和文入力に関する一考察*

保原　信**　山本　稔**

1. はじめに

日本語は，世界でも数少ない表意文字による言語である．平安初期から，漢字以外に，かな文字を併用するようになったが，中国文化との関連，あるいは漢字自体の持つ表意文字としての便利さ，造語能力などのために，漢字はその後も生き残り，現在では，教育漢字881字，当用漢字1,850字などの字数制限を受けたり，機械化文明に対処して，漢字廃止論，日本語のローマ字化なども聞かれるが，漢字の否定は日本語の否定，さらには，日本文化の否定につながり，むしろ最近の国語審議会などでは，制限字数のわくを広げるという意見が強くなりつつある，とにかく，漢字を使用すること自体には，将来も変わりがないと思われるのであるが，前述のとおり，日本語は機械による情報処理という面から考えると，欧州言語にはない固有の問題を多く持っている．たとえば，英語と比較しても，つぎに掲げるような特徴が，直ちに考えられる．

（1） 漢字を使うために，文字の種類がきわめて多いこと．

（2） 同音異字（異義）の単語が多いこと．

これらは狭い意味の情報処理という立場から考えたものであり，いわゆる，言語学的・文法的に考えるとさらに多くの特徴が挙げられよう．

近年，電子計算機の発達に伴って，種々の情報処理装置が登場し，あらゆる分野に浸透しつつある．そこで問題になるのが，日本語を含む情報処理の問題である．とくに，新聞社・出版社などのように，大量に，しかも迅速に日本語を処理しなければならないところでは，切実な問題である．現在，すでに高速漢字プリンタ，漢字電信テレプリンタ，自動植字装置などが実用化されているが，これらのシステムを使用する場合共通して問題になるのは，能率的な和文入力装置が欠けていることである．

パタン認識の実際的応用として，印刷体の漢字を読み取る機械については，1964年IBMで実験†されているが標準パタンを記憶するだけでも，膨大な記憶容量を必要としている．また，手書きとなると，簡単な制限手書き数字ですら，現在満足に実用機として使える機械は存在していない．そこで，将来の問題は別としても，さしあたり，とくに時間をかけて練習する必要がなく，だれもが気軽に，しかも能率的に使用可能な和文入力装置，あるいは和文タイプライタが強く望まれてくる．

2. 現用和文タイプライタの問題点

汎用和文タイプライタとして必要な文字の種類は，当用漢字，ひらがな，かたかな，英文アルファベット，数字，その他の記号を含めると，およそ2,150種に近い字数となる．現用のタイプライタは，これに幾分かの漢字を加え，ほぼ2,500字程度を用意しているが，これだけの文字を並べた中から，タイピストが所望の活字を選択印字するシステムとなっている．これを欧文タイプライタと比較すると

（1） 欧文タイプライタの場合は，10本の指を使用して打鍵するのに対して，和文タイプライタの場合，本質的に1本の指しか使用していない場合と同等である．

（2） 原稿と鍵盤の両面を視覚的に1つ1つ確認しなければならない（欧文タイプライタの場合には，原稿だけに視線が集中されている）．

（3） 熟練に相当の時間を必要とする．

などの欠点が指摘される．これらの原因で印字速度が

---

* A Note on a Method for Inputing the Japanese Sentences to Machine, by Makoto Yasuhara and Minoru Yamamoto (The University of Eledtro-Communications)

** 電気通信大学

† 漢字を少なくとも 25×50 の matrix に分割する必要があるという．したがって，漢字1字の標準パタンとして，少なくとも 1,250 ビット，すなわち，1語32ビットと考えると，39語の記憶容量を要し，仮に，2,500字程度の漢字を扱うとすると，標準パタンだけで，100K 語近い膨大な記憶容量となってしまう．

遅く非能率的でタイピストが非常に疲労しやすい．このようなことからも，少なくとも，現在の欧文タイプライタ程度の気軽さで，タイピングのできるシステムが完成すれば，実用としては充分であると考えられる．

3. 和文入力装置として必要な条件

いま仮に，和文タイプライタとして必要な漢字の字数を，多く見積って 4,096(=2^{12}) 字とすると，タイピストがこのタイプライタを用いて，1字印字するためには，12 ビットの情報を与えてやる必要がある．また，欧文タイプライタの字数を 64(=2^6) と考えると，この欧文タイプライタを用いて1字印字するためには6ビットの情報が必要である．したがって，もし，この欧文タイプライタ程度の鍵数を持ったシステムを考えれば，2回の打鍵を行なわないと，漢字 4,096 字のうちの1つを選択することができない．したがって，もし，漢字1字に対して，平均2回の打鍵を許すことによって，実用上充分な字数を，すべて指定できる欧文タイプライタ程度の鍵数を持ったシステムができれば，原理的にも全く合理的である．

