漢字の入出力処理について

坂本義行　（電子技術総合研究所）

1. はじめに

日本語テキスト（漢字かな混り文）を計算機へ入力する方法については、すでに種々の方式が開発されている。ここで紹介するのは、対象となる利用者として研究者（素人）自身が小論文（数千文字から成る）等のテキストを直接入力する方法についてである。すなわち、大型計算機に接続されているＴＳＳ画像端末から会話形式により、簡単な規則で確実に打鍵入力が可能な方法である。

打鍵は、画像端末の英数字鍵盤によりローマ字記述で行なうが、字種の識別を行なうため、大小シフト・キー等数種の識別符号を挿入する。ローマ字かな変換処理では、種々の表記法を可能な限り許すこととし、例外的なものについては、パラメータの変更により表記する方法が用いられている。かな漢字変換処理に関しては、対象となるテキストの特性から、汎用的な辞書として恒久的に利用する固定辞書とテキストあるいは利用者個有の一時的に利用する浮動辞書の２種類を設けている。この辞書の検索方法は、会話形式により、同音異字については、画面に表示し、選択追加処理を行なう。同時に、「語音」の処理が可能である。

なお、当システムで作成されたテキストの編集処理は、独立に開発された漢字テキスト・エディタにより行なうことができる。さらに、テキストの横組編集出力を行なう漢字RUNOFFシステムを開発中である。

プログラムは、処理の簡便さからTSS FORTRANで記述されている。辞書は、Tree構造で表現されており、固定辞書の蓄積データとして、当用漢字音訓表の「字音」とその漢字符号、が用いられている。漢字符号は、汎用性の面から整数10進4桁符号が割り当てられ、文字パターンは、大型計算機のディスク上にベクトル方式で、online ランダム・ファイルとして約3,000種の文字が記憶されている。

2. 方法

2.1 入力の諸方式

漢字のかな混りテキストを計算機へ入力する方法としては、

（1）漢字コードを直接打鍵入力
（2）かな（ローマ字を含む）漢字変換
（3）漢字認識（OCR）

の三種に大別される。

（1）の方式は、漢テレ等を用いるフルキー方式と呼ばれるもので、数千種の鍵盤から選択し打鍵するため、かなりの訓練を要する点で操作性に問題がある。

（2）の方式は、（1）に比較し操作性では優れているが、変換を自動化するには、ぼう大な辞書と複雑な手続きを必要とし、現在までのところ完全な変換結果は得られていない。

（3）の方式は、試験的な段階であり、英数字、かな、一部の漢字や特殊字形については、処理システムが開発されているが、実用という面からは将来の問

題であろう。

以上いづれの方式にも特徴を有しており，便用面から有用性が考えられる。

テキストの性質と容量によって，つぎの2種類について上記の（1），（2）を検討すると，

（a）大量のデータを熟練者に委託する。

（b）小量のデータを研究者（素人）が入力する。

（a）に対しては，（1）の方法を用いると，off lineで紙テープ等の媒体上に漢字符号を直接入力できる。この方式は，熟練者では鍵盤検索操作が安定し，打鍵速度が平均50字/分程度となる。漢字への変化処理による誤差を発生しない点，大量のデータの一括入力処理に適している。打鍵誤りの発生率は，30万字の実験で，0.01%（4ヶ所），その内容は，第1表のような傾向を示し，二回の修正処理により完全なデータが得られた。

第1表　打鍵誤りの内容

誤りの内訳	件数	比率(%)
欠落　文字	90	21.7
〃　機能記号	49	11.8
過剰打鍵	41	9.9
原文不鮮明によるもの	63	15.2
隣接キーを誤り打鍵	76	18.3
類似語	13	3.1
同音異義語	5	1.2
外字機能記号誤り打鍵	5	1.2
文字種誤り（ひらがなとカタカナ）	4	0.9
〃　（漢字とかな）	3	0.7
その他	65	15.7
計	414	

（b）の方法では，素人という点，操作が容易で訓練を要しない点がもっとも大きな問題となる。すなわち，（2）の方式で同音異字の処理を人間の判断に委ね，正しい漢字を選択する。また視認性をよくするため，画像端末上に漢字パターンを表示する方法をとる。

