日 本 語 情 報 処 理

東京芝浦電気株式会社 総合研究所 森

健 一

漢字、かな、カタカナ、英数字、記号など、多種類の文字種を含む日本語情報を処理するために研究開発されている入出力装置、漢字パターン発生器、カナ漢字変換、漢字オフィスコンピュータなど、最近の技術動向について解説した。

1. ま え が き

日本語は、漢字、ひらがな、カタカナ、英数字、記号を含む多種類の文字種を使用する国語であるため、日本語の情報を計算機やその他の機械で取扱うには種々の技術的障害があり、広く普及するところまでなかなか発展しなかった。最近になり、LSI 技術、画像工学技術、計算機利用技術の進歩により、ようやく一般社会にも日本語情報処理の研究成果が利用される展望がひらけてきた。

昭和 53 年 1 月には、日本語情報処理の基本ともなるべき漢字を含む約 6,800 字種の文字符号に関する日本工業規格 (JIS C 6226)¹⁾ が制定された。選挙速報には計算機による候補者名や得票予測が漢字パターン発生器によりテレビに表示され、大手出版社の単行本、文庫本は写植機によって印刷されるものが多くなり、オフィスでは漢字による宛名印刷が用いられ、漢字オフィスコンピュータ、日本語ワードプロセッサも使用され始めた。

特にオフィスの機械化は日本が欧米諸国に比較して大変遅れている分野である。日本の誇る工場の高い生産性に反して、事務所のホワイトカラーの事務能率は欧米のその数分の 1 程度だといわれている。文書の作成、情報の蓄積・検索・伝達など事務のあらゆる面で、使用している言語が日本語であるために、これまで機械化がむずかしかったことから、最近の日本語情報処理の研究は、オフィスオートメーション、ワードプロセッシング、日本語による情報検索などと結びつけられて研究されている²⁾。

本稿では、最近の日本語情報処理の技術開発に関して、日本語情報の入力装置、出力装置、処理装置に分けて解説を試みた。さらに、本稿に述べられた技術を基にして

近い将来にどのような応用システムの研究開発が進められているかについて項目を挙げた。

2. 日本語情報入力装置

日本語情報処理をするには、まず日本語の情報を入力する手段がなければならない。欧米では一般事務で用いる文字種が多くて 100 字以下であるので、タイプライターが発達し、これを基本にして情報処理用の入力装置が開発されている。日本では文字種が多いため、なかなかタイプライターが普及せず、事務所の誰もが邦文タイプライターを使うというわけにはゆかない。現在までにオペレーターが用いる日本語情報の入力装置として研究開発されたものは数多くある。その原理も多種多様であるが、現実的に効果的なものは表 1 にあげた 5 種類の方式である。以下に各入力方式について特長を紹介する。

2.1 漢字テレ鍵盤穿孔機

昭和 30 年代に実用化された最も古い日本語情報入力装置で、新聞社、出版社などで現在でも広く使われている。200~500 個の鍵盤上に 2~15 字の文字が収容されており、右手で文字鍵盤を押し、左手で 1 つの鍵盤上にある複数の文字のうちの入力したい文字の位置を選択す

表 1 日本語情報入力装置の代表的な方式

方 式	特 長	入力速度 (字/min)	収容文字 数 (字)	タッチ打 鍵/目視 打鍵	必要と する熟 練度
多段シフト方式 (漢テレ鍵盤式)	現在も使われている両手使用の入力	40~80	2,000 ~3,000	目視打鍵	大
タブレット方式 (ペンタッチ式)	小型軽量でペン指示入力	30~70	2,000 ~4,000	目視打鍵	中
邦文タイプライター方式	邦文タイプストが使え	30~50	2,000 ~4,000	目視打鍵	大
2 打鍵文字コード方式	最も高速に入力できるプロ向型	120~200	2,000 ~10,000	タッチ打鍵	大
カナ漢字変換方式	カナ文を漢字混り文へ自動変換する	50~100	2,000 ~7,000	タッチ打鍵も可能	小

