計算機による日本語の処理*

高橋達郎** 広田広三郎**

1. はじめに

計算機の応用分野が広がり，言語情報処理の比重が高まってくると，当然問題となってくるのは日本語の処理である．これは広くいえば各国語の問題であり，計算機の普及浸透とともに，特にアルファベットと異なる言語を使用する国々における共通の問題であろう．これはヨーロッパにおいて発達した近代印刷術が各国に導入され，それぞれの国において独自の自国語用のシステムが作り出されたことと同様と考えられる．また，印書機械にしても，現在各国で自国語用のタイプライタが開発使用されている現状である．現在の計算機の機能をもととして，日本語をそれに合わせるべく改変するより，日本語に合わせた機能を開発すべく努力した方が，計算機側にとってもはるかに有利なことであろう．

特に，マン・マシン応答の普及により，不特定多数の利用者を対象とする場合や，アウトプットを多数複製(印刷)して配布する場合を想定すれば，日本語処理の重要性はますます重くなるものと思われる．これは欧米においても全く同じ事情であって，本来の言語により近い形を要求する声が高まりつつある．

機械翻訳・情報検索などの言語処理の研究段階においては，言語の形体よりそこに内在する法則の究明が主目的であるから，適当な記号系により変換して処理を行なえばよいが，これを実用に供する段階においては，当然本来の姿での処理が問題となってこよう．かかる意味において，日本語の処理はその実用面に重点が置かれる問題である．

日本語の処理において最も問題が多いとされているのは，その字種の多いことである．これがアルファベットと最も大きな違いであるが，これを分析的に考察して，その利害得失をみきわめることが，まず必要なことである．

2. 日本語と英語

情報処理という場合，処理の対象となっているのは表記されている言語ではなく，情報の内容である．したがって，同じ内容の情報を英語と日本語で表記した場合，どちらが計算機処理にとって有利であるかという点から考察することができる．

たとえば，情報という語をとると，日本語では「情報」は2字であるが，英語では“Information”で11字である．後に述べるように，漢字のコードとして1字12ビットを使用するとすると，日本語の場合は24ビットであるのに対し，英語の場合は1字6ビットとして66ビット必要となる．一般の文については，大量の英文和訳・和文英訳を行なっている当情報センターの平均値として

日本語：英語　　1字：2.5字

という値が出ている（筆者がある技術文献について調べたところでは1：2.8である）．この場合においても，ビット数でいって日本語が有利である．

次にインプットについては，英文タイプライタのタイプ速度は熟練者で大体

200～240 字/分

漢字テレ・タイプライタで大体

60 字（漢字）～120 字（ひらがな）/分

である．前述の字数の率を乗ずれば大体インプット速度は同程度とみてよいであろう．英文タイプは素人でもある程度の速度に達するので，過大評価されがちであるが，専門のパンチャで比較すると，上述のようにほぼ同程度である．この問題は植字におけるモノタイプの速度の比較によってもみることができる．ある新聞社の例によると

邦文モノタイプ　　100 字/分

ライノタイプ　　180 字/分

で日本語の方が速くなっている．

アウトプットについては，印字の品質の問題があるので一概にはいえないが，現在開発されている日本語の高速プリンタは，中速のラインプリンタ程度の印字速度は持っている．将来，プリンタは電子式のものが

---

* Rrocessing of Japanese Characters by Computer by Tatsuro Takahashi and Hirosaburo Hirota (The Japan Information Center of Science & Technology)

** 日本科学技術情報センター

普及するであろうから，その時点では英語も日本語も同等と考えてよいであろう．

現在欧米では計算機による文の処理，たとえば編集・植字や索引の編集が急速に普及しつつあるが，以上の考察からも明らかなように，これと同等の処理を日本語で行なうことは，決して不利でないといえよう．

3. 漢字，かな，ローマ字

3.1 かな書き文

言語はそれが話された場合でも書かれた場合でも，同じ情報を伝達しうるはずである．特に，最近のように文章の口語化が一般化されてくると，話し言葉と書き言葉は表現のニュアンスに多少の違いはあるにしても，同一とみなしてよい．文章として書かれたものを朗読して聞き取ったとしても，何の抵抗もなく受入れられる．しかもその文中には漢字で書かれた語が多数混在し，それが音読みされているものが多い．このように音声で情報が間違いなく伝達されることを前提として，文字の問題を論ずることができる．

かな書きとして一般に提案されているのはカタカナ書きである．カナ書きに対する反論としておもなものは

（1） 分かち書きが必要である．
（2） 読みにくい．
（3） 同音異義語の区別がつかない．

などがあげられる．分かち書きとは表音文字のみで綴った場合，語の区切りを明示するため分けて書くものであるが，問題となっているのは，分かち書きの規則が一定していないことである．最も実用的といわれる東大システムの原則は

