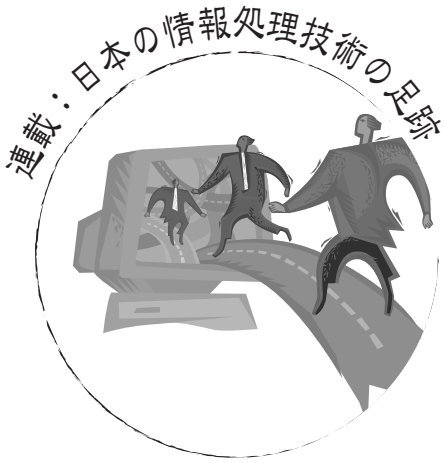

漢字・日本語処理技術の発展：日本語の入出力と処理

浦城 恒雄
東京工科大学メディア学部
uraki@media.teu.ac.jp

漢字・日本語処理の変遷

今日，コンピュータやPCでは数千種以上の漢字を高度のカナ漢字変換技術によって鍵盤から入力し，多種の輪郭フォントを用いて多種のサイズでインクジェットプリンタやレーザプリンタに出力できるが，これらは1990年代に入ってからようやく実現したものである．1950年代に入り米国を中心にコンピュータの商用化が始まったが，一般のデータ処理では大文字26字に記号を加えて47～63字あればよく，テキスト処理でも88～94字程度の文字種が扱えればよかった．これを前提にソフトウェアや入出力装置が開発された．

我が国においては1950年代の終わりに商用機が市場に出たが，その頃の入出力装置は電信系の穿孔タイプライタを改良した入出力タイプライタと光電式テープ読み取り機が代表的なもので，シフト機能によりカナ文字（48字）を扱うことができた．高速印刷に必要なラインプリンタはドラムの径を大きくしたり，チェーンやベルト方式の採用によりカナ文字を印刷できるようにした．以来1970年代の半ばまで一般のデータ処理の分野では英数字を主体にカナ文字を補助的に使う英数カナの時代が続いたのである．

しかし日本語の表記は漢字仮名混じり文であり，日本語表記が本質的に必要な分野で，漢字への取り組みが先行的に始まり，1980年代に入ると一般企業にまで広まっていった．利用の立場から漢字・日本語処理の変遷を概観してみよう．

■ 1950～1960年代：

1950年代の終わりに新聞社などにより日本語文の電信送受信装置の利用が始まった．1960年代の半ばから官庁および外郭団体の一部で日本語情報処理が始まり，印刷業界でも漢字自動写植システムの利用が始まった．

■ 1970年代：

1970年代半ばに入ると高速漢字レーザプリンタが登場し，保険業界など一部の民間企業において住所，氏名，会社名，品名，項目名などに漢字を用いる応用が始まり，地方自治体などにおいても住民サービスのため漢字の利用が始まった．

■ 1980年代：

汎用的な端末システム（ディスプレイとプリンタ）で漢字を容易に扱えるようになり，そのうちPC機能を取り込んでインテリジェントなオフィス端末として発展し，ようやく日本語処理がオフコンを含めてコンピュータの当たり前の機能となった．またワードプロセッサ（以下ワープロと呼ぶ）が普及し，社内文書や個人文書の作成に広く利用されるようになった．PCも普及し始めたが国産PCは漢字フォントをROMとして持って日本語機能を実現し，国際仕様となったPC/ATとは互換性はなかった．

■ 1990年代：

1991年に漢字フォントをソフト的に処理するPC/AT互換のDOS/Vの共通仕様が発表されるとNEC（1997年に転向）を除く国産主要PCメーカはDOS/V路線に転向した．海外PCメーカの日本市場への参入が始まり，PCの低価格化が進み，企業のみならず個人への普及が進んだ．オフィス端末は次第にPCに吸収され，日本語処理を担う主役はPCとなった．

本稿では日本語処理が本格化（1980年代以降）する以前の歴史を振り返り，漢字仮名混じり文である日本語の入力および出力（主としてプリンタ）にどのように取り組んできたかを述べてみよう．