"Japanese Information Processing" by Kenichi Mori (Toshiba Research and Development Center, Kawasaki)

るためのシフト鍵盤を押すことにより特定の文字が入力される。習熟したオペレーターは左手の選択キーはタッチ打鍵（鍵盤を見なくても指が基本位置から正確に動いて打鍵する方法）ができ、右手も頻度の高い文字では速く動かせるように文字配列を工夫することにより、40～80 字/min の入力速度が実現されている。この鍵盤を使いこなすには相当の経験を積まなければならない点が問題となり、以下の多くの方式が研究されるようになった。

2.2 タブレット型入力装置

平面上のタブレット上に 2,000～4,000 字の文字を配列しておき、ペンで文字を指すと装置がペン先の位置を検出して、指された文字座標を文字コードに変換する方式である。タブレット板とペンの位置関係を検出するには、タブレット板内に直交して配列された多数の水平軸検出線と垂直軸検出線とペンとの間の静電容量が最大となる位置を検出する方法、電磁誘導を用いる方法、タブレットの文字盤の下に発光ダイオードを埋込み順次発光させてゆき、ペンが指している文字が発光するとペンに出力信号が現われる方式など、多くの方式が開発されている。軽量小型で、キーなどの機械的な可動部がないのが特長である。文字は部首順や、音訓の読み順に並べられているので、オペレーターは使い方を覚え易い。しかし小さな文字をペンで指さねばならず、速度はあまり早くはできない（図 1）。

2.3 邦文タイプライター改造方式

邦文タイプライターを打てる人は全国に 10 万人ぐらいいるといわれているが、これまで日本語文書の清書用として事務所で使われてきた邦文タイプライターを改造して日本語情報入力に用いようとするのがこの方式である。

代表的な方法は、タイプ活字の側面に、その活字の文字コードを白黒の帯状のパターンとして貼付けておき、タイプする直前に光電変換装置によって文字コードを読

取る方法である。邦文タイピストの入力速度はあまり早くなく、タイプライターの信頼性に問題はあるが、価格が比較的安く実現できること、入力と同時に入力情報の記録がとれること、邦文タイピストの訓練は従来通りのものでよいことなどに特長がある。

2.4 2 打鍵コード入力方式³⁾

漢字はそれぞれ音訓の読みを持っているが、1つの読みに対し同音異字がきわめて多い。コウやショウの読みを持つ字は 100 字以上もある。そこでカナ文字 48 字を 2 音組合せてできる 2,304 組に別々の漢字を割当てておけば同音異字が発生することなく、カナ鍵盤を 2 回打鍵することにより 1つの漢字を指定することができる。2,304 字の中には当用漢字、カナ、英数字記号を含めることができる。1つの文字に対応する 2 音の読みを憶える必要があるが、タッチ打鍵ができるように訓練すると、その入力速度、正確度は他の方式に比較して圧倒的に高い。われわれが「空気、空模様、空部屋、空っ風」などの単語を紙に書こうとすると、意識の方では「空」の文字を「クウ、ソラ、アキ、カラ」などと読んでいるが、手の方は無意識に同じ筆順で文字「空」を書いている。「空」の文字をみると反射的にカナ鍵盤上の 2つの鍵盤に指が動くように訓練するには 200～400 時間程度の練習を必要としている。

2,304 字以外の文字についてはシフトコードを挿入することにより、2,304 文字を単位として新しいコード表を対応させることができるので、原理的には何万字でも入力できることになる。2 音と 2,304 字の対応表はいくらでもできるが、指の動き易さ、文字の出現頻度などを考慮して、カナ鍵盤上の文字位置と 2 音対応表を作成せ