1. 自立語は分けて書かない．
2. 付属語で種々の品詞の自立語につくものは独立させて書く．

の2箇条のみであることから推して，分かち書きの規則を明文化するのは非常に困難なことであろう．なお漢字かなまじり文では，漢字が視覚的に区切りを行なっており，また話し言葉では適当な区切りとアクセントで解決している．

読みにくさの点は，まず，なれていないことがあげられる．たとえば

アイスクリーム

とカナで書いた場合，このように普及している語に対してはほとんど抵抗がない．また，現在では文章はひらがな書きが一般であるから，これのカナ書きも一つの大きな原因である．したがって，かな書きにあたって，カタカナとひらがなを共用するのも一法であろう．外来語（漢字の音読みを含む）はカナで表示するなどの方法も考えられる．

同音異義語の問題は話し言葉の場合と同じである．人間が理解するときは前後の文脈から判断してほぼ混乱はないが，機械処理を行なう場合には，機械に前後の文脈を判断させなければならない．たとえば「キカン」という綴りに対して

帰還，機関，気管，汽罐，器官，期間，季刊，
旗艦，貴官，貴翰，……

など 10 以上もの同音異義語が存在する．漢字で表現されていれば文字の比較で識別できるものが，カナ書きではそれぞれ文脈を調べなければならない．一般にカナ書きの場合，この識別を容易にするため，語の書き変えを行なっている．

創造する→作り出す
想像する→考える

このように同音異義語が多数存在する理由は，漢字の字種に対して音が少ないことに帰因する．当用漢字について調べると

音読み延べ字種	1,924 字
音	220 種
平均同音異字数	8.7 字/音

となる．最も多いものはコウ，ショウの 61 字で，コウショウなる語は 40 以上にも達する．これに反し訓読みの場合は

訓読み延べ字種	1,112 字
訓	971 種
平均同訓異字数	1.2 字/訓

であり，訓読みで置換することにより同音異義語は大幅に減少する．

以上カナ書きの問題点をまとめたが，単に情報を伝達するための文としてカナを用いる場合には，なれによって解決されるべき性質のものであろう．これに該当する調査として，小説文について調べた結果によると，1900 年には 1,000 字中約 40% が漢字であるのに対し，1950 年代には約 30% 弱へと減少していると報告されている[1]．この傾向は将来も続くであろうから，漢字の使用はますます少なくなるであろう．しかし，これは漢字の消滅とは別の問題である．

なお，漢字のかな書きについて一つの提案がなされている[2]．漢字を訓読みで使用する場合にはかな書きとし，漢字は音読みでのみ使用するというものであ

る．これは外来語としての漢語のみを漢字で表記するもので，現在欧米からの導入語をカナまたは原綴りで表記するのと同じ発想である．かかる表記法の利点としては

1. 漢字の訓読みを覚えなくてよい．
2. かな書きによる読みにくさは増大しない．
3. 送りがなの問題がなくなる．

などが考えられる．前述の漢字の減少の傾向も，多分にこの方向によって実現されているものと思われる．なお，漢字のあて字はかな書きにすべきことは当然である．

3.2 かな表記

次に漢字は表意文字であるといわれている．たとえば

配達，配給，配管，配車，配色，配水，配船，
配線，配属，配置，配電，配当，配布，配分，
配役，配列，……

なる語は，それぞれなんらかの意味でくばることに関係があり，また，それぞれの語の表わす意味は漢字から想像がつく．これは一般にいわれていることであるがその発生段階においては表意的であったとしても，現在においては意味が直観的に対応しなくなっている語も多い．

自然，科学，管理，事故，機関，……

この表意の意味が失われている代表的なものが固有名詞である．本来の音に漢字を当てたものは論外として，明らかに意味のあるものでも，その実体とはなんらの関係のないのが普通である．

東京，京都，日本，……

人名の姓にいたってはほとんど意味と関係がない．このように漢字が表意でないものについては，かな書きの方がむしろ正常であろう．人名の名については表意的ともとれるが，実体との関係でないことは明らかである．以上の論旨は外国の場合も同じであって，固有名詞に意味がないことを前提とすれば

トウキョウ，ニューヨーク
カトウ，スミス
カズオ，ロバート

など同列に扱ってよい．したがって，固有名詞はかな書きの方が合理的である．この利点をあげると

（1） 漢字の字種を人為的に押さえることができる．現在漢字問題で，最も困難なのは固有名詞の字種である．
（2） 難読の問題がなくなる[3]．
伊達(だて)，指宿(いぶすき)
神戸（かんべ，かんと，かんど，こうど，ごうど
こうべ，かのと）
（3） 分かち書きの問題がない．
（4） 配列の問題が解消する．

