計算言語学 25－3
（1981. 2. 6）

MDSによる日本語入力と編集処理について

坂本義行（電子技術総合研究所）

1. はじめに

日本語の計算機処理において、他の欧米語と異なる主たる点は、文字種の形と個数、語の構成といった形態素部分である。この点に関して、これまでの処理方法は、必ずしも日本語のもつ特長を十分に活用したものとはいえない。漢字が有する「文字」の視認性と「読み」が有する操作性を最大限利用した入出力処理方式を採用すべきである。

一方、マイクロコンピュータ、ディスプレイ等のデバイス技術の進歩に伴って、端末装置は、分散処理の考えに基づく処理機能、低廉・高品質のディスプレイを備えた知能端末が実現される傾向にある。ここでは、計算機と対話していく上で、複数・多重画面を用いることで、複雑なテキストを並列的に、同時に眺めることができ、画面の切り換え、複数画像を重ね合せることにより、より効率的な入出力、編集処理が行える、新しい知能を有する日本語ワードプロセッサについて述べる。

2. 処理の方法

2.1 日本語のワードプロセッシングとは

言語テキストを計算機で処理する場合、大きく次の2つに分けられる。

i) 1次資料としての印刷物を計算機により作成する。

ii) 2次資料を計算機で作成する。

i)においては、計算機はその言語内容を理解する必要はなく、極論すれば、2次元的に白黒のドットに分解し、蓄積し、出力時に復元可能であればよい。

ii)では、文字、語、文といった単位で認定できなければならない。すなわちテキストをストリング列に分解し、その言語のもつ形態情報、構文情報、意味情報を抽出するといった、i)と比べて、はるかに高度な言語処理手続きを必要とする。ここではii)を満足するようなプロセッサを目標とする。

2.2 文字の処理

日本語は、他の欧米語と異なり、漢字、ひらがな、カタカナ、英数字、記号等多くの文字種が用いられている。1昨年、情報交換用漢字符号体系というJISが制定され、これは通常の国語文に用いる図形文字について、情報交換に用いる符号の規格であると述べられている。しかし実用上その用字の範囲を決定することは難しく、分野により片寄りがある。一般に知られているものとして、

康熙字典　42,170　　当用漢字　1,850

大漢和（諸橋）48,902　　常用漢字　1,926（当用漢字－19，＋95）

この他に、国語研究所が調査したものに、

雑誌90種（S31）3,228（延 43万），新聞（S41）3213（延200万）

しかし、あらゆる分野における分野別の正確な統計はない。この雑誌と新聞について比較してみると、両方に共通なものが、2,831字、雑誌のみが476字、新聞のみが382字、全体では3,690字となっており、これは標本が異なるというだけでなく、その差は、国語政策、時代の違いによる面が大きいとおもわれ

る。

これに対して，JIS では第1水準として，2,965，第2水準として，3,384字を設定し，これは，情報処理に用いられてきた字を中心に選んだものだと説明されているが，漢字処理システムを作成するときの目安程度にしかならない。また，漢字処理の一分野をなす住所録ではとくに古い姓名に特別な字があり，その処理を困難とする。さらにJISでは，字体の明確な定義はなされていない。これも姓名では問題となる点である。姓の調査によると約3,500字との報告があるが，その数値は，JISの第1水準に近い値であるが，JISに含まれていない字が多くあり．一般のシステムにおける使用頻度を考えると用字の範囲を決定することは容易でなく，共通部と変化部を設ける必要があり，その操作が容易でなければならない。

漢字処理システムにおける書体，字形，表記法の問題も，使用目的によって決定すべきである。ドット表示か，ストローク表示か，1字を表わすドット数を，25×25，36×36，64×64，128×128 といった表示法が多く用いられている。どれを採用するかは，処理効率に大きな影響を与えることとなる。その処理が一時的な確認といったディスプレイ的なものでは，ストローク表示，品質が要求される印刷物では，64×64 以上のドット表示といった使い分けが必要である。