図 1 日本語情報入力装置の比較

図 2 ラインプット鍵盤および各段と指の使用率

図3 英文タイプ（アメリカ）とラインプットの習熟曲線

ねばならない。この原理にもとづいて開発されたのが図2のようなラインプット鍵盤である⁴⁾。この装置では出現頻度の高い文字ほど左右の手が交互に動き、指は図2のような使用率になるため、高速入力を無理なく長時間続けることができるようになっている。図3にはラインプットの200時間までの練習時間に対する習熟曲線を示している。英文を1語5打鍵とし、日本語の1字2打鍵と比較のため同じ図上に示した。この結果からみると、英文タイプの習得曲線とラインプットの習得曲線はほぼ同じで、2打鍵コード方式が特に習熟に困難であることは示されていない。しかしながら、日本ではタイピストに200～400時間の練習を課する習慣が少ないため、一般的には習熟が困難であるとされている。

2.5 カナ漢字変換方式²⁾

2打鍵コード入力方式が漢字の全く新しい読みを憶えさせるのに対し、われわれが普通に読む発音通りに日本語文章をカナタイプライターで入力し、計算機ソフトウェアでカナ文字を漢字混り文へ自動変換するのが、カナ漢字変換方式である。漢字には同音異字が多いが、熟語単位にするとその数は少なくなり、文法解析を援用すればさらに少なくなる。例えば「コウショウスル」というカナ文に対し、「コウ」「ショウ」は100字以上同音異字があるが、「コウショウ」は約20語ぐらいになる。さらに語尾に「スル」がついて動詞になり得るものは、「交渉、考証、口承、公称、哄笑、高唱」などに限られてくる。

一方、われわれが日常用いている単語は、使用する事務分野ごとに相当に偏った、比較的数少ない単語であることが多い。例えば国語辞典には6万語程度の単語が採用されているが、画像工学に関する論文とか、新聞の経済欄にはその数分の1の単語しか出現しないのが普通である。「コウショウスル」についていえば、工学の論文では「交渉、公称」ぐらいしか出現しない。この点を利用して、カナ漢字変換の単語辞書に用語の分野別、個人

図4 カナ漢字変換の処理フロー

文節式 入力原文

いんさつもじにんしきそうちは かんじ、ひらがな、カタカナ、アルファベット、すうじ、きごうから なる 2せん しゅるいいじょうの もじを つかって いんさつ または タイプされた にほんごぶんしよを まいびょう 100 じの スピードで よみとる ことが できる そうちである。

文節式 変換結果

印刷文字認識装置は漢字、平仮名、カタカナ、アルファベット、数字、記号からなる2千種類以上の文字を使って印刷またはタイプされた日本語文書を毎秒100字のスピードで読み取ることが出来る装置である。

図5 カナ漢字変換の入力文と変換例

別の使用頻度を学習する機能を付加することにより、必要な単語のみを参照し、同音異義語の出現率を小さくすることができる。

カナ漢字変換方式は、カナタイプを打てる人なら誰でも簡単に日本語情報を入力することができるので、最も将来性の大きい方式であると考えられている。また、カナ文の入力は、原稿のみをみて打鍵するタッチ入力方法を用いることができるので、従来の漢テレ鍵盤、邦文タイプライターより高速な入力が可能である。

カナ漢字変換方式の処理フローを図4に示した。入力のカナ文は適当に文節に分ち書きするか、漢字にすべき部分を指定して入力される(図5参照)。後者の入力方法では、多くの場合、分ち書きはされていないので、自動文節抽出の処理によって、文節に分割される。1つの文節は次のような構造をもっている。

{[接頭語]{自立語}{接尾語}}*[付属語]*

ここで*印は重複してもよいことを示している。自立語としては名詞、副詞、連体詞、接続詞、感動詞などの体言と、動詞、形容詞、形容動詞などの用言を含んでいる。

日本語には漢字熟語を連結して複合語を作る能力がきわめて高いという性質がある。例えば「画像再構成演算方式」という語には、少なくとも4つの単語が含まれている。このように造語力の強い日本語のあらゆる可能な複合語を、すべて辞書に格納しておくことは困難である。したがって辞書には「画像、再構成、演算、方式」など