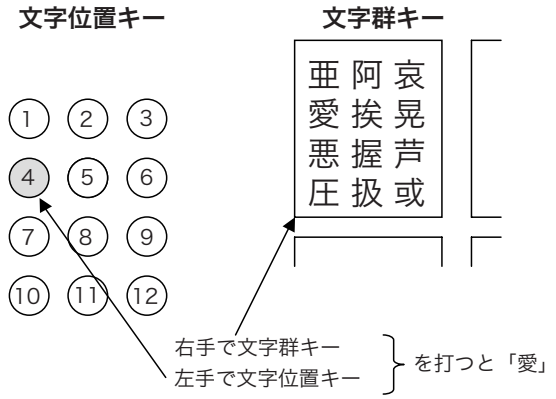

図-1　多段シフト方式の原理

夜（ヨル）	山（ヤマ）	読み
母（ママ）	髪（へ a）	意味，外来語
公（ハム）	化（i ヒ）	字の形
皮（ヒフ）	意（i ミ）	熟語

図-2　連想コードの例

漢字・日本語の入力

多段シフト方式

漢字が扱える入出力装置は新聞社でのTTS（Tele-Typesetting System）のための漢字電信印字機（漢字テレタイプ．通称，漢テレ）から始まった．読売新聞社が防衛庁と共同研究を行い，日本飛行機製作所に発注したものが1954年に完成した．続いて1955年朝日新聞社と新興製作所との共同研究による試作機が発表され，1958年に製品化されて広く新聞業界で使われた漢テレの原型となった．その後沖電気も参入した．

この漢テレに用いられた入力方法は多段シフト式のキーボードである．1つのキーに3列×4行の12文字を入れ，キーを24列×8行配置すると2,304字を扱うことが可能で，右手で文字群のキーを，左手で文字群の中の所望の文字の位置を選択するキーを打鍵する（図-1）．熟練者で70～100字／分の入力が可能であった．1960年共同通信社による加盟地方通信社へのニュース記事配信のTTS化が始まり，全国の新聞社で使われるようになった[1)]．

一段選択方式

■和文タイプ方式

従来からあった和文タイプライタに文字コード発生機構を付加したものである．この方式では文字の選択に機械的移動を伴うため手や腕の疲れを生じ，打鍵速度も30～50字／分とあまり速くなかった．文字コードは活字位置から機械的に発生させたが，位置に関係のないコードを付与するために活字にバーコードを付加してバーコード読み取り機で読み取る方式が考案された．

■ドラム（漢字表）方式

1965年に漢字の表をドラムに巻きつけ，左手でハンドホイールを回しながら左右に移動し，所望の文字のところに針を持っていき，右手でキーを押すとコードが発生するドラム式漢字鍵盤送信機が沖電気により開発された．類似の方式のものが東芝タイプライタからも製品化された．

■タブレット（ペンタッチ）方式

1字あたり数ミリ角の文字を60列35行程度並べた文字盤上で，文字をペン状のもので軽くタッチすると電子的にコードを発生する機構を持つもので，1970年代に入ると各社で開発され，この時代の代表的な漢字入力装置となった．この方式は片手操作が可能で，タッチも軽く疲労も少なく素人向きであったが，入力速度は30～70字／分と多段シフト方式と比べるとかなり遅い．コード発生の方式としては電磁結合方式（日立，東芝など），静電結合方式（ぺんてる，沖電気，富士通など），光電方式（日電漢字システム，三菱など）があった．後に圧電シートや感圧導電シートを用いたものも開発された．

2ストローク方式

漢字をカナやアルファベット2文字の組合せと1対1対応させたコード体系を作り，2ストロークの打鍵で入力する方法で，1972年に開催された日米コンピュータ会議で，ラインプット社の川上晃（裁判用速記タイプとして広く使われたソクタイプの発明者）によって最初に発表された．コードを覚えやすくするために連想的な対応付けを多く用いた（図-2）．

効率のよいタッチ法による入力を可能にするために左

右の手による交互打鍵や各指の負荷の合理的分配などを考慮した独自のキー配列になっており，100 ～ 125 ／分という高い入力速度を可能にした．しかしラインプット社はこのコード体系や教則法を非公開とし，自社以外に利用させなかった[2]．