2.2　ローマ字表記とカナ変換

汎用性を目的としているため，特殊なカナ鍵盤を用いず通常の英数字鍵盤を有する画像端末からの入力を考えた。これは打鍵速度の点で，文字数の増加分は，鍵盤の操作性，訓練の必要性という面からカナによる入力と差がないとおもわれる。

ローマ字表記には，ヘボン，訓令，日本といった諸式があるが，外来語，外国の人名，地名に関しては，表記が一定していない。ここでは可能な限り種々の表記が行なえる「ゆるやか」な規則とし，表記が重なるものについては，スイッチを設け，パラメータによる選択方法をとっている。漢字，カタカナ，ひらがなによる表記の異なりを第2表のように5個の仮引数 BOIN, SHIIN, KANA, JP1, JP2

第2表　選択表記規則

		BOIN	SHIIN	KANA	JP1	JP2	WI	WE	WO	TU	DU
ひらがな	a_1	CBOIN	CSHIIN	0	0	0	お	え	を	つ	づ
	a_2	〃	〃	0	1	0	うぃ	うぇ	を	と	ど
	a_3	〃	〃	0	1	1	うぃ	うぇ	うぉ	とぅ	どぅ
カタカナ	b_1	SBOIN	SSHIIN	100	1	1	ウィ	ウェ	ウォ	トゥ	ドゥ
	b_2	〃	〃	100	1	0	ウィ	ウェ	ヲ	ト	ド
漢字	c	〃	〃	100	0	0	井	江	ヲ	ツ	ヅ

で選択する。なお，第3表に漢字についての規則表を示した。その結果，表記に関する自由度が減少した。

第3表　漢字字音のローマ字表記

ア	イ	ウ	エ	オ	キャ	キィ	キュ	キェ	キョ	ガ	ギ	グ	ゲ	ゴ
a	i	u	e	o	kya	kyi	kyu	kye	kyo	ga	gi	gu	ge	go
カ	キ	ク	ケ	コ	ギャ	ギィ	ギュ	ギェ	ギョ	ザ	ジ	ズ	ゼ	ゾ
ka	ki	ku	ke	ko	gya	gyi	gyu	gye	gyo	za	{zi / ji}	zu	ze	zo
サ	シ	ス	セ	ソ	シャ	シィ	シュ	シェ	ショ	ダ	ヂ	ヅ	デ	ド
sa	{si / shi}	su	se	so	{sha / sya}	syi	{shu / syu}	{she / sye}	{sho / syo}	da	di	du	de	do
タ	チ	ツ	テ	ト	ジャ	ジィ	ジュ	ジェ	ジョ	バ	ビ	ブ	ベ	ボ
ta	{ti / chi}	{tu / tsu}	te	to	{zya / ja}	zyi	{zyu / ju}	{zye / je}	{zyo / jo}	ba	bi	bu	be	bo
ナ	ニ	ヌ	ネ	ノ	チャ	チィ	チュ	チェ	チョ	パ	ピ	プ	ペ	ポ
na	ni	nu	ne	no	{cha / cya}	cyi	{chu / cyu}	{che / cye}	{cho / cyo}	pa	pi	pu	pe	po
ハ	ヒ	フ	ヘ	ホ	ヒャ	ヒィ	ヒュ	ヒェ	ヒョ	ファ	フィ		フェ	フォ
ha	hi	{hu / fu}	he	ho	hya	hyi	hyu	hye	hyo	fa	fi		fe	fo
マ	ミ	ム	メ	モ	ビャ	ビィ	ビュ	ビェ	ビョ	ヴァ	ヴィ	ヴ	ヴェ	ヴォ
ma	mi	mu	me	mo	bya	byi	byu	bye	byo	va	vi	vu	ve	vo
ヤ		ユ	イェ	ヨ	ピャ	ピィ	ピュ	ピェ	ピョ	ツァ	ツィ		ツェ	ツォ
ya		yu	ye	yo	pya	pyi	pyu	pye	pyo	tsa	tsi		tse	tso
ラ	リ	ル	レ	ロ	リャ	リィ	リュ	リェ	リョ	クァ	クィ	クゥ	クェ	クォ
ra	ri	ru	re	ro	rya	ryi	ryu	rye	ryo	qua	qui	quu	que	quo
ワ	ヰ		ヱ	ヲ	デャ	ディ	デュ	デェ	デョ	グァ	グィ	グゥ	グェ	グォ
wa	wi		we	wo	dya	dyi	dyu	dye	dyo	gwa	gwi	gwu	gwe	gwo
ン					テャ	ティ	テュ	テェ	テョ	クヮ				
n (')					tya	tyi	tyu	tye	tyo	qwa				
ン					ニャ	ニィ	ニュ	ニェ	ニョ					
m (')					nya	nyi	nyu	nye	nyo					
					ミャ	ミィ	ミュ	ミェ	ミョ					
					mya	myi	myu	mye	myo					