内部コードをどのように与えるかは，配列の問題も含めて困難な問題である。JISの符号は，ASCII 2文字組みが採られている。通常の漢字を表現するには十分であるが，いくつかの問題点をも含んでいる。たとえば，プログラムを書く上での漢字データとの識別ができないといった問題があり，電総研では，10進4桁の数値を用いている。これは処理を容易にしている反面，記憶容量を多く必要とする。

配列を決定するには，その内部処理とは独立に決めるべきだと思われる。情報交換用には，外部索引用にJISといった共通の符号に統一し，内部では，個々の処理向きのコード体系を採用すべきであり，その間のコード変換は，現在のハードウェアを用いれば，大きな問題とはならない。

漢字以外の文字にも多くの問題がある。仮名（ひらがな，カタカナ），英数字，ギリシャ，ロシア，記号（たて書き用，よこ書き用），制御コードの範囲とその配列の問題がある。

漢字かな混じり文を処理するための最小単位を文字とすれば，この符号化とその処理機械の最小単位が処理効率に大きな影響を与える。従来の計算機では，character（6ビット）または，バイト（8ビット）を単位にソフトウェアができており，これは英数字をとりあつかう場合に便利である。しかし，漢字の場合は，上述のように 3,000 以上の符号を必要とし，ビットまたはバイトの組み合せによって表現せねばならない。すなわち，SNOBOL4のようなストリング処理用言語では，バイト単位でサーチが行なわれるため，漢字の区切りを挿入するといった前処理を必要とする。この点から，漢字一字をワード（16ビット）とするようなハードウェアとソフトウェアが必要となる。

2.3 入力の具備要件

日本語のテキストを計算機に入力する場合，大量のデータを一括して処理する場合と，小量の種々異なるデータを散発的に処理する場合がある。前者は熟練者が，後者は素人が処理を行なう点で異なる。以下は，後者の場合について必要な条件を項目別に列挙してみる。

i) 入力の操作性 — 小形，軽量，労働性（目，腕，指の運動）
ii) 教育 — 容易性，普遍性
iii) 信頼性 — 視認性（入力確認のためのモニタ，ベリファイ）
iv) 外字処理 — 標準の文字種，外字の入力の容易性
v) 校正 — 速時性

2.4 編集の具備要件

ワードプロセッシングにおける最も重要な機能の1つがこれである。これは，テキストのドキュメンテーションをいかに透明にするかであり，そのために必要な項目を列挙する。

i) テキスト表示 — 文字・図形の表示速度と視認性
ii) ファイル管理 — テキストの蓄積，検索，ページング処理
iii) 編集機能 — コマンド，メニューとその入力，カーソル表示，校正前後の情報
iv) 書式機能 — 用紙設定の自由度，禁則処理，図式の処理，ディフォルト処理