写真1 漢字認識装置

の単語単位で格納したとすると、「ガゾウサイコウセイエンザンソウチ」というカナ文字列が与えられたとき、機械はどこが単語のされ目であるかを判定する機能をもたねばならない。

したがって、カナ漢字変換方式は、①文節、単語へのカナ文の分割、②辞書との照合、③文法解析、④頻度学習機能、⑤同音異義語の選択などの機能から構成されている。

2.6 漢字認識装置⁶⁾

日本語の情報で印刷されている書物や新聞などの文字を読取り、計算機へ入力する装置として、漢字認識装置が開発されている（写真1）。この装置は8～12ポイントの活字で印刷またはタイプされた書物や文書を、100字/secの速度で読取る能力をもっている。文字行は縦書でも横書でもよく、用紙は特許公報などのように普通紙に印刷されたものでよい。

漢字を読取る技術は、パターン認識および画像処理最先端の技術が応用されている。文字のセンサーはCCDアレイであり、1,700素子2個によりA4判の用紙を約20本/mmの解像度で走査している。文字パターンの特徴により約2,000字種の文字の中から100字以下の候補文字が選ばれ、混合類似度法と呼ばれるパターン認識理論にもとづく方法によって候補文字群から正しい文字が認識されている。

3. 日本語情報の出力装置⁷⁾

日本語情報の出力には、字種が多いことが原因となり多くの技術的な困難さがある。その最大の問題は価格で、これがために普及が限られていたといっても過言ではない。この他に、文字品質、書体、縦書／横書、外字処理などの問題との間のトレードオフも考慮されねばならない。

3.1 漢字パターン発生装置

写植装置など高品質の文字を要求する場合には、字母

表2 ドット漢字パターン発生装置

ドットパターン記憶方式	文字発生速度(字/sec)	収容文字数(字)	適用出力装置
フロッピーディスク+RAM	5～50	～3,000	端末プリンター、ディスプレイ
磁気ディスクメモリー+RAM	20～100	～25,000	ラインプリンター、ディスプレイ
LSI-ROM	～30,000	～6,000	ディスプレイ、写植機、ラインプリンター、端末プリンター
磁気バブルメモリー	100～1,000	～10,000	ディスプレイ、ラインプリンター

板などのアナログ的な漢字パターン発生装置が用いられていたが、最近の日本語情報処理装置ではデジタル式のドット漢字パターン発生装置が主流となっている（表2）。1文字をあらわすドット数は16×16、24×24、32×32、48×48程度のものが多い。

写植機では128×128ドット程度の高品質文字が要求されているが、情報処理分野では24×24程度で充分であるとされている。JIS標準の6,800字種を記憶するには、24×24の場合でも約400万ビットのメモリーを必要とする。

現在、最も安価な記憶装置はフロッピーディスクメモリーである。これは速度も遅いので主として端末プリンター用に用いられている。1枚のフロッピーディスクの収容文字数は約3,000字である。磁気ディスク装置を用いる場合には、収容文字数は最も小さい磁気ディスクでも約25,000文字を格納することができる。

LSI技術の発達で価格、速度とも最も期待されているのがLSI-ROMである。LSI-ROMを利用するためには漢字のドットパターンの字形を規格化する必要がある。さらに、磁気バブルメモリーの利用も期待されているが、素子の低価格がどこまで達成されるかがキーポイントになっている。

3.2 漢字ディスプレイ

漢字ディスプレイには、コードリフレッシュ型とパターンリフレッシュ型がある。英数字のディスプレイは前者のものが多いが、漢字ディスプレイでは後者の形式のものが多い。

前者は、キーボードや計算機から与えられた文字コードの情報を記憶しておき、画面のリフレッシュのたびに、所定の位置に文字コードから文字パターンへ変換して表示しなければならないので、漢字パターン発生装置が高速でなければならない。画面に500字表示し、リフレッシュサイクルを50Hzとすれば、25,000字/secの文字発生速度を必要とする。表2の中ではLSI-ROMだけがこの速度をもっている。