鍵盤としては通常のカナ鍵盤を用い，コード体系を公開した入力方式が，新興製作所，カンテック，九段コンピュータセンタなどにより開発された．

この方式は教育訓練でコードを覚えた熟練者の入力に適しており，大量の漢字入力を行うセンターなどで利用された．タッチ法の推進に熱心なグループにより強く支持され，初期のワープロの一部にカナ単漢字変換と併用して採用されたが，誰しもが手軽に使えるものではなく，カナ漢字変換の高性能化につれてほとんど使われなくなった．

カナ漢字変換方式

今日最もよく使われているカナ漢字変換方式の原型は九州大の栗原らの研究が最初といわれ，1963 年に特許出願された．カナ文を文節分かち書きで入力し，単語辞書による照合，構文解析，意味解析などカナ漢字変換に必要な基礎的手法を提案した．

この研究をベースに沖電気の黒崎らが 1967 年にカナ漢字変換システムを試作した．1971 年日本ソフトウェアの藤井らは外電のローマ字電文を対象にした変換プログラムを作成した．カナ漢字変換で問題になる同音異義語に関しては 2 つに絞って併記し，扱う人に判断させる方式であった．1973 年 NHK の相沢らはニュース文に限定した実験システムを試作した．

1970 年代後半に入ると高性能化したミニコンピュータを利用し，変換の効率および精度の向上に向けた研究が大学や企業の研究所において本格化した．1976 年東芝の河田らは単語辞書の学習的構成法を提案し，分野を限らない一般文章を対象とした実験システムを開発し，引き続く日本語ワープロ開発の母体となった．1978 年 9 月の東芝による JW-10（価格 630 万円）の発表を皮切りに各社がカナ漢字変換方式によるワープロを製品化した．

カナ文字列を漢字文字列または漢字仮名混じり文字列に変換しようとするとき完全な自動化は不可能で，できるだけ正しい日本語に自動的に変換したのち，選択あるいは修正する必要がある．変換を漢字単位の辞書を用いて行う単漢字変換方式では取り扱う漢字に対応して数千字分の辞書があればよいが，同音異字の出現頻度（ある調査では 1 字に対して 13.6 字）が高く，自動化の成功率はきわめて低く，入力者による選択が前提となる．

これに対して熟語単位の変換方式の場合は同音異義語の出現頻度（ある調査では平均 2.3 語）が低く，文法的処理などと組み合わせることによりさらに精度を上げることができるが，熟語辞書が少なくとも数万語は必要である．さらに辞書との照合方式が問題となる．最長一致法といわれる方式は辞書の見出し語と入力文字列とを比較し，一致する最長の語を見つけて変換の候補とする方法で，最長が見つかっても次の文字列との接続条件（動詞のあとの活用変化など）が不適当であれば次に長い一致語を候補とする．

この方式に対しあらゆる可能な一致語について次の自立語や付属語との接続条件を調べる総当り法がある．いずれも同音異義語をいかにうまく選択するかが中心的課題であり，頻度や使用状況による優先順位の変更や文法的処理（自立語と付属語との接続関係の規則性を利用）や意味論的処理（複数の文節を対象に単語の意味的なつながりを利用）による精度向上の研究が行われた[3]．

カナ文をいわゆるべた書きで入力すると入力速度は速いが変換率が悪い．変換処理をしやすくするためには入力段階で工夫をする方法が検討された．文節単位にスペースを挿入する文節分かち書き方式に加えて，漢字部を［ ］でくくる漢字指定方式や，ローマ字表記の標準である単語単位にスペースを挿入する単語分かち書き方式や，自立語と付属語との間にもスペースを入れる自立語分かち書き方式などが提案された．制限を付ければ付けるほど変換処理は容易になるが，入力者にとっては考えながら打鍵する必要があり，タッチ法から程遠いものになり入力速度が落ち，誤入力率が高くなることからトレードオフの難しい課題であった．

同音異義語の処理に次ぐ重要な処理に特殊文節処理といわれるものがある．日本語は接頭語，接尾語に加えて 2 つ以上の熟語をつなげて複合語を作る能力に富んでいる．複合語や接辞（数に接する「円」，人に接する「様」，地名に接する「市」など）付の語をすべて単語辞書に登録するわけにはいかず，特殊な処理が必要になる．また数詞や固有名詞の扱いも特殊処理に含まれる．接辞をグループに分けて単語と接辞の連結関係の強弱によって優先処理をする方法が，1974 年沖電気の松下らによって提案された．

