漢字の機械への入力*

斉藤勝久** 永井悦夫** 山田興三** 大迫昭和**

Abstract

In the field of recent information retrieval, the remarkable trend has been viewed the computer becomes to treat Japanese language in processing. One of the most important problem in this Japanese information processing should be how to put the Kanji characters as the input to a machine. This problem is due to the composition of Japanese language which applies a great number of ideographic Kanji characters and due to the fact that there is no simple and practical machine like an English typewriter.

In this paper the practical method of Kanji character encoding and pattern input method is described, where Kanji Display System is applied. On the encoding, the practical method by pronunciation is described and further the possibility of the other practical system where all Kanji characters can be applied is described.

1. まえがき

西洋文明と東洋文明の大きなちがいである文字は，現在の計算機時代においても，人間と機械の対話の手段として著しい差異がある．欧米のアルファベット文字は字種が少なく，鍵盤を用いることにより，タイプライタにみられるようなきわめて簡単な入力方式で，機械に自国語を入力することができ，機械の内部の言葉としても自国語を使用できる．わが国では，計算機開発の当初から機械とのインタフェースに欧米語を用いており，多種類の漢字を含む日本語の文字を取り扱うことはなかった．しかし，計算機技術の進歩により最近では，日本語で表現されている情報の検索，日本文の多量な処理などの目的で計算機を導入することが行なわれるようになった．このような漢字を含む日本語で表わされる情報の処理システムにおける第1の問題は，数千種類にのぼる漢字を容易に速く機械に入力する手段である．本文では，このような漢字の実用的入力手段として，漢字ディスプレイを用いた方式について述べる．漢字の入力はコード化とパターンの2つについて述べる．

2. 漢字のコード化

日本語は，表音文字を使用している欧米語と違って表意文字であって，その種類が多く，コード化のための文字選択が重要な課題である[1),2),3)]．漢字のコード化として考えられるものは，大きくわけて次の3通あり，第1は漢字を形としてとらえる方法，第2は漢字を発音で選択する方法，第3は英数字の組合せで選択する方法であるが，従来検討されてきたものは，ほとんど第1の方法である．これには自然の形そのものでとらえる方法と，漢字を構成している要素に分解してとらえる方法である．前者は鍵盤で単純な操作ではあるが，文字盤が大きく選択が大変である．後者は選択要素は少なくなるが，その組合せが複雑になる．この中には，構成要素にコードを与えて，その組合せで1漢字を表わす方法もあるが，自然言語で表現された情報を機械に入力するためコード化するとき，その言語を構成している1字に1つのコードを与えるのが，最も自然な方法である．このように漢字のコード化については，いろいろ検討されているが，一長一短で決定的ではなく，漢字のもつ複雑さを物語っている．

本文では，この漢字コード化の実用的手段として，漢字を発音により分類して，漢字ディスプレイを用いて選択する方法について述べる．

漢字を発音で選択するという問題は，日本語では同

---

* Kanji Character Input to Machine by Katsuhisa Saito, Etuo Nagai, Kozo Yamada and Akikazu Osako (Oki Electric industry Co. Ltd.)

** 沖電気工業株式会社

第1表　同音異字の種類

同音異字数	音読みの数	訓読みの数	音訓併用	同音異字数	音読みの数	訓読みの数	音訓併用
1	56	514	546	21	4	1	3
2	47	81	118	22	1		1
3	31	24	54	23	1		1
4	20	15	27	34	4		4
5	22	9	19	25	0		1
6	12	5	15	26	0		0
7	16	2	22	27	1		2
8	12	2	15	28	2		2
9	12	2	13	29	1		1
10	2	5	4	30	2		4
11	9	2	9	31	1		1
12	6	1	6	35	1		0
13	4	1	8	36	1		1
14	4	0	4	43	1		2
15	1	2	3	46	1		0
16	0	0	3	51	0		1
17	0	0	1	52	0		1
18	1	1	0	61	2		2
19	4	0	2				
20	4	0	5	合計	286	667	901