以上の諸点を考慮したワードプロセッサについて以下に述べる。

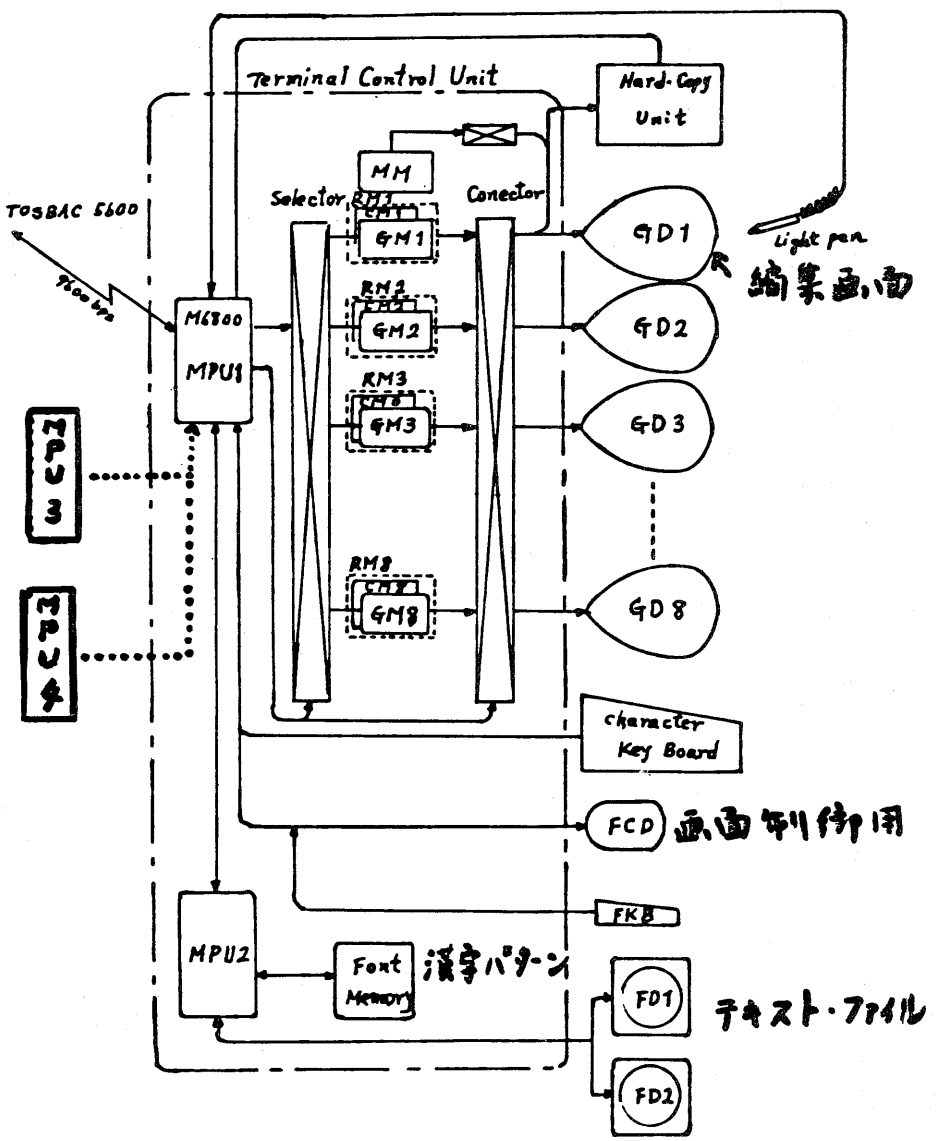
3. ハードウェア構成

本システムのハードウェア構成図を第1図に示す。複合多重な画面制御をMPU1ファイル管理と漢字処理をMPU2といった2台のマイクロコンピュータで処理する分散処理方式を採用している。MPU1はMDC（Multi Display Controller）部の制御，すなわち，グラフィック・ディスプレイ（GD，最大4台まで可能），いづれのGDとも接続可能なリフレッシュ型のグラフ・メモリー（GM），GD1に接続され重複表示可能なメッシュ・メモリー（MM）等とMPU1間の切り換え，ハードコピー，ライトペン，キャラクター・キーボード，画面管理用ディスプレイ（FCD）とキーボード（FCB）の制御および大形のホストコンピュータとの交信の役割を果す。

MPU2は，2台のフロッピー・ディスク（FD1，FD2）上のテキストの蓄積検索とROM（128KB）上に合成フォント方式で蓄積されている漢字パターンの検索を行なう。

ドット方式と比較して，その記憶容量，伝送速度，使い安さの点からストローク方式を採用した。フォントの記憶は文字を直線のストロークに分解し，その座標を記憶しておき，この座標を出力装置まで伝送し，その装置が有するストローク・ジェネレータにより直線を復元する方法である。この方式は，文字のサイズを変える，あるいは回転（たて書きにおける欧米字）の表示等が簡単に行なえるといった利点がある。さらにフォント・メモリーを大幅に節約するために，作成しようとしている文字の部分形状が他の文字の部分形状と位置を含めて全く等し