などが考えられる．なお，かな書きによって同姓同名が多くなることは確かであるが，致命的な問題ではない．

かなは字種も少なく，音と文字が対応しているので簡単に覚えられると一般にいわれている．しかし，小学生が言葉を覚えるのに，かなよりも漢字の方が早く覚えられ，小学1年生で200字程度の漢字を覚えるのは容易であるとの実験もある[4]．いずれにしろ，教育漢字程度の字種を小学校で覚え込ませるのは容易であろう．したがって，教育面からかな書きを推進しなければならない必然性は少ないようである．なお，日本語のかなは音と字が対応しているので，発音どおりに綴ればよいが，英語，フランス語などでは，語の綴りを覚えるのが学習上の大きな負担となっているとのことであり，この点日本語の方が漢字を入れても，学習が容易であるのかもしれない．

3.3 その他

かな書きと同じ理由でローマ字書きも提案されている．これも表音文字による表記法であるから，かな書きと同じ問題が存在する．ローマ字法はかなに対してローマ字綴りが一意的に対応しているから，ローマ字書きとかな書きとは全く同じことである．ローマ字で書いても国際的に通用するわけでないから，ローマ字書きのメリットは英文タイプを使用しうることのみであろう．この他かな書き，ローマ字書きに関しては種々の論点があるが，本論からそれるので触れないこととする．

字数については，各方式により同一文を表記した場合の字数の統計の報告から，結果を示しておく[5]．

漢字まじり文	1 字
かな書き文	2.06 字
ローマ字書き文	2.8 字

この値は入出力の能率に関するデータとして重要である．

一般に漢字まじり文というと漢字とひらがなとからなると解釈されがちであるが，最近では外国語のカナ書き表示が多くなり，また，科学技術文献では翻字されない記号などが多数含まれているので，上述の結果はそのまま適用はできない．当センターの抄録誌から

抄録を任意抽出して 33,563 字について調べた結果は次のとおりである．

	字数	%
漢字	12,585	37.5
かな	9,866	29.4
カナ	4,157	12.4
欧文字	2,535	7.6
記号	2,696	8.0
数字	1,724	5.1

また，中国からの外来語を漢語として，そのまま使用しているのと同じく，欧米からの外来語を原綴りのまま取り入れて，英字まじりかな書き文という形式も考えられる．実験の結果によれば，科学技術文献などでは，かな書きよりもはるかに読みやすさの点ですぐれていることを付記しておく．

漢字かなまじり文においては，漢字とかなの役割が大体定まっており，実質的な内容を表わす自立語は漢字，文法的な関係を表わす付属語はかなで表記されるのが普通であり，漢字の字種を制限する場合も，この方向に従って行なわれている．また，英字，カタカナも自立語とみなされる．インデクスの見出し語（キーワード）は，自立語の代表的なものであるので，ある索引誌を分析した結果を次に示す（キーワードの第1字の字種）．

ひらがな		3
カタカナ		80
英字		50
漢字	(音読み)	509
	(訓読み)	4
計		646

この結果から文の内容分析を行なう場合，漢字が一つのてがかりとなることがわかる．また，漢字の訓読みがほとんどないことからみて，訓読みのかな書きに一つの根拠を与えている．

以上漢字とかなの問題をひととおり述べたが，結論として，日本語を無理にかな書きにする理由も認められないが，また，漢字を制限する必要もあり，適切な字種を定めることにより，機械処理のメリットと読みやすさのメリットを高めることができよう．文字記号系の大きさの問題は，日本のみでなく世界共通，つまり人類に共通の問題であろう．人間の視覚による識別能力，記号の記憶能力，表記の経済性など多くの観点から考察を進めるならば，最適な大きさが求まるかもしれない．これは，2 進法，10 進法，12 進法などの問題と同じことである．

4. 漢字

4.1 字種

漢字の問題において，最も力点が置かれているのはその字種である．各種漢和字典に記載されている字数は次のとおりであるが，

大漢和字典（服部）	8,626 字
新漢和大字典（宇野）	9,111 〃
新修漢和大字典（小柳）	10,827 〃
新漢和大辞林（児島）	14,270 〃
大字典（上田 他）	14,924 〃

現在では昭和 21 年に告示された当用漢字（1,850字）が常用されている．字種に制限を加えずに使用することは，印刷の面から，また人間の学習および記憶の面から弊害が大きく，ひいては，文化の発展にまで影響を及ぼすおそれがあることは確かであろう．したがって，漢字字種を制限しようとする動きは古くから起こり，現在の当用漢字に至っている[6]．

大正 12 年	常用漢字表	1,962 字
昭和 6 年	常用漢字表（修正）	1,858 字
昭和 17 年	標準漢字表	2,528 字
	常用漢字	1,134 字
	準常用漢字	1,320 字
	特別漢字	74 字
昭和 17 年	標準漢字表	2,669 字
昭和 21 年	当用漢字表	1,850 字
昭和 23 年	当用漢字表別表（教育漢字）	881 字