他の漢字パターン発生装置はこの速度がないため、高

速のランダムアクセスメモリー (RAM) に一度移してからでなければコードリフレッシュ方式を用いることができない。

パターンリフレッシュ方式では1画面分のリフレッシュメモリーをもっており、文字パターンは所定の位置にうめ込むとき、一度だけ読出されるので、その漢字パターン発生速度は高速性を必要としない。外字の合成や図形の表示も可能である。

漢字ディスプレイ装置は、従来、漢字パターン発生装置が高価であっ

たために、なかなか普及しなかったが、フロッピーディスクメモリーなど非常に安価な記憶装置と、小容量のRAMを用いたバッファメモリー、およびパターンリフレッシュメモリーの組合せによって、その価格を大幅に引下げる技術が開発された結果、漢字オフィスコンピューター、日本語ワードプロセッサ等に用いられるようになった。

漢字ドットパターンの規格化が行われれば、LSI-ROMによるコードリフレッシュ方式の漢字ディスプレイが主流になると予測されている。この場合、通常のCRTの解像度では、24×24ドットが最適なドット構成であると考えられている(写真2)。32×32では文字品質は向上するが、1画表に24×12行程度の文字数しか表示できなくなる。一方、16×16ドットでは明朝体の可読性の良い文字をデザインすることが困難であるからである。

3.3 漢字プリンター

漢字プリンターには非常に多くの方式があるが、最近になって、価格、速度、品質、コピーなどの点から注目されているものには、次のようなものがある。

高速プリンター：電子写真プリンター

中速プリンター：インクジェットプリンター、ワイヤドットプリンター

表3 代表的なドット式漢字プリンター

記録方式	特長	記録速度	分解度 (本/mm)	用紙	問題点
電子写真 プリンター	超高速印字 高品質文字 文字寸法可変	0.05~2μs/ドット 1,000~25,000字/sec	8~20	普通紙	価格 感光体寿命 転写定着技術
インクジェット プリンター	中速印字 文字寸法可変 低騒音	10~30μs/ドット 30~120字/sec	8~12	普通紙	インク粒子制御 同時複写 価格
ワイヤドット プリンター	中速印字 同時複写可	0.5~1ms/ドット (並列) 10~50字/sec	5~8	普通紙	ピン寿命 文字寸法可変 文字品質
感熱記録 プリンター	低速印字 文字寸法可変 低騒音	5~20ms/ドット (並列) 5~20字/sec	7~20	感熱紙	印字高速化 同時複写 文字改ざん性

低速プリンター：感熱記録プリンター

これらの方式による漢字プリンターの比較表を表3に示した。電子写真式のものとしては、レーザー光を用いたものやファイバーオプティックス管(FOT)を用いたものがある。高印字品質、高速であるが、高価な点が欠点となっている。中速度のプリンターとしては、ワイヤドットプリンターの技術開発が進み、価格も50万円程度となっている。騒音が少なく高品質の漢字プリンターを実現するには、インクジェット方式が有力であり、コピーを必要としない応用に用いられよう。

さらに、感熱記録紙に微小発熱素子アレイの熱記録ヘッドを接触させ、加熱した画素が発色する感熱記録プリンターが注目されている。この印字方式は乾式であり、機械部分が小さく、小型軽量で取扱い易い点が特長である。

用途によってこれらの漢字プリンターが例い分けられており、今後さらに価格の面で技術開発が進められよう。

4. 日本語情報処理装置

日本語情報の入出力装置をそなえた計算機応用システムは、これまでにいろいろ開発されてきたが、いずれも高価格なものであった。出版社、新聞社、信託銀行など、どうしても漢字処理をせねばならない業種が、高価であっても日本語情報処理システムを導入していた。最近になって2つの新しい安価な日本語情報処理装置が市販され、事務所や一般企業でも手軽に日本語による計算機利用、事務の効率化をはかれるようになってきた。漢字オフィスコンピューターと日本語ワードプロセッサの実現がそれである。