そこで，たとえば，漢字を従来どおり，そのまま漢字として入力した場合と，全部“読み”で入力した場合の打鍵数を比較してみよう．漢和辞典を調べてみると，平均して後者は前者の2倍程度であるという予想は，一見して明らかである．このように考えると，漢字を“読み”で入力するという考え方が，きわめて合理的であることがわかる．

ところで，このようなシステムを考えた場合，まず第1に問題になるのは，一連の文章のうち，どの単語を漢字に変換すべきであるかを判断することであり，第2は前にも述べたように，日本語には，同音異義の単語（発音の際にアクセントで区別している）がきわめて多く，これをすべて漢字で区別している(むしろ，これは逆であって，漢字を使うために，同音異義が多いのであるが）ために，単語の意味，および文章の内容を理解できない場合には，“読み”を正確に漢字に置き換えることは不可能である．しかし，もし，このようなシステムが完成すれば，端末部としては，せいぜいかなタイプライタ程度でよく，タイピング速度も上がり，したがって，能率も向上し，さらに，タイピストの疲労も軽減され，誤りも少なくなることは論をまたない．

4. Human assisted compiling

前にも述べたように，このようなシステムを考える場合，2つの問題があった，その第1は，一連の文章の中から，漢字に置換されるべき単語を抽出することであり，第2は，同音異字をどう取り扱うかであった．第1の問題の最も実際的な解決策の1つは，タイピングの際に，漢字の部分に適当なラベルを付けて，他と区別してやる方法である．実際には，括弧でくくるなり，スペースを両端にそう入することになるが，これによって，タイピングに大きな支障をきたすおそれはあまりなさそうである．

第2の問題は，学問的にも興味深い問題を含んでおり，現在機械翻訳の研究者が，あえて避けている Semantics の問題に直結している．

FORTRAN, ALGOL, COBOL などと呼ばれる言語は，人間に理解しやすい表現法による Problem oriented language であり，これを機械語に翻訳するための Compiling software が必要である．これらの言語は，いわゆる Formal language であり，厳密な文法だけを規定し，したがって，この文法に従って正しく表現された文章であれば，翻訳プログラム自身だけで，正確に処理することができ，人間がこれに関与する必要は全くない．それは，これらの言語が Syntactics のレベルまでの問題を限定的に取り扱っているために，いわゆる決定論的な手法だけで処理可能であるからである．これに対し，Semantics の問題がはいると，いわゆる機械側だけの判断だけでは，処理不可能な問題が生じてくるために，人間が機械になんらかの形で関与するという Mode を，新しく考える必要が生れてくる．この考え方は，従来の CAD (Computer aided design) や，CAI (Computer aided instruction) などのように，人間側に主体性があるという考え方ではなく，Software 側に主体性があり，したがって，人間が手助けをするという意味で Human assisted (aided) compiling ともいうべき考え方である．

5. 読み-漢字変換のための Hardware System

いわゆる Hardware として必要な Computer の core memory の容量は，それほど大きなものが必要ではないと思われるが，これ以外に，辞書および中間結果，最終結果を記憶するための大容量のディスク，

あるいはドラム††，修正あるいは校正の際に必要な Soft copy のための日本語 display，および On-line かなタイプライタが必要である．このように Hardware としては，相当大がかりなものとなるが，これらのシステムは，この問題専用に用いられる必要は全くなく，明らかに他の一般的 Computer system と共通的であるため，すでに Computer system が用意されている場合には，それをそのまま利用し，必要な端末部をそろえるだけで充分である．さらに，TSS の技術を応用すれば，1台の Computer に，数台のタイプライタが，同時に access することには全く問題がない．

6. 読み-漢字変換の方式

6.1 入力方式

第4節ですでに述べたが，漢字としての単語・熟語を，どのように区別して入力するかという問題を，もう一度一般的に考えてみよう．まず，漢字の機能を分類してみると，一応つぎの6種に分類される．

（1） **単語**　辞書の見出しとして採録されるもので，たとえば，速度，映画，原稿，…….

（2） **熟語**　2種以上の単語が結びついて，1つのまとまった意味を表わし，常に，そのまとまりとして用いられ，辞書の見出しとして採録されるもので，たとえば，白血球，精進料理，…….