漢字の制限に対してもちろん多くの反対があり，当用漢字を増加させよとの意見も多いが，それらの意見をまとめてみると，せいぜい 300 字程度の増加で満足させられるといわれている．一般的な固有名詞の漢字を考慮に入れても，3,000 字もあれば十分であるといわれる．この数字を一応の上限とみれば，漢字処理の問題にも目安がつく．

漢字を制限する方法としては，次のようないろいろな方法がとられる．

（1） かな書きにする．

鯉→こい，於ける→おける

（2） 意味的に同じか近い語で置き換える．

醍醐味→妙味，梗概→大要

（3） 意味的に同じか近い字で置き換える．

蒐集→収集，附録→付録

(4) 意味的に同じか近いかなの語で置き換える.

嘲弄→あざける，巷説→うわさ

特に(3)の方法は漢字の基本義に由来するもので，これを押し進めると，さらに一層の制限を行なうことができる[7]. たとえば

伏，服，副

はともに「くっつける」という意味であるから，これをすべて伏で代用させることも可能である.

洋服→洋伏，副官→伏官

漢字の字種で困難な問題は固有名詞（特に人名）である. 東京都の電話番号簿を作成するのに，11,000種の漢字を必要としている. これについては**3.2**でも述べたとおり，固有名詞はすべてかな書きで処理する方法が，いろいろの点からメリットがあろう.

字種の調査としては，国立国語研究所が行なった「現代雑誌九十種の用語用字」[8]が大規模なものであり，この結果については別稿で紹介されることになっているが，二，三の点に触れておく. これは昭和31年の雑誌について行なったもので，丁度当用漢字が告示されてから10年目の結果である.

当用漢字	1,850 字
度数8回以下	177
度数0回	15
表外漢字	1,493 字
度数9回以上	322
度数50回以上	28

この結果からみると，当用漢字の中から，さらに削除してもよい漢字が確かに存在することがわかる. たとえば，度数0の字は次のとおりである.

蚕，弐，丙，嗣，墳，弧，斥，朕，殉，璽，
疫，痘，諮，迭，陪

また，表外漢字としては固有名詞として表われたものが多いが，その他にも，たとえば

頃→ころ，云→いう，袖→そで，衿→えり
誰→だれ，廻→回，糎→センチ・メートル(センチ)
或→ある，筈→はず，頁→ページ

などが度数50回以上にはいっているが，これらは矢印右のように書き換えることにより，漢字を使用しなくてもすむ性質のものである（音読みがない）.

4.2 情報センターの調査結果[9]

本稿では科学技術関係についての調査結果を詳しく紹介しよう. これは日本科学技術情報センターが，その抄録誌（科学技術文献速報）の編集・植字工程を電算化するにあたり，入出力字種を決定するために行なった調査結果である. 文献速報各編から計1,297の抄録を選び，その中から99,771字の漢字を抽出して分析した.

抽出抄録数	1,297
延べ字数	99,771
	(当用 99,476，表外 295)
異なり字数	1,528
	(当用 1,378，表外 150)

もともと抄録は当用漢字以外を使用しないたてまえで行なっており，また，外国文献が主であるから，日本の固有名詞もほとんど含まれていない. したがって，ここに現われた表外漢字は，誤って使用したもの（稼，廻，迄，僅，殆など），および学術用語として慣用されているが当用漢字にないもの(嵌，勾，溝，漬，槽，滌，澱，脾，泡 など）が大部分である.

当用漢字で出現しなかったものは500字で，他の調査と比較すると多いが，これは抽出もれがみうけられること，科学技術という限られた範囲であるため，愛哀，慰，詠，恩，祈，喜，……などが出現しないことによるものである.

字種と延べ使用率との関係は，国研の調査による実用通俗科学雑誌および婦人雑誌と比し，漢字が集中的に使用されていることがわかる. 以上の結果をもととして，固有名詞を対象外として漢字字種を次のように決定した.

(1) 当用漢字 1,781 字

(2) 表外漢字 凸，凹の2字のみを認める. それ以外は適当な書換え，またはかな書きとする.

科学技術分野を対象とした字種調査としては，国立国会図書館が行なった「雑誌記事索引・自然科学編における漢字のひん度調査」[10]がある. これは1963年8月号について，科学技術編と医学編についてそれぞれ調査結果を示しているが，前者に関しては上述の結果とほぼ同様であるが，後者に関しては，医学特有の表外漢字が上位にランクされている（カッコ内は使用順位).

腫(18)，癌(28)，腎(51)，瘍(58)，娩(143)，
蛋(147)，腺(180)

上位500字中にかかる字種が37字出現している. 医学関係の漢字をどう取り扱うかは，今後の大きな課題である.