いとき，それを借用して合成する方法で，多くの文字に共通性のある扁，旁など約90種を合成フォントとして用いている。[1]



第1図 システム構成図

4. 処理機能とその手続き

4.1 入力処理

4.1.1 ローマ字入力

汎用性の面から，特殊なカナ鍵盤を用いず通常の英字鍵盤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，訓練が不要という点からカナによる入力と差がないと思われる。[2]

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。

字種による分かち書きで入力する方法を採った，これは視認性と操作性を考慮したものであり，字種の識別キーとして下のようなシフト・キーを定めているが，漢字，ひらがなに関しては，大小の英文字を与えているため，英文での操作と大きな違いはないと思われる。

i）漢字 ― 英小文字で表現　ii）ひらがな ― シフトキーによる大文字で表現　iii）カタカナ ― 「!」記号でその両端を囲む　iv）英数字・特殊記号 ― 「?」記号でその両端を囲む，ただし，「?」自身は，!，?，@，英数字を除く記号で?を囲む。

4.1.2 仮名漢字変換

仮名漢字変換では，仮名文をある単位（語，文節）に区切る手順と同音異義語の中から適正な1語を選ぶ効果的な手続きが開発されねばならない。ここでは，分かち書きされた小文字列に対する自動変換システムとして，点字文のために開発したシステム[3]をこのシステムに応用するとともに，半自動変換システムとして，人間による選択表示方式が併用されている。自動変換で誤った変換がなされた語に対し，語単位あるいは文字単位で修正を施したものが，自動的に「学習辞書」に登録され，以後同一テキスト内では優先的に選択される。

利用者が処理対象とするテキストの大きさは，高々数千ないし数万字からな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数（漢字列）は，実験によると第1表のような結果が得られ，2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結果が得られている。

第1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

4.1.3 語音補正と速記処理

文字単位の読みで"ある「字音」が認められてなく，語としての読み「語音」が認められている場合がある。たとえば，「日本語」において，「日」は当用漢字音訓表では，「に」という字音はない。しかし「ニホンゴ」という語音は認められている。この補正が行なえるように，文字の連鎖とその語音登録の機能が付加されている。すなわち，

日本語　nichi, hon, go = nihongo　により表現できる。

又，同様な手続きにより，決り文句，長い固有名詞等は，その略称を学習辞書に登録することにより，以後その利用することができ，入力速度の向上に役立つ。

電子技術総合研究所　denshi, gizyutsu, sougou, kenkyu, sho
= densouken

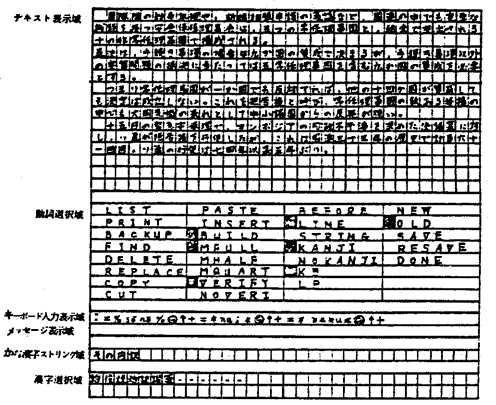
4.2 編集処理

4.2.1 編集機能

GD1を編集作業用マルチ・ディスプレイとする。その画面構成を第2図に示す。

テキスト表示域には，メッシュパターンにより，原稿用紙が表示される。これは，メッシュの書き換えにより，変更が可能である。編集コマンドは，動詞選択域にメニューとして表示され，ライトペン，又はキーボードのいずれからでも指定できる。

同時に修正データの表示および修正データが漢字を含み，かつ同音異字を含む場合には，ローマ字で入力すると下端に同音異字が表示され，ライトペン指示により，選択され，テキストが修正される。



