同様に，広く利用されているJISカナを含む8ビットコードの空き領域を用いて漢字を利用可能としたいとの自然な要求が生まれ，シフトJISがパソコンとほぼ同時期に誕生した．

すなわち，符号化方式としては，JIS，各社のEBCDIC拡張，EUC，シフトJISがネットワーク上に併存することとなった．

一方，1983年にJIS漢字コードが改正され，83JISと呼ばれる規格が生まれた．83JISでは，常用漢字表の訓令・国字への対応を主とし，(1)非漢字の拡張，(2)漢字の拡張，(3)第1水準と第2水準の一部漢字の入れ替え，(4)常用漢字などの政令漢字についての字形の変更，(5)字形の変更を行った．

83JISへの対応は，この5つのどれに対応するかが各社別々のものとなった．たとえば，日立は83JIS全部を受け入れ，NECは（78年に出された正誤表へは対応しない）78JIS初刷を基本とし，(4)の政令字形への変更だけを行った．また，NECが83JISである日立製のプリンタをビジネス用に出したことによって，新旧JIS問題が発生する．

DOS/Vのもととなり，その後のマイクロソフト漢字コードの基本となったOADGの漢字コードを決めたIBMのシフトJISでは，(1)および(2)だけを採用することとなった．

謝辞　最初のJIS漢字コードの委員長であり，1997年の最新の改正に関しても詳細に読まれ，修正意見をお寄せいただいた東京大学名誉教授森口繁一先生が2002年10月に亡くなられました．森口繁一先生の逝去を悼むとともに，森口繁一先生がJIS漢字コードを含め，日本の情報処理技術に多大な貢献をされたことに謝辞を捧げます．

参考文献

1) 和田　弘，高橋　茂：コード会のコードについて，情報処理，Vol.1, No.2, p.107-109, 情報処理学会（1960）.
2) 芝野耕司編著：増補改訂 JIS 漢字字典，日本規格協会（2002）.
3) 芝野耕司：JIS X 0221（ISO/IEC 10646）の目指すもの－文字コードと日本の国際対応－，情報処理学会情報規格調査会 NEWSLETTER, No.40（Dec. 1998）, http://www.itscj.ipsj.or.jp/topics/tp40.html
4) 芝野耕司：JIS X 0208:1997“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”の改正，情報処理学会情報規格調査会 NEWSLETTER, No.35（Sep. 1997）, http://www.itscj.ipsj.or.jp/topics/tp35.html
5) 芝野耕司：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）の制定，標準化ジャーナル，Vol.30, pp.3-7, 日本規格協会（Mar. 2000）.
6) 和田　弘：情報処理と標準化，情報処理，Vol.12, No.2, pp.63-67, 情報処理学会（Feb. 1971）.
7) 情報処理学会企画委員会編：電子計算機の国際標準化　ISOの動きとわが国の歩み，情報処理学会（1971）.
8) 情報処理学会漢字コード委員会：標準コード用漢字表（試案），情報処理学会（Oct. 1971）.
9) 行政管理庁：行政情報処理用標準漢字選定のための漢字の使用頻度および対応分析結果，行政管理庁（Mar. 1973）.
10) 行政管理庁：行政情報処理用基本漢字に対する符号付与に関する調査研究報告書〔付表〕，行政管理庁管理局（Mar. 1975）.
11) 日本情報処理開発センター：漢字整理番号表，日本情報処理開発センター（Mar. 1975）.
12) 日本情報処理開発センター：情報交換のための漢字符号の標準化に関する調査研究報告書，日本情報処理開発センター（Mar. 1975, Mar. 1976）.
13) JIS C 6226-1978,“情報交換用漢字符号”，日本規格協会（1978）.

（平成14年10月28日受付）