{ }：いずれの表記の可

字種の識別を行なった後に，各々の字種による変換処理を行なうために，操作性を考慮して，数種の簡単な記号ヒシフトキーを定めている。

(1) 漢字 ― 鍵盤上に英文字の大小によるシフトキーが備えられているので，これを利用し漢字部分について小文字で表現する。

（2）ひらがな ― シフトキーにより大文字で表現する。
（3）英数字，特殊記号 ― 「?」記号で両端を囲んだ文字列で表現する。ただし，「?」自身は，!，?，@，英数字を除く記号で?を囲む。
（4）カタカナ ― 「!」記号で両端を囲んだ文字列で表現する。

2.4 固定辞書と浮動辞書

かなから漢字への変換処理は，辞書による自動変換と人間を介して変換する同音異字処理の部分からなっている。ここでは確実に変換することを主たる目的としており，処理速度は考慮されていない。

かな漢字変換のための辞書として，
（i）固定辞書
（ii）浮動辞書
の二種類が設けられている。

（i）の辞書とは，汎用性のある基本語について作成されている辞書で，半永久的に検索のみに使用されるものを指す。一字訓読み辞書，当用漢字音訓辞書等で構成されている。一字訓については，野村氏[1]の調査によれば，第4表のように一字訓が訓読みの語において大多数を占める。また同音異字が音読み結合形に比較し，極端に少ない値を示している。異なり語音[2]501項目について，その語音と漢字符号を辞書として登録した。なお，同音異字が最大5個のものが存在した。当用漢字音訓辞書は，約3,500項目が登録されている。

第4表 漢字音訓の用法分析 （数字は延べ度数 ( )内の数字は%）

用法	音	訓	計
自立	12,280(1.9)	80,489(53.9)	92,769(11.0)
結合	514,822(81.9)	54,089(36.2)	568,911(73.1)
接辞的(前部分)	13,926(2.2)	2,293(1.5)	16,219(2.1)
接辞的(後部分)	87,829(14.0)	12.511(8.4)	100,340(12.9)
計	628,857(100.0)	149,382(100.0)	778,239(100.0)

（言語生活 1975.6 より）

第5表 テキスト内での出現漢字語の分析

データ	延べ語数	異なり語数	異なり漢字語数	λ(%)	μ(%)
1	1,241	525	201	42.3	16.2
2	1,662	721	281	43.4	16.9
3	2,618	1,100	574	42.0	21.9
4	2,637	885	415	33.5	15.7
5	3,208	1,208	607	37.6	18.9
6	3,425	1,116	513	32.5	14.9

λ＝(異なり語数/延べ語数)×100
μ＝(異なり漢字語数/延べ語数)×100

（ii）浮動辞書は，本システムの特色をなす部分で，利用者（研究者）自身で作成し，検索を行なう辞書を指す。

利用者が処理対象とするテキストの大きさは，高々数千から数万字から成るものと考えられる。そのテキスト内で出現する異なり語（漢字列）は，実験から第5表に示すような値が得られている。すなわち，二文字以上で構成されている語（漢字）での同音異義語は，ほとんど出現しない。一字の場合も訓読みとして出現し，同音異義語として現われない。この点から，テキスト内で始めて出現した語（漢字列）を辞書に一回登録することにより，以下の処理では，自動的に検索が行なわれ，漢字への変換が行なわれる。