第2図 編集作業画面の構成

4.2.2 画面制御とファイル管理

GD1からGD4の4台のマルチ・ディスプレイは，フロッピー上のファイルから，指定したテキストの必要なページをGD1に表示でき，同時に，他のページをGD2～4に表示することができる。又画面間の移動も任意に行なえる。LISTというコマンドでは，テキストの原稿が，ページ単位でスクロール・アップするように移行される。

画面制御とテキストのページングを行なうために，画面とテキストのページを対応づけした漢字テキスト・ディレクトリと編集作業画面を制御するための漢字カレント・バッファ・ページを設けている。

4.3 書式処理

テキストを一定の書式に従って出力するためのRUNOFFサブシステムをもっている。

改頁，改行，文字サイズ，文字ピッチからヘッディング，頁付け，センターリングといった書式を指定するためのコマンドが多数設けられている。これを第2表に示す。

これらコマンドをテキスト中に挿入するのは，エディタを用いて行なうことができる。

第2表 書式指定の制御語

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	

この書式化されたテキストをファイルに蓄積または，表示，出力するために，つぎの3種のPerform命令がある。

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

センターリング{

巻頭言

会長就任に際して

穂坂　衛*

脚注

私、今回当学会の会長に推薦されまして、驚き大変な当惑を感じ、初めは固辞したのでありますが、諸般の事情から自分の仕事をある程度犠牲にしてもこの大役を引き受けざるを得ないのではないかと思うに至りました。もちろん身にあまる光栄でありますが、今は責任の重大さをひしひしと感じております。歴代の会長はこの分野の先駆者、すぐれた学者、またトップの管理者であられ、正に会長にふさわしく会の運営に必要な影響力を十分にもっておられて、学会の発展に尽くしてこられました。私はその何れでもなく、影響力の持ち合わせのないことは自分が最もよく知っている所であります。しかしお引き受けした以上は、誠意と努力をもち、会員、理事役員、事務局員の皆様の御協力を得まして、本学会の目的を常に見失わず、広い見解と先見性をもって、会員の要求に答え、光栄ある学会の歴史を受けつぎ発展させてゆきたいと決意しております。

* 本会会長　東京大学宇宙航空研究所教授　←文字サイズ

525　ページ付け

第3図　書式出力例

コマンドの中で，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文字に関する禁則処理を，フルピッチ，ハーフピッチの文字の組合わせによって行なう。第3図に，いくつかのコマンドにより書式化された出力例を示す。

5. あとがき

MOSのOSを現在開発中で，本システムは，現在，各機能別に処理可能な状態であるため，本プロセッサの操作性，処理効率についての結果が得られていない。またテスト中に挿入すべき図表の編集機能は含まれてなく，現在設計の段階である。

さらに，日本語ワードプロセッサの本来の目的である日本語テキストの有効かつ経済的な蓄積，加工，検索を行なうためには，漢字の特徴を有効に利用したワード処理として，以下のような機能を開発する必要がある。

i) 入力処理として，操作性，経済性の面ばかりでなく同音異義語等の漢字がもっている言語特性が得られる機能を付加する。

ii) 単語に分割されていない日本語文から自動的にキーワードを抽出する機能．さらに文節あるいは文を規格するために，正書法（送り仮名）を確立するに必要な自動文節認定システムを開発する。

iii) テキストの構造は単に文字のシーケンシャルな列ではなくコンテクスチュアルである。ワードプロセッサの機能は，文字列である一次資料から構造化された二次資料をつくりだすかであり，語句，文句の結合関係を抽出できる機能を付加する必要がある。

参考文献

1) 大岸洋他；「汎用端末装置による漢字フォントの作成」，IECタ-ム，電子通信学会，1979.4.26

2) 坂本義行；「漢字入出力処理の一方式について」，第13回情報科学技術研究集会論文集，JICST，1976

3) 坂本義行；「日本語の点字情報に関する計算機処理エ-仮名点字自動作業システム-」，第16回情報科学技術研究集会論文集，JICST，1979.