音異字が多いということで従来行き止っていた．この方法は，簡単なカナ鍵盤あるいはローマ字を用いれば英字鍵盤ですみ，日本人であれば筆順あるいは要素による漢字の分解よりも常識になっており，操作が単純で手数が少なくてすむ利点がある．漢字の読み方は，わが国では音読みと訓読みがあり，音読みの場合は一義に定まらない字がほとんどであり，訓読みでも30%以下で，問題はこの同音異字の処理である．この一義的に定まらない文字を漢字ディスプレイに表示し，この中から，目的の文字を選択する方法について検討する．同音異字を当用漢字1,850字と補正字28字について調べてみると，**第1表**のようになる．ただし，2つ以上の読み方のある字は重複している（角川国語辞典による）．音訓併用の項は，たとえば，アと発音する漢字は音では

“ア” 亜

だけであるが，訓では

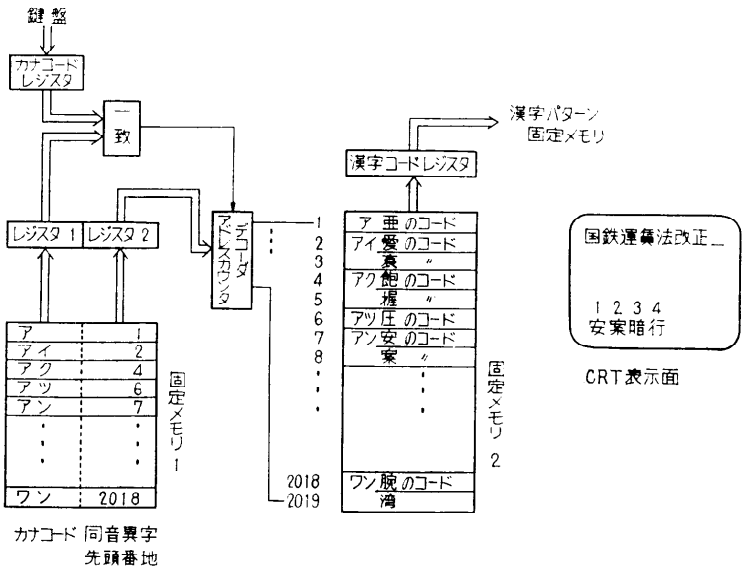
“あ” 合，会，飽，明，上，揚……

とありこの両者を合わせた数である．

音読みだけの場合は，286の発音種類になるが，同音異字が多く60字程度になるものがあり，選択する字数が多くなる．訓読みでは逆に発音種類は多いが，一義的に定まる字も多く同音異字は少なくなる．しかし，訓読みのない字は全体の40%もあり，また実際にこの方法を使用する場合，音だけ，訓だけでは不都合なので，音訓いずれの発音も許すとすると，発音種類は900以上になり，同音異字も増し選択する字数は多くなる．また，漢字のカナ表現は，**第2表**のようになり，4字以上は30字程度で訓読みの特殊なものであるから，これを無視すれば3字で表現できる．この発音で，漢字を選択するシステムの1例を**第1図**に示す．固定メモリ2は，ディスプレイが収容している全漢字のコードを発音順に記憶しているもので，同音異字のコードは連続したアドレスにはいっている．固定メモリ1は，発音に対応したカナコードとその発音に対応する同音異字のコードがはいっている固定メモリの先頭番地が，対になって記憶されている．選択したい漢字をカナ鍵盤で発音に従って打鍵すると，そのコードはカナコードレジスタにセットされ，固定メモリ

第2表　漢字発音のカナ表現

カナ数	漢字数
1	59
2	620
3	195
4	28
5	3



第1図　発音による漢字デコーダ
Kanji Decoder by Pronunciation

1を順次読み出してレジスタ1と一致をとる．一致したときのレジスタ2の内容を固定メモリ2のアドレスカウンタにセットし，指定された発音の漢字をすべてCRT に表示する．同音異字がない場合はそのまま一義的に決まる．たとえば，第1図の CRT 表示面においては，"案"なる文字を入力したい．カナ鍵盤でアンと打鍵するとアンと発音する4種類の漢字が CRT 上の下部に表示される．この中から目的の漢字 "案" を選択する．