2.5 漢字テキストの編集処理

この処理は，本システムとは独立に，EPICS（TOSBAC 5600 大型計算機システム）のTSS環境下において，漢字データを含むテキスト・ストリングを編集する目的で，東芝に外註して作成したものである。ここでは，本システムと結合して使用するため，その特色を簡単に述べるにとどめ，詳細は，その解説書にゆずる。

2.5.1 漢字符号と入力表現

漢字符号の割付けは，JIS 標準化の作業が現在すすめられている段階で，多

くの問題を含んでいるわけで，一意に決定することは困難であるが，プログラム上でのデータ表現，一定の順序配列，約一万種の文字表現等の点から，10進4桁で表現した。ただし，JIS が決定されれば，この符号との変換を可能なものとする。

入力表現は，英数字のみからなるテキストで用いられている α-mode のデータを一定の条件のもとで，混在することを許すため，漢字データの区切りを示す，開始記号「'''' 」と終了記号「"""" 」で囲まれた数字列で表現する。

2.5.2 漢字データの出力

漢字データの出力形式には，つぎの4種が備えられている。

(i) α-mode による出力では，入力表現と同じもので，開始，終了記号を含む数字列からなる。

例　　'''' 1682 1065 0542 0070 0534 """"

(ii) PRINT 出力では，(i) の場合の開始，終了記号が除かれ，漢字データの文字区分を明示するために，4桁の先頭の数字に「＋」を重ね書きする。

例　　＋682＋065＋542＋070＋534

(iii) DISPLAY 出力では，漢字パターンで表示される。

例　　漢　字　デ　ー　タ

(iv) EXHIBIT 出力では，漢字符号での表示および漢字パターンで表示され，第1行を漢字符号列（PRINT表示），第2行を漢字パターン（DISPLAY表示）を並記する表示方法である。

例

＋682＋065＋542＋070＋534

漢　字　デ　ー　タ

なお，1行は最大18個（漢字）表示可能であり，1画面当りの収納行数は，数字列で35行，漢字パターンで17行収納できる。

2.5.3 編集の機能

編集処理を行なうための命令は，機能を示す動詞の部分と，機能を修飾するオペランドとの組合せから成り，その形態は，第6表に示す8種類であり，その動詞の種類は，第7表に示すものが可能である。

第6表　編集命令の形態

①	VERB
②	VERB m ; r
③	VERB m : st
④	VERB m : st ; r
⑤	VERB m : st , st
⑥	VERB m : st : st
⑦	VERB m : st , st ; r
⑧	VERB m : st ; r : st

注

m　：S, L, null（モード表示）

r　：n, *（繰返し回数）

st（ストリングフィールド）

第7表　編集制御の種類

BACKUP or B	COPY	RUNOFF	NONVERIFY
FIND or F	CUT	DISPLAY	ASCERTAIN
PRINT or P	PASTE	CASE	NOASCERTAIN
DELETE or D	BUILD	STANDARD	CONFIRM
INSERT or I	LINE	VERIFY	NONCONFIRM
REPLACE or R	STRING	KANJI	NORMAL
EXHIBIT			

2.6 漢字テキストの表示処理

漢字テキストを一定の様式に従って出力する処理は，前節で述べた漢字テキスト・エディタ同様，EPICSと結合して働く漢字RUNOFFサブシステムが，現在開発中である（'86年12月，稼動予定）。その概要を紹介すると，

改行，改頁，頁付け等のテキストの様式を指定するための制御語が多数設けられている。これを第8表に示す。

第8表　様式指定の制御語

①	.CHSIZE	n	⑳	.PAGE	x, y, n
②	.CHPITCH	n	㉑	.PAPERLENGTH	n
③	.LNPITCH	n	㉒	.PARAGRAPH	
④	.BEGINPAGE	n	㉓	.POINT	n
⑤	.BOTTOMMARGIN	n	㉔	.REFERENCE	n
⑥	.BREAK		㉕	.SCOREUNDER	n
⑦	.CENTER	n	㉖	.SINGLESPACE	
⑧	.COMMENT		㉗	.SPACE	n
⑨	.DOUBLESPACE		㉘	.SUBHEADER	x, n
⑩	.FILL		㉙	.SUBFOOTING	x, n
⑪	.FOOTING	x, n	㉚	.SUBPARAGRAPH	n
⑫	.HEADER	x, n	㉛	.TOPMARGIN	n
⑬	.INDENT	n	㉜	.UNDENT	n
⑭	.LEFTDENT	n	㉝	.TABULATE	n, ..., n
⑮	.LINELENGTH	n	㉞	.NOTAB	
⑯	.MARGIN	t, b, l, r	㉟	.BOLDFACE	n
⑰	.MULTISPACE	n	㊱	.HALF	
⑱	.NODENT		㊲	.FULL	
⑲	.NOFILL		㊳	.JUSTIFY	