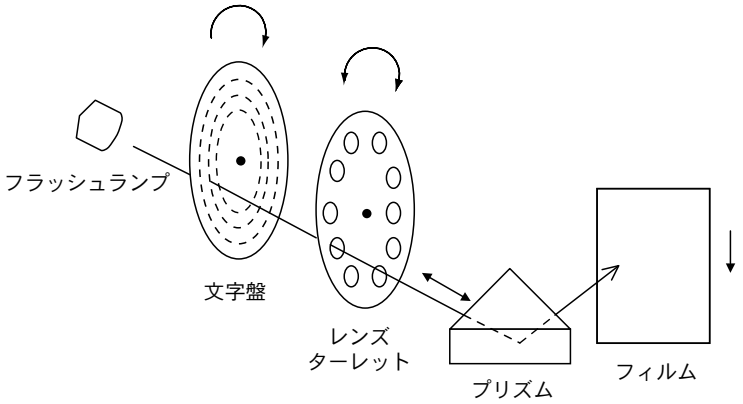

図 -3　移動文字盤方式の原理

パターン認識入力

文字認識や音声認識などパターン認識技術による入力についても研究が進められた．音声認識は 1970 年代の終わりには特定話者の単語登録方式で 100 語程度を認識する装置が実用化されたが，特殊な環境でのデータ入力としての限定的応用にとどまり，日本語入力としては程遠いものであった．文字認識には印刷文字認識と手書き文字認識があるが，1970 年代に入るとタブレット盤上で文字を書くと直ちにオンライン的に認識するオンライン文字認識方式の研究がいくつかの研究機関で進められた．文字を書くときの筆順を把握しやすいので通常の文字認識よりも認識が容易で，比較的簡単なアルゴリズムで 1,000 ～ 2,000 字の認識が可能であり，誤認識に対してその場で再入力ができるので CAD 図面の入力などで実用化された．

漢字・日本語の出力

漢字はアルファベットやカナに比べ，文字の種類が非常に多く，1 文字のパターンが複雑であるので漢字出力においては文字フォントの発生法が重要であり，そのコストは 1970 年代までは漢字処理システムにとって大きな比重を占めた．そのため文字パターンの圧縮や文字フォント発生装置をシステム（複数台の漢字ディスプレイや漢字プリンタ）で共用する方法がとられた．文字フォント発生の方式としては字母型を用いるアナログフォント方式とドットまたはストロークで漢字を表現するディジタルフォントに大別される．半導体メモリの低価格化に伴い，次第にディジタルドットフォントが主流になった．

アナログフォント方式

■活字方式プリンタ

最初の漢字入出力装置である漢テレは新興製作所と沖電気によって商品化されたが，いずれも約 2,500 の文字を扱うことができた．新興製は一定速度で回転するタイプホイール（活字輪）を使用し，所定の活字が印字位置にきたときに印字する．沖製はタイプパレットケースに各々復帰スプリングを持つ活字を 4 段 24 列入れ，これを放射状に組み立てたものを差動歯車を使った加算機構によって所定の活字を選択し印字位置に持っていって印字する方式であった．印字速度は共に 130 ～ 160 字／分であった．

■移動文字盤方式プリンタ

1966 年写研が SAPTON-N という写真植字機を開発した．漢字を 12 列同心円上にネガの形式で配列して文字盤とし，回転中に印字すべき文字がきたとき電子的にフラッシュランプを点火し，35mm の穴なしロールフィルムに写植する方式で，2,304 字種で 300 字／分の能力を持つ国産最初の自動写植機であった（図-3）．

1967 年日立が米国 ITEK 社より技術導入して開発した H-8247 型プリンタは原理的にはチェイン式ラインプリンタのチェインの代わりに多字種の文字を縮小して収容したフィルムベルトを使用し，ハンマーの代わりにキセノンランプの閃光をライトガイドで導いてエレクトロファックス用紙に光学的に印刷する方式である．180 行／分と当時としては非常に高速で，外務省に納入され情報検索業務に使用された [4]．