（3） **接頭語**　単語として単独に用いられることがなく，他の単語の前について，その語にある意味をつけ加えるもので，たとえば，御親切，打揃う，…….

（4） **接尾語**　単語として単独に用いられることがなく，他の単語の後について，その語にある意味をつけ加えるもので，たとえば，私達，彼等，平均的，…….

（5） **合成語**　熟語のように，常にまとまった形で用いられることがなく，適意2つ以上の単語が組み合わされて用いられる．したがって，辞書の見出しとして採録されない．たとえば，情報伝達手段，…….

（6） **固有名詞**

（5）項の合成語は，漢字独特の造語機能のために，新らたに意味を定義して用いる必要がなく，自由自在に使えるきわめて便利なものであるが，機械による情報処理という面から考えると，きわめて扱いにくい．

†† 当用漢字，およびそれ以外も含めて2,500字（新聞・雑誌に現われる漢字の種類は，およそこの程度であるといわれている）から成る漢字によって作られる単語・熟語の数は，およそ32,000語である．辞書を作成するのにあたって，1単語に計算機1語（32ビットとする）を用いるとすれば，32 K 語程度の記憶容量が辞書のために必要である．

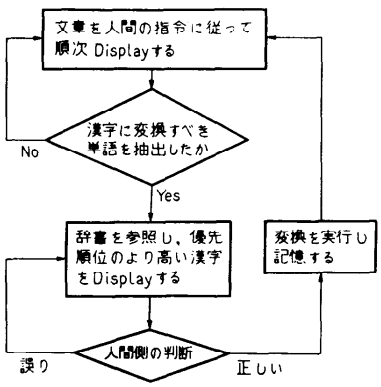
単語と単語を，たとえば，入力の際に（じょうほうでんたつ　しゅだん）のように明瞭に区別††† しておけば，読み-漢字変換プログラムの方はきわめて簡単になり，処理時間も短縮されよう．もちろん，辞書の方を適当に整備しておけば，必ずしもこのように，スペースで区切らなくても可能ではあると思われるが，処理時間が長くなるだけでなく，修正の段階で，人間側の荷重が大きくなり，かえって，能率がおちるという結果になりかねない．

それから，（6）の固有名詞についても，適当なラベルで区別しておく必要があると思われる．しかしながら，熟語内の単語間，接頭語と単語，および単語と接尾語の間は，必ずしもスペースで区切る必要がないようにするべきである††††．

6.2 同音異字の取扱いおよび修正

辞書を適当に整備し，同時にタイピスト側の若干の譲歩を許せば，漢字の単語を一連の文章から抽出することは可能であるという一応の見通しは立てられるが，同音異字については，前述のように，どうしても Human assisted の形式を採用せざるを得ないと考えられる，そこで，たとえば，つぎのように考えてみよう．

まず，タイピストは，前節で述べたような約束に従って，かなタイプライタを用いて文章を打鍵するが，計算機はこの Mode では，打ち込まれる文章をそのま



第1図　読み-漢字変換モード

††† 英語，独語などでも，このように名詞を連らねて用いることが許されているが，英語は単語と単語をスペースで区切り，独語では区切らないで用いている．

†††† たとえば，"教職員家族" の場合は（きょう　しょく　いん　かぞく），（きょう　しょく　いん　かぞく），（きょうしょくいんかぞく）*，（きょう　しょくいん　かぞく）などの区分の仕方が考えられるが，これらはどれも正しく変換されうるようにしなくてはならない．しかし，＊印のついた場合が，最も能率的な区分法であるようにしたい．

ま，一旦記憶してしまう．打鍵が終了したところで，漢字への置換作業を開始するが，それは**第1図**のフローチャートに従うものとする．人はスコープディスプレを見ながら，提示された漢字が正しく変換されたものであるかを判断し，正，あるいは誤のボタンを押すだけでよい．このような形式での人間の関与を考えれば，打鍵の際の誤り，文字の脱落や文章のそう入は，きわめて容易であり，したがって，校正も同時に完了するので，この面からも能率的である．問題は，辞書の整備，あるいはプログラム方式により，いかに人間側の荷重を少なくするかということになる．

最後に，初めにも述べたように，漢字の自動識別には，独特の困難な問題があるが，以上述べたシステムが完成すれば，この問題は，印刷体かな文字の自動識別という問題に置き換えられることになる．また，将来