4.1 漢字オフィスコンピューター

英数字、カナを用いるオフィスコンピューターは小売業、事務所を中心に、すでに日本全国に普及しており、売上管理、在庫管理、集金業務、報告書作成など多くの

写真2 漢字ディスプレイの表示例

業務に使われている。これらのコンピューターは計算センター等に設置されるのでなく、主として事務所におかれるので、その利用者はプログラム言語などの特殊な知識がなくとも使えること、すなわち日本人であるなら日本語で使えることが強く望まれていた。

前述したように、日本語入力装置、出力装置が安価になりマイクロコンピューターが普及するにつれて、顧客名、住所、商品名などを漢字で入出力できる機能をもった漢字オフィスコンピューターが実現するのは当然のことであった⁹⁾。

漢字オフィスコンピューターでは、漢字パターン発生装置として、フロッピーディスクと RAM の組合せを用い、漢字プリンターと漢字ディスプレイの両者を制御している。フロッピーディスクは顧客、在庫、集計等のデータファイルメモリーにも使うことによって全体の価格をおさえている。漢字プリンターは同時複写の必要性から、ワイヤドット漢字プリンターを用いているものが多い。

4.2 日本語ワードプロセッサ

これまでの計算機が数値データを処理するのに対し、文字情報やテキストの処理を行うのがワードプロセッサである。ワードプロセッサは欧米で、事務所の生産性向上、省力化、合理化の有力機器として発達してきた。

日本語ワードプロセッサは、日本語の入力方法として

写真 3 日本語ワードプロセッサ

前述したカナ漢字変換方式や漢字タブレットなどを用い、出力装置として漢字ディスプレイとワイヤドット漢字プリンターを備え、さらに、テキストのファイルを記憶、検索するためのフロッピーディスクを含む独立型の日本語情報処理装置(写真3)である⁹⁾。

テキストの編集機能として、文章の入力、訂正、挿入、削除の基本機能の他に、2つの文の合成、さし込み、置換、移動、作表、タブレーション、印刷様式の変更、などの多くの機能をもち、手紙、仕様書、論文、申請書などの文書の作成や、日本語情報ファイルの管理に用いることができる。日本でも今後の事務合理化の有力な手段となり得るものである。

5. む す び

日本語情報処理の研究・技術開発の最近の動向について解説した。短い紙数なので、音声による日本語情報の入力、手書文字読取による入力、音声による出力などの入出力装置や、日本語情報検索システム、出版システム、住民情報システム、証券代行システムなどの応用システムにおける日本語情報処理技術、さらに、機械翻訳、自動抄録、質問応答、コンサルティングシステムなど、研究中の高度な日本語の理解システム¹⁰⁾については割愛せざるを得なかった。

(昭和54年2月2日受付)

(参 考 文 献)

- 1) 情報交換用漢字符号系 (JIS C 6226), 日本規格協会 (1978)
- 2) 日本語情報処理技術の動向, 53-C-349 (1978)
- 3) 山田: 日本語テキスト入力法の人間工学的比較, 東大情報科学科報告, 78, 06 (1978)
- 4) 川上ほか: タッチ打法による漢字入力, 情報処理, 15, 11 (1974) 863
- 5) 森ほか: 計算機への日本語情報入力, 信学会研究会資料, EC 78-23 (1978)
- 6) 森ほか: 日本語文書を読取る印刷漢字認識装置, パターン情報処理システム講演会論文集 (1977) 33
- 7) 長谷川: 漢字処理装置, 情報処理, 19, 4 (1978) 350
- 8) 松倉: オフィスコンピューターによる漢字処理, 東芝レビュー, 33, 11 (1978) 943
- 9) 日本語ワードプロセッサ JW-10 カタログ, 東芝, 19, 10 (1978)
- 10) 長尾: 自然言語の理解, 情報処理, 19 (1978) 952