固有名詞用の漢字は，もちろん当用漢字の中に多数存在するが，それ以外にも使用率の高いもの，および

その性質上欠かせないものと（たとえば県名）があるので，別途調査を行なった．まず，人名については，日本化学総覧（日本文献の抄録誌）Vol. 39 の著者索引に記載された著者全部の姓のみを対象として調査した．表外漢字は計 270 字であったが，上位 20 位までの漢字を採択することとした．

岡，崎，塚，阿，柴，菅，垣，栗　など

地名については，地方名，県名および人口 5 万以上の都市名に用いられている表外漢字は全部採択した．このほか和雑誌に慣用されている表外漢字は採択することとした．

彙，輯，叢，纔　など

以上の結果をもととして，固有名詞，慣用語を対象として表外漢字を次のように決定した．

表外漢字　　78 字

これと，前述の当用漢字および凸，凹を合計して，計 1,861 字を採用することとした．

以上が情報センターで漢字字種を決定するまでの経緯であるが，抄録文をインプットするにあたって漢字がない場合には，かな書きにすることにしている．これは特に人名において問題となるものである．以上の結果からみて，現状で漢字処理をするにあたっての問題点を列挙してみよう．

（1） 漢字は，適用分野によって字種が大幅に異なる．

（2） したがって，各適用分野によって，効率を高めようとすると，独自の字種を選択しなければならない．

（3） これが独自のシステムを生み互換性を失わせている．

（4） よって字種とコードの統一を早急にはかり，漢字処理の普及の障害を取り除かなければならない．

（5） これをバックアップするため，人名・地名・法人名などの当用漢字，またはかなへの変換を促進させること．これと関連して当用漢字の入換えを行なってもよいのではないか．

4.3 字　体[11]

漢字処理にあたって字体もまた問題である．当用漢字の制定にあたって，字体も大幅に簡略化されたがまだ問題は多い．漢字を記号としてみれば，方向としては簡略化であろう．記号として混乱が生じない限度で，また，識別しやすい点から，ある程度の冗長性を持たせて決定することとなろう．現状をみてみると

日：目，水：氷

の区別に困難を生じないことからみて，思いきった簡略化は可能であろう．この場合，へん，つくり，かかり，かこみなどを整理して規格化すること，つまり，最小限の構成要素から組み立てることが必要で，これに関しては，OCR およびディスプレイからの要求を十分取り入れることが，将来の問題として重要である．

漢字の字体を別の面からみることもできる．**2.** において日本語と英語の比較を行なったが，それによると英語は日本語に比して，約 2.5 倍の字数を要することがわかる．この理由は英語がアルファベット 26 文字に対し，日本語が約 2,500 字であるため，日本語の 1 文字の意味内容が大きく，したがって，字数が少なくてすむことであるが，一方，別の面からみると，2,500 字を識別するために，字画は，はるかに多くなっている．したがって，文字を印刷した場合，アルファベットで 8 本，漢字で 18 本以上の解像力がないと明瞭に識別できないといわれている．よって，それぞれの文字で書かれた情報を，同じ明瞭度で記録する場合，英文字の方が 2.25 倍多く記録でき，前述の 2.5倍のデータと相殺されて，単位面積あたりの情報記録容量はほぼ同じとなる．英語における綴りの冗長度，および漢字の字画の冗長度が独立であるとすれば，この一致は偶然であるかも知れないが，同じ情報内容を記録するのに視覚に訴える記号を用いれば，任意の記号系を用いても，結局同じことになるとも考えられる．これはアナログ記録としてのマイクロフィルム，ビデオレコードなどでは重要な問題であるが，また，プリンタの問題でもある．いずれにしろ，字画を簡略化することは有利なことであり，簡略化の場合，現中国の略字は大いに参考となろう．

4.4 配　列[12]

漢字の配列も確定していない問題の一つである．これは索引誌・電話番号簿・人名録などのような，インデクスにとって重要である．現在一般に用いられている方法は，次のようなものである．

（1） 漢字の発音をかなで想定し，かなの五十音順に配列する．

（2） 漢字の発音をローマ字で想定し，アルファベット順に配列する．

（3） 漢字をその読みで類別し，五十音順に同一漢字をグループにまとめて配列する（電話番号簿）．

これらの方法はそれぞれ一長一短があるが，（1）と（2）は全く発音順であるから，各文字がバラバラに出

現し，眼で読む場合，各文字のパターンを識別に利用できない難点がある．また，漢字単位にまとまっていいなため，サーチしなければならない範囲が広く，探索に時間がかかる．この点(3)は漢字の特性を生かした有効な配列法である．しかし，(1)～(3)のいずれも発音が明示されていなければ配列できず，また，計算機で配列するにはかなをつけ，そのかなをキーとしてソートし，漢字を出力するという方法をとらなければならない．特に，発音は固有名詞の場合，とりわけ人名の場合，ふりがながついていなければ，ほとんど不可能である．