このデコーダ部の固定メモリは

音だけの場合	固定メモリ1	25ビット	300 語
	固定メモリ2	12ビット	2 K語
音訓併用の場合	固定メモリ1	25ビット	1 K語
	固定メモリ2	12ビット	3 K語

となる．この漢字コード化方式の金物としては，上記の固定メモリよりなるデコード部以外に，ディスプレイされた同音異字の選択手法の問題がある．入力をカナ鍵盤で行なうものとすると，ディスプレイされた文字の選択をライトペンで行なうことは不自然で，むしろ対応する番号を入力する方がよい．同音異字の選択を重視すれば，ライトペンの方が好ましい．この場合には鍵盤をすべてなくして，ディスプレイの一部をカナ鍵盤のかわりに用い，CRT 上だけで漢字を入力する方が操作が容易である．漢字入力用の CRT を情報表示用の CRT とわけて，入力用 CRT は平面上におくことも考えられる．このようにこの方式の実用化には固定メモリのコストと同音異字の選択法の2つが重要であるが，固定メモリについては，技術の進歩によりコスト低下が見込まれるので，選択法が第1の問題になる．選択法に関しては，さらに選択字数の問題がある．選択の比較的容易な同音異字 20 字までの占める割合は音だけの場合 65%，音訓併用の場合 72% で，20 字以上のものは特殊な取扱いをして，字を使用頻度順に並べて何行かを使って表示するか，何画面かにわけて表示する．また，別の方法として，漢字の形からの特長，たとえば，きわめて常識的で代表的な部首約 20 種を併用すると，第 1 表において同音異字 61 字である "コウ" と "ショウ" は 25 字前後に減らすことができる．このように同音異字の多いものについては，さらに部首情報を加えることにより選択操作は簡単になる．あるユーザで用いられている漢字テレタイプ用文字 859 字について調べた結果によると，音読みだけで同音異字 20 字をこえるものはわずか1種類（同音異字21字）で，あとはすべて 20 字以下であり，10 字以下のもので全体の 70% 以上をカバーしている．このような場合には，充分実用性があることがわかる．将来カナ単音の音声入力が可能になれば，この方式による漢字入力手段は，CRT ディスプレイの高速表示と併用することにより，充分実用性が出てくることが期待できる．

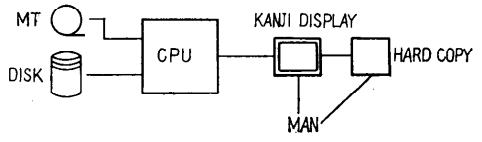
3. 漢字パターンの機械への入力

漢字の種類は約 30,000 字あり，現在わが国で使用されている字は 8,000 字といわれている．最近の新聞に関する文字調査によれば，2,500 種の文字により表現されない量は全体の 0.003% 以下で，使用される文字が限定される傾向は，今後ますます強まるものと思われる．このように 2,500 種程度の文字で，日本語の情報はほとんど取り扱えるので，現在実用されている漢字を取り扱うシステムでは，入出力機器で取り扱う字種を限定している．

しかしながら，いかに実用上さしつかえなくても，使用頻度の少ない漢字鍵盤にない字を表現したいという要求は，日本文を取り扱うときどうしても生ずる．このような使用頻度の低い字は，使用分野・地名・人名などに特に関連しており，時代・季節によって変化し固定して用意する必要は少ない．入出力機器の規格を統一し，できるだけ簡単化しコストを下げ，すべての漢字を取り扱うことができるようにするため，使用頻度の高い第1種文字（たとえば 2,500 字）とその他の第2種文字を別な取り扱いをする漢字ディスプレイシステムについて，この項では検討する．

3.1 漢字パターン入力システム

ここで考える漢字処理システムは，**第2図**のような構成になっている．前述のような観点に立って，この漢字ディスプレイは使用頻度の高い第1種文字については，大容量のパターン固定メモリをもち，使用頻度の低い第2種文字については数字分のパターン一時メモリをもっている．パターン一時メモリは IC シフトレジスタなどを用い，その漢字ディスプレイの文字発生方法に従って文字のパターンを任意に書き換えることができるものである．この数字分のパターン一時メ