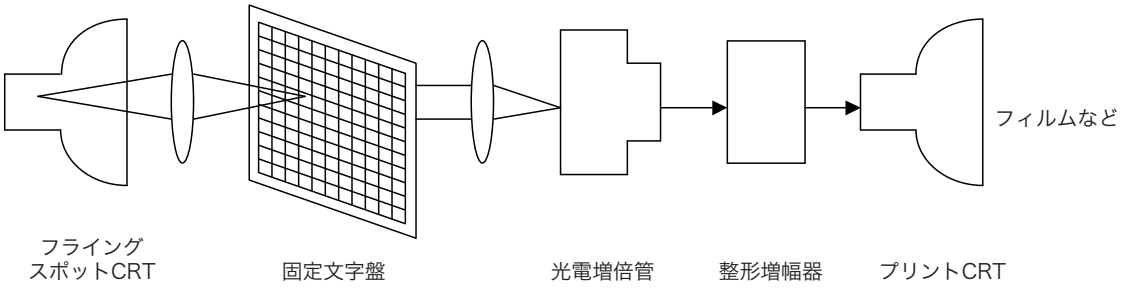

図-4　固定文字盤方式の原理

■固定文字盤（フライングスポット）方式プリンタ

文字盤における文字の選択を電子化して速度を上げることは可能で，1967年日本電子産業（JEM）によって開発され，日本科学技術情報センター（JICST）に納入されたJEM-3800漢字プリンタはフライングスポット方式を採用した．400～1000種の文字を800mm角の文字マトリックス（高解像度写真乾板を使用）に収容し，フライングスポットCRTのビームが文字コードに対応した文字マトリックス盤上の文字を選択するようにX-Y偏向回路を制御し，CRTのビーム操作によって選択された文字の形の光信号は光電増倍管により増幅され，整形回路を経てプリントCRT上に1字ずつ記録され，フィルムまたはエレクトロファックス紙にプリントする方式である．フライングスポット系を増設することで文字種を増やすことが可能であり，2,688字を最大3フォントまで可能であった（図-4）[5]．

ディジタルフォント方式

文字のフォントをディジタル的に記憶する方法としてドット方式とストローク方式があった．輪郭方式は1980年代半ばから写植機などで使われはじめ，1990年代に入って広く用いられるようになった．

ドット方式は，文字を格子状のドットマトリックスで表現するもので，当初は記憶装置が高価であったことから，何とか漢字が表現できる最小限のドット数として，横×縦が15×18や16×18のものが用いられた．しかし，当用漢字でも一部の文字は略式表現にならざるを得ず，ディスプレイでは許容されるとしてもプリンタ出力する最終文書としては問題視された．24×24以上を用いると曇や鷹のような画数の多い字も表現でき，明朝のような太さに変化のあるフォントもかなり表現できる．一般には32×32で十分であるが，活字に近い品質を得るには64×64以上のものが必要であった．1970年代に入ると主流の方式となった．

ストローク方式は文字を直線の集合として記憶する方式で，1文字あたりの記憶容量はドット方式に比べてやや少なく，文字の拡大，縮小，回転などを演算処理で容易に行えるが高速のドット展開が必要なものには不向きで，プロッタや蓄積管ディスプレイやベクトル機構を持ったグラフィックディスプレイなどで用いられた．

■静電式プリンタ

1967年富士通により開発され朝日新聞社に納入されたFACOM6501Aプリンタは1文字を15×18ドットで表現し，磁気ドラムや固定記憶装置に記憶しておき，270本の細いピンを1列に並べ，ドットに対応して各ピンにパルス状高電圧を与えて絶縁処理を施した特殊用紙上に電荷の潜像を作り，粉末インクで現像し，熱定着する方式である．2,688字を扱い，15字／行で93.7行／分の速度であった．

1967年共同通信社が試作し，1969年東芝によって製品化された漢テレファックスは1文字を24×24ドットで表現し，2,592字が固定記憶装置に記憶され，静電マルチスタイラストナー方式による電子印刷を行い，速度は当時普及し始めた200bit／秒の伝送に対応して500字／分であった．

■ワイヤドットプリンタ

複数本の細い金属製のワイヤを電磁アクチュエータで駆動し，ドットマトリックス文字を印刷する方式で，1960年代初めにIBMのカード用プリント機構に採用されたが，1970年代に入って米国Centronics社のModel101が発表されるとインパクト式シリアルプリンタの分野で急速に使われるようになった．英数字の場合は7本のワイヤを用い，5×7ドットの文字を165字／秒の速度で印字した．この成功に刺激されて国産各社で18本のワイヤ