これらを解決するには，漢字と発音とコードを一意的に固定し，機械的にソートするより方法がない．これを英語の場合についてみてみると，全く同じ事情である．たとえば

it 〔it〕
item 〔áitem〕
iterate 〔itəreit〕
ivory 〔áivəri〕

のように発音順に配列されていない．全くアルファベットの字順であり，しかも，iに対して2とおり以上の音が対応している．これを漢字に適用すると，漢字の字順を一意的に定め，すべてその字順に配列することである．字順としては，画数・部首別・音順などいろいろ考えられるが，実用性からみて音順一画数順が適当と考えられる．これによりソートが機械的に簡単にできること，インデクスの場合，読み方がわからなくてもひけることなど，メリットが大きい．

5. 入力システム

計算機による日本語の処理において，最も論争の的となるのは入力システムである．漢字の入力装置としては，従来から新聞・通信社で漢字テレタイプライタが，情報の伝送用として用いられてきたが，英文タイプの場合には，字種が少ない関係上，その操作に抵抗は少ないが，漢字の場合には字種の関係上，その操作を容易にする工夫が種々試みられている．ただし，操作が容易ということと，能率がよいということとは別問題である．これはまた専任者（パンチャ）用の装置と不特定多数用の装置を，別個に開発しようということともみられる．漢字の入力には，次の各種の方法が試みられている．

（1） 漢字のパターンをそのまま用いて入力する．
（2） 漢字の構成要素パターンを用いて入力する．
（3） 漢字の音を用いて入力する．
（4） 漢字をコードに変換して入力する．

(1)は在来の漢字テレタイプの型で，字種の数だけのキーが必要である．したがって，キーの記憶に熟練が必要で，主として専任者用といえよう．キーの配列には二とおりあり，部首別・画数順など漢字の音に関係なしに，パターンとして配列するものと，音順に配列するものとがある．前者は漢字を知らないパンチャ(たとえば外人)にも操作できるが，後者は一応の漢字教育を前提としている．(2)は Chicoder がその代表で米国で考案されたものである．フレクソライタをそのまま用いてキーの数を限定し，漢字が読めなくても打鍵が可能である．(3)は，いわゆるカナ入力漢字処理システムといわれるもので，不得定多数の利用者を対象とし，将来のマン・マシン応答用として注目されているものである．(4)は，たとえば，数字コードに変換するもので，打鍵は簡単であるがコードへの変換が大変である．

5.1 漢字テレタイプライタ

従来から用いられている伝送用の入力装置であるが，現在ではモノタイプ用の入力装置としても利用されている．また，和文タイプを改良した簡単な装置も使用されている．漢テレの問題点は，鍵盤とコードの標準化および精度の向上であろう．

標準化はデータの交換を行なう新聞社・通信社にとって，特に重要なものであり，また，自動植字を行なう場合には，キャスタとも統一することが望ましい．この問題に関しては，昭和 34 年共同通信社が中心となり，新聞6社および漢テレ，モノタイプ・メーカが協議し，CO-59コードが定められている[13]．字種 2,304 字（予備 192 字），配列は使用度数によって4ブロックに分け，各ブロック内は部首別配列である．しかしこのコードが，全面的に各新聞社に採用されている状況ではない．

以上の標準化は漢テレ本来の利用法，およびホットメタルのキャスタを前提として定められたものであり，情報内容を機械的に処理することを前提としてはいない．漢字のソートとか，電算植字を行なう場合には，それに応じたコード変換が必要となろう．また，別の立場からすると，計算機処理を前提とする場合には，入力コードは統一されなくても，出力コードを統一しておけば，伝送上の問題はないとの意見もある．

前章でも述べたように，日本語の場合，文の内容によって字種は非常に異なっているので，編集・植字の

自動化のみを目的としたクローズド・システムにおいては，ある程度独自の方式を開発して，効率を上げることもやむを得ないと考えられる．当情報センターでも，抄録誌の自動編集・植字システムに用いる漢テレの字種を次のように決定し，文字の配列も独自に設定した[14]．

字種	字数
漢字	1,861
かな，カナ	158
英字	52
欧字	112
数字	30
記号	199
スペース	6
予備	78
計	2,496

ここに欧文字，記号が多いのは，科学技術の抄録文を植字するためである．しかし，漢字を当用漢字に限定し，しかも，その中から使用度数の少ないものを削除すれば，2,304 文字（または2,496字）ですべての文に共用できる字種とコードを設定することは可能であろう．ただし，固有名詞用の表外漢字は，度外視することが前提である．