第2図　漢字ディスプレイシステム
Kanji Display System

モリを用いる意味は，使用頻度の低い文字についてはディスプレイ1画面で生ずる字数は数字程度で充分であるということと，必要に応じて文字のパターンを書き替えるためである．漢字ディスプレイを用いてパターンを入力するには，通常の使い方で表示したい文字を記憶しておくリフレッシュメモリを利用するのが簡単で便利である．このシステムに漢字を追加しようとするとき，人はディスプレイ面に追加したい文字のパターンとそれに対応するコードを入力する．入力された内容は，ディスプレイとハードコピーだけで使用するときは，単に，パターン一時メモリに入れて使用する．また，この数字のパターン一時メモリに固有のキーを用意しておき，その内容を CRT 上の一部に表示して使用したい文字を確認できるようにすれば，このキーを用いることにより追加された文字は，あらかじめ収容されている第1種文字と同様に取り扱うことができる．システム全体で取り扱う字を追加する場合は入力されたパターンとコードは計算機に送り，ディスクメモリに文字パターンを追加する．ディスプレイにない漢字は，このような手段でディスクメモリに入れられているので，計算機からディスプレイする情報に第2種文字が含まれているときは，その第2種文字のコードとパターンをディスプレイに送り，追加文字コードレジスタとパターン一時メモリに入れて表示を可能にする．計算機が第1種文字と第2種文字の区別を容易につけられるように，漢字コードの1ビットをこの判定に用いる．

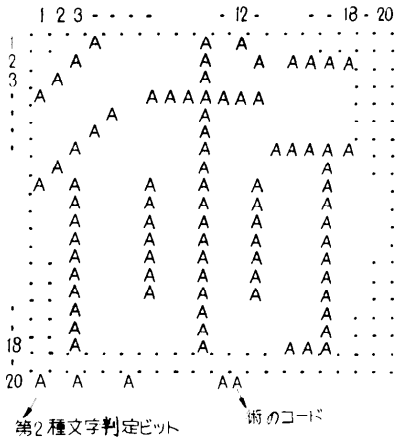
このシステムを使用すれば，同じコード体系を使用している日本文の情報を処理する場合，どうしても必要な漢字，たとえば，人名・地名などは，必要に応じて漢字ディスプレイで編集し直せば，最終データとしてのハードコピーは望みのものが得られる．

3.2 ドット式漢字ディスプレイを用いる文字パターン入力

前項で述べたようなシステムでは，文字パターンの入力は次のような事柄が重要である．

①手法が簡単であること．②入力しているパターンを CRT で見て確認できること．③この実現のための金物が簡単であること．

以上の項目を考慮して，ドット式漢字ディスプレイの場合には，リフレッシュメモリを利用して，通常の CRT 表示の1文字を入力パターンの1ドットに対応させる方法が最も簡単で有用である．たとえば，1文字のドットパターンは 18×18 ドットで表現されており