これ等の制御語をテキスト中に挿入するのは，2.5で述べた漢字テキスト・エディタを用いて行なうことができる。また，これ等の制御語を9,000番台の数字によって表現することもできる。各制御語の説明は省略する。

この様式化されたテキストをファイルに蓄積または印字出力するために，REFORM命令があり，つぎの3種類の形式がある。

i) REFORM ファイル名1, ファイル名2, PRINT
ii) REFORM ファイル名1, ファイル名2
iii) REFORM ファイル名1,, PRINT / DISPLAY

制御語の中で，とくに JUSTIFY は禁則処理を行なう命令で，横組における，右端の2文字，左端の1文字に関する禁則規則を，フルピッチ，ハーフピッチの文字の組合せによる処理を行なうことが可能である。

漢字のフォントパターンは，その符号配列として，特殊記号，数字，英字，ギリシア文字，ロシア文字，ひらがな，カタカナの順に 0～599 の範囲に割付け，601 から上位に，当用漢字，人名漢字，補正漢字を部首別で割付けられている。漢字パターンは，文字を直線に分解し（最大62），端末へ伝送するベクトル方式がとられており，ドットによる直接伝送方式より，伝送速度の向上がはかられている。

3. 実験と結果

3.1 辞書の蓄積と検索

固定辞書となる一字訓読け，当用漢字音訓のデータは，その字音をローマ字表記し，漢字符号をかっこに囲む，つぎのような表現で入力される。

例；kan（1177, 787, 1776, 1608, ……, 1781）

なお，同音異字は「,」で区切り連続して入力される。

辞書の内部では，ローマ字からカタカナに変換され，文字を単位節とする Tree 構造で，漢字符号は Table の形で蓄積されている。このため，辞書の修正は容易である。

浮動辞書への未登録語の蓄積は，固定辞書を検索するために，字音または字音と「+」記号の形で入力する。

例；一字訓　aida /(CR)
音訓表　kan+/(CR)　　(CR) 改行キー

同音異字の選択は，画面の下部に表示されたものの中から選んで，その番号を打鍵する。蓄積は，語を単位として処理されるため，「=」記号を加える。

例　kan+/(CR)　；ji+=/(CR)　；漢字

この実例を第1図に示した。選択された語は，テキストに収められるとともに，浮動辞書に登録される。

複合語の表現は，

例　漢字　漢字入力　漢字入力処理

に対して，

(((漢字)入力)処理)