漢テレの打鍵速度は熟練したパンチャにおいては

60 字（漢字）～120 字（ひらがな）/分

といわれ，新聞社ではこの程度の速度は確保されている．当センターのデータでは長時間での値として

18,000 字/日，50 字/分

となっている．この値は数式などを含む特殊な文であることを配慮して考察しなければならない．一般の文で平均的に

60～70 字/分

の速度とみてよいであろう．

漢テレのエラーとしては，打鍵に対して正しいコードがさん孔されないさん孔エラーが主で，某新聞社のデータによると，次のような値となっているが

ミスパンチ	10^{-3}
さん孔エラー	10^{-4}

初期不良ははるかに大きい値が出ている．いずれにしろ計算機の入力装置として用いる場合には，精度の向上が望まれる．また，現在漢テレを用いる場合，コードのビット構成は6ビット2列を用いているが，紙テープの読取りにあたっては，1字分のペアを確保することが絶対に必要で，このエラーを防ぐため，当センターの漢テレでは，1列目と2列目を判別するビットを設けてある．

5.2 カナ鍵盤漢字入力システム[15]

漢字の構成要素として音を用いる方式である．したがって，漢字が正しく読めることを前提としている．構成はカナタイプライタとディスプレイ装置からなりカナで日本文を入力すると，漢字まじり文となってディスプレイされるが，これを最終的にチェックして入力するものである．カナで入力する場合，次の三とおりの方法が考えられる．

（1） 入力形式に規定を与えない方式

カナ文字の単なる系列として入力する．この方式は漢字に変換する語を機械的に分離抽出することが困難で，実用的でない．

（2） 入力形式に十分な規定を与える方式

漢字に変換したい部分に前もって記号をつけるとか音訓で注釈をつけるなどして入力する．これは入力側の負担を重くするが，変換は容易となる．

（3） 入力形式にある程度の規定を与える方式

一般の分かち書き程度の規定を与えるもので，実用的であるが，分かち書きをあまりきびしくしてはならない．

次に多くの同音異義語から，目的とする語を選択し決定する問題が重要である．Chicoder と同じく，同音語すべてを表示して人間が選択してもよいが，ある程度のアルゴリズムで選択しうる部分は，計算機に負担させるべきであろう[16]．第9回大会の報告によれば機械的に 70%，人間が介入すれば80% を，誤りなり変換しうると報告されている．

本装置は大量データの入力用ではなく，計算機の端末として，マン・マシン応答用に使用すべきものであり，日本語による問合わせを可能とするものである．なお，ディスプレイの方式は出力装置と同じことである．

5.3 漢字コード入力方式

これは前もって漢字を数字コードに変換して入力する．変換はコードブックなどにより人手で行なうから漢テレによって直接入力するより，余分な労力がかかると思われる．利点としては安価な既製の数字せん孔機で入力できることである．この方式は，社会保険庁で採用されているもので，人名の照合がその目的である．前にも述べたとおり，人名の読み方はかながついていない限り非常に危険である．したがって，照合用としては，漢字のパターンと一義的に対応するコード

化が最も安全であるからである．また，この例では漢字の出力は行なっていないから，コード化法は字種の制限を全くうけないという利点が大きい．

コード化法の最大の問題点は，漢字のコードをもし暗記できるとすれば，この方式が最も有利となることである．これは実験によってたしかめる価値がありそうである．

以上入力システムの現状を簡単に紹介したが，将来の問題として，主として漢字の使用度数をもととして字種をさらに制限するとともに，字体を極力簡略化し，使用漢字の標準化をすすめ，低コスト高能率の機種を開発するとともに，OCR への道を開くよう努力すべきであろう．

6. 出力装置

6.1 現 状

日本語の処理において，入力システムとともに出力装置も今後の重要な開発課題である．現在，計算機では，一般にラインプリンタが用いられているが，これとの大きな違いは，字種が 20～30 倍も多いことであり，この字種で同程度の速度を得るには，メカニカルなプリンタではほとんど不可能なことである．従来から漢字プリンタとしては，機械的に印字する漢字テレプリンタと，写植方式によるものとが実用に供されてきたが，いずれも印字速度が 2～5 cps と遅く，計算機の出力装置としては不十分なものであった．

わが国において日本語の出力装置が問題となり始めると時を同じくして，欧米においてはディスプレイ装置と計算機植字が脚光を浴び始め，結局のところ，これらは同一レベルの問題であることがはっきりしてきた．計算機植字とは印刷工程における編集と植字の工程を計算機により自動化する試みで，新聞社のように迅速性を第一とするシステム，電話番号案内局のように改訂を繰り返すシステムにおいては，強力な武器となる．この場合，これらシステムはあくまで印刷であるから，従来の印刷物と同じ品質を要求され，しかも字体（ローマン，ゴジック，イタリックなど）やポイントの変化を要求される．したがって，字種も欧米で 1,000 字程度は必要となり，速度もラインプリンタ程度のものが要求される．かかる要求を満たすべく開発されたものが，Photon Zip 900, Linotron, Videocomp などの装置であるが，上述の理由から，これら装置はフォントを変えることにより，日本語の出力装置としても可能性があるわけで，近く日本語用としてわが国に輸入されることになっている．