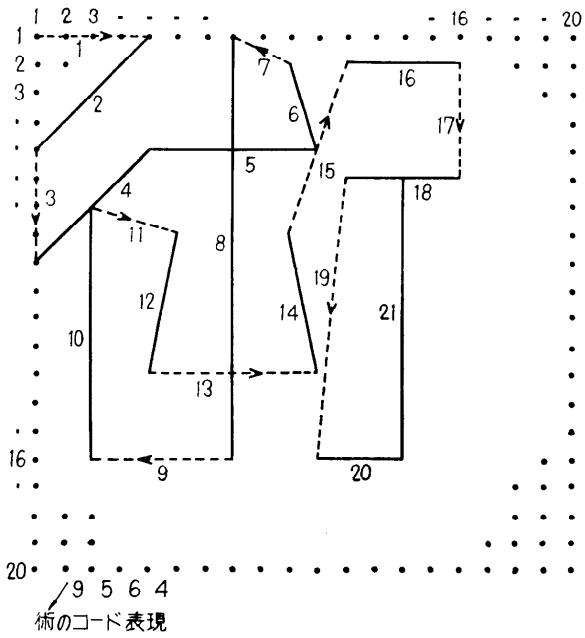
第3図　ドット文字パターンインプット
Input of Dot Pattern

1画面は20字20行で表示されているとする．パターン入力は任意の文字，たとえばAの文字を，文字パターンの輝点の部分に，通常文字を入力するのと全く同じようにして**第3図**の例のように入れる．デザイン上の類似性を出したい場合は，文字Aの代わりに●印のパターンを1字として，鍵盤に用意しておけばよい．コードは第2種文字を英字か数字の組合せで入力するか，第3図のようにマークにAを対応させてコードを入力する．このようにリフレッシュメモリに入れられたパターンは，順次読み出されて，デコード回路を通して文字パターン一時メモリに入られる．計算機とのパターンの受け渡しは，この文字パターン一時メモリとの間で行なわれる．

3.3 ベクトル式漢字ディスプレイを用いる文字パターン入力

ベクトル式の場合，通常の文字表示と文字パターンの対応がドットの場合ほど類似性がないので，パターンを入力しながら文字を CRT 上に表示させるのが多少複雑であるが，やはりリフレッシュメモリを用いるのが簡単である．ベクトルによる漢字表示には，いくつかの方法があるが，1字を記憶している容量が少なくてすむので，コネクテッドベクトル方式を採用する．ドットの場合と同様に，文字をインプットするときのアドレスに対応して，CRT 上に 20×20 メッシュを表示する．文字パターンの入力は一筆書きで，現在の輝点の位置から次のストロークの端を順次指定してゆく（**第4図**参照）．

指定の方法には“たて”，“よこ”の番号を打鍵する方法とライトペンによる方法がある．ライトペンを用

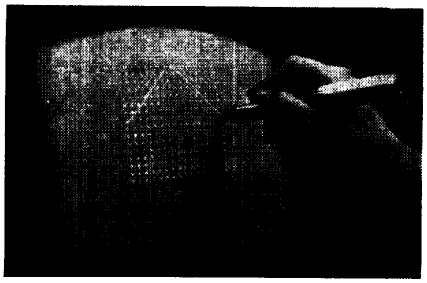


第4図 ストローク文字パターンインプット
Input of Stroke Pattern

いるときは，指定したベクトルの入力は，そのときのリフレッシュメモリのアドレスをパターン一時メモリに入力してゆけばよい．また，このときそのストロークを光らせるか否かも指定する．そのパターンをCRT 上に表示するには，通常の文字表示間隔を1ステップにして描いてゆけばよい（**第5図**参照）．

5. むすび

漢字の機械への実用的入力方法として，漢字ディスプレイを利用する漢字のコード化とパターンの入力の2つについて述べた．コード化については漢字の複雑さから，その構成は発音でも形でも表音文字のようなシンプルな合理性がなく，機械だけでその操作を単純化することは困難であるという観点に立って，ディスプレイを用いる方法を検討した．日常の漢字の使用において，常識になっている発音と少数の部首により入力したい漢字のはいっている集合をできるだけ小さくして，この中から人が選択する方法の実用性について述べた．



第5図 ディスプレイによる文字パターンインプット
Character Pattern Input by Display

パターンのインプットについては，このようなシステムがどの程度有用性があるか，システムとしての評価が必要である．この結果いかんによっては，今後漢字に関する各種の標準化の問題に対して，使用頻度による第1種文字と第2種文字という取り上げ方が有益であることが期待できる．

謝辞 本文は通産省工業技術院の大型プロジェクトの一環として行なわれた漢字ディスプレイの開発研究をもとにしてまとめたものである．

日ごろご指導いただいている電気試験所野田部長，西野室長，当社研究所林原室長に深謝する．

参考文献

1) G. L. Walker, S. Kuno, et al.: Chinese Mathematical Text Analysis, IEEE Trans. on EWS, Vol. EWS-11, No. 2 (1968), p. 118～128.
2) 坂井，長尾，寺井：部分パターンによる漢字の記述，信学会オートマトン研究会資料，1961-01.
3) H. Hayashi, S. Kuno, et al.: Graphical Input/Output of Nonstandard Characters, Communication of the ACM, Vol. 11, No. 9 (1968), p. 613～618.

（昭和44年4月8日受付）