の形で3語として登録される。この3語の検索は，

例； kanji　　kanji, nyuuryoku　　kanji, nyuuryoku, shori

の形で行なわれる。
つぎに，語音補正処理は，

例　nichi+/㉔ ，hon+/㉔　go+=nihongo/㉔ ；日本語

の形で打鍵する。
検索全体の手順を第二図に示す。

（電子技術総合研究所）

1。はじめに　日本語テキスト

日本語テキスト（漢字かな混り文）を計算機へ入力する方法については、すでに種種の方式が開発されている。ここで紹

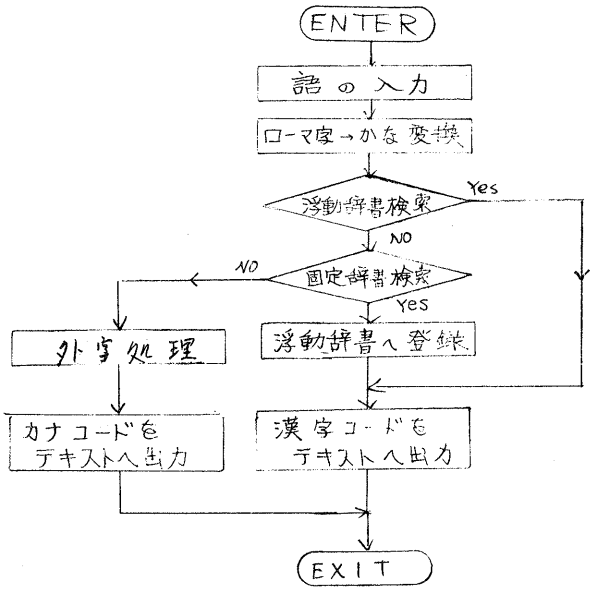
=kai+=

01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
介灰回会快戒改怪海悔皆界械開階絵塊解懐壊街貝
=01

日本
=go+=nihongo

01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
五互午呉後悟娯御碁語誤護期
=10

第1図　画面による選択表示の実例



第2図
検索の手順

3.2 テキストの蓄積と編集

1個のテキストを蓄積する単位をファイルと呼ぶ，1ファイルは N個のレコードで構成され，1レコードは，最大文字数を指定することにより構成される。蓄積時におけるレコード単位の修正は，バッファに一時的に記憶されており，つぎのレコードが入力された時点で，ファイルに蓄積される。ファイル単位の修正は漢字テキスト・エディタにより処理を行なう。この出力例を第3図に示す。

テキストの様式化は，漢字RUNOFFサブシステムを用いて，画像端末あるいは，他の出力装置により，印字出力を行なう。

3.3. プログラムと結果

主プログラムは，TSS FORTRANで約1,500行から成っており，漢字パターンの表示は，ユーティリティ・プログラム PLOT-10を主プログラムから呼ぶことにより行なう。一時訓，音訓，浮動辞書，テキスト・ファイルは，ディスク上のonline-fileとして蓄積され，検索処理開始前にコア上に読込まれる，その容量は約4KW（1W=36bit）である。

固定辞書の検索速度は，EPICSのTSS利用者の利用度に大きく影響され，とくに，同音異字が多い場合，その文字の漢字パターンを端末へ伝送，表示するのに，数十秒から数分を要する。さらに，画像端末として，蓄積型が用いられているため，画面の部分消去が不可能であり，固定辞書を検索するたびに画面全部の書きかえを行なう度で多くの時間を要している。

4. あとがき

今後の問題として，固定辞書の拡張，パターン伝送の高速化，画像の部分制御，種々の様式が可能なハードコピー等の改良とともに，新しい方式として，長谷川氏[3]の英字における分かち漢字変換方式の並置を検討中である。

応用として，既存の大量データの編集，活用処理[4]，Concordance[5]の諸プ

```
-DISP.*
い技本かつてる□
つ子日字にれな□
に電　漢法さと□
理（に（方発象□
処　めトる開対□
力じスすが、□
出行はキ力式は□
入義。テ入方の□
の本1語へのる□
字坂□本機種す□
漢　□日算種介□
　　）　計に紹て
　　所□をでで し
　　究ト）すこと
　　研ス文、と者
　　合キりは。用
て□総テ混てる利
術語ないい□
end of file
```

第3図　漢字テキストエディタの出力（DISPLAY）

にグラと結合し，日本語テキスト分析をすすめる。

最後に，本プログラム作成を助けて下さった，電気通信大学の佐藤雅之君，漢字パターンを提供下された人間機械研究室，共通ソフト開発で検討会に参加して下された諸氏および東芝の担当者の方々に感謝いたします。

参考文献

1）野村雅昭；新聞の文章につかわれた漢字，P35，言語生活，1975.4，筑摩書房。
2）林四郎；漢字の役割，P25～26，言語生活，1975.6，筑摩書房。
3）長谷川貞夫；英字符号による漢字を含む普通文字の印刷，日本特殊教育学会第13回大会発表論文集，1975年9月。
4）坂本義行；日本語の活用処理，CL5-2，計算言語学研究会資料5，1976.5，情報処理学会。
5）坂本義行；岡本哲也；日本語のコンコーダンス，CL2-1，計算言語学研究会資料2，1975.7，情報処理学会。