を用いた 16 × 18 ドットの漢字プリンタが開発された．数枚の複写が取れるメリットもあり，比較的低価格の漢字プリンタとして漢字・日本語処理の普及に大きな貢献をした．ワイヤを 12 本ずつ 2 列に千鳥状に配列し，タイミングをずらして印字することにより 24 × 24 ドットで 40 字／秒程度のものが 1970 年代の終わりには広く使われるようになった．

■インクジェットプリンタ

1976 年 IBM が発表した 46/40 型インクジェットプリンタは，インクのジェット流を一定電界中に通しておき，インク粒子の荷電量を制御してインク流を偏向させて文字を描く方式で，77 字／秒で印刷できる高速高品質低騒音のシリアルプリンタとして注目された．引き続き日立，東レなども製品を開発した．荷電量制御方式のものは連続インク粒子流を噴射し，印字しないものは回収する機構が必要であるが，東レが採用したインクオンデマンド方式は，速度は遅いが回収機構が不要であり，その後マルチノズル化による速度向上も図られ，多色印刷の道も開けてインクジェットプリンタの主流の方式となった．

■レーザ（電子写真式）プリンタ

1975 年 IBM によりレーザと電子写真技術を用いた 3800 印刷サブシステムが発表され，その後のレーザプリンタブームの引き金となった．電子写真プロセスは複写機の世界で Xerox により古くから実用化されていたが，強度変調したレーザ光を回転鏡を用いて走査して感光ドラム上に文字パターンの潜像を作り，帯電トナーを吸着させ紙に転写したのち熱定着させる方式である．ラインプリンタと同様の連続普通紙を利用でき，英数字では 1 万行／分を超える高性能であった．発表当時は漢字データの印刷はできなかったが横方向には 240 ドット／インチの印字密度を持ち，文字フォントさえ拡大できれば容易に漢字プリンタになるもので，1977 年に漢字機構が付加された．

これに刺激されて日立は日立工機と共同で開発に着手し，1977 年に 7,000 行／分（英数字の場合．漢字も印刷可），1979 年には 15,000 行／分のモデルを出荷した．富士通と日電は西独 Siemens 社のものを導入して漢字化を行った．これらの高速漢字ラインプリンタの登場により，事務処理の分野における漢字の導入が本格的に始まった．

一方キヤノンはカットフォームの用紙を扱う低速高品質の小型レーザプリンタの開発に取り組み，1976 年には湿式プロセスを用いた高品質機（288 ドット／インチ）を開発し，引き続き低価格機を狙い半導体レーザと乾式プロセスを用いた LBP-10 を製品化し，その後端末システムや OA の分野に小型レーザプリンタが広まるきっかけとなった．

おわりに

漢字・日本語処理の歴史を振り返るとき 1978 年は画期的な年で，JIS「情報交換用漢字符号系」の制定，東芝による日本語ワープロ「JW-10」が発表された年である．またコンピュータの主要メーカから漢字・日本語処理の体系化が発表された年でもある．この年を機に 1980 年代以降本格的な普及に向けて歩み出し，今日に至ったといえよう．漢字・日本語処理の揺籃期ともいえる 1980 年代以前において，最も大きな問題であった漢字仮名混じり文の入力と出力という課題への先人たちの取り組みを紹介した．

参考文献

1) 安田寿明：我が国の新聞社における漢字情報処理，情報処理，Vol.10, No.5, pp.340-347 (Sep. 1969).
2) 川上　晃他：タッチ法による漢字入力，情報処理，Vol.15, No.11, pp.863-867 (Nov. 1974).
3) 森　健一他：かな漢字変換，情報処理，Vol.20, No.10, pp.911-916 (Oct. 1979).
4) 長井　担：漢字入出力装置の動向と技術的問題点，情報処理，Vol.10, No.5, pp.320-332 (Sep. 1969).
5) 長谷川実郎：高速漢字プリンタ，情報処理，Vol.10, No.5, pp.279-284 (Sep. 1969).

（平成 14 年 9 月 9 日受付）