現在わが国で開発されている高速漢字プリンタは，すべて電子式のものである．また，漢字ディスプレイ装置も開発されているが，この画面を写真にとれば，これもまたプリンタということができよう．漢字プリンタは，文字の品質によって 2 種類に分けることができる．一つはモニタプリント，またはゲラ用の装置でラインプリンタとほぼ同じ用途を目標としたものである．したがって，品質を犠牲にして低コスト，高印字速度を目標としている．1 例として F 社で開発されたプリンタの性能を次に示す[17]．

字　　種	2,688 字
文字の大きさ	3.8×4.3 mm
印字速度	100 行/分 (15 字/行)
文字パターン	15×18 (270) ドット/字
記録方式	電極静電記録

他はいわゆる高速植字機といわれるもので，まだ活版または従来の写植に匹敵する品質のものは開発されていない．当センターでも抄録誌の自動編集・植字システムの一環として，高速植字機の性格を持つプリンタを導入しているが，その性能は次のとおりである[18]．

字　　種	3,071 字（明朝，ローマン，ゴシック，イタリック）
文字の大きさ	6, 7.5, 10.5 ポイント
印字速度	200～600 cps
解像力	20 本/mm 以上

印字の質はまだ従来の植字のそれには及ばず，特に字並びに改善の余地が残されている．この点については，米国の高速写植機 Videocomp などは，従来の印刷と同等またはそれ以上の品質を得ている点からみて，今後の開発がまたれるところである[19]．

7. むすび

本稿では計算機による日本語処理の現状と問題点を特に漢字，入出力装置について概観した．詳細については，別稿でそれぞれ専門の諸氏が執筆されるが，全体の見通しとして役に立てば幸いである．計算機と日本語は，現状においては大きな異和感があるが，計算機が真の情報処理システムとして成長するには，越えなければならない一つの関門であろう．

入出力装置については，今後の技術的開発努力によって解決される問題であるが，日本語そのものに内在する不合理性は，一朝一夕に解決するとも思えない．常識的な方法ではあるが，機械的処理が可能な分野ま

たは効果の大きい分野から，徐々に導入することにより認識を深めさせ，国語問題を合理的に改革するということであろう．カナ書きが合理化であるというような，理論の飛躍におちいることなく，分析的に解明しなければならない問題が，まだ未検討のまま山積している現状である．

参考文献

1) 安本美典："漢字の将来"，言語生活 (137)，46～54，1963.
2) 平山健三："漢字のかな書きについて"，言語生活 (156)，68～73，1964.
3) 東京電話番号案内局編："難読姓氏辞典"，pp. 521，1966；荒木良造編："名乗辞典"，pp. 306，1960.
4) 宇野精一："漢字の問題"，続日本語を考える，77～118，1969.
5) 木沢 誠："情報処理機械の立場からみた印刷技術"，印刷雑誌，49 (9)，2～8，1966.
6) 文部省編："各種漢字表 字種一覧"，pp. 219. 1968.
7) 藤堂明保："漢字の意味"，続日本語を考える，151～170，1969.
8) 国立国語研究所編："現代雑誌九十種の用語用字(2)漢字表"，pp. 256，1963.
9) 森田 朗他："文献速報自動印刷システムのための入出力字種の選定"，第5回情報科学技術研究集会論文集，7～16，1968.
10) 国会図書館編："雑誌記事索引自然科学編における漢字の頻度調査"，pp. 47，1967.
11) 林 大："漢字の新字体について"，続日本語を考える，119～150，1969.
12) 西村恕彦：文字列の配列順序についての問題，情報処理，10 (1)，21～25，1969.
13) 滝沢 順："漢テレ・全自動モノタイプ方式の使用符号について"，印刷雑誌，49 (9)，8～15，1966.
14) 新興製作所：SCK-201形漢字けん盤さん孔機仕様書，1968.
15) 黒崎悦明："カナけん盤漢字表示方式"，情報処理学会第9回大会予稿集.
16) 栗原俊彦，他："仮名文の漢字混り文への変換について"，九大工学集報，39 (4)，659～664，1967.
17) 岩井麟三："電子計算機出力としての漢字プリンタ"，情報処理学会第9回大会予稿集.
18) "情報センターの自動編集機とその周辺"，印刷雑誌，51 (6)，2～5，1968.
19) 高橋達郎："電子計算機植字の現状"，情報管理，11 (5)，245～254，1968；"電子植字"，"印刷雑誌臨時増刊 52，1969.

(昭和44年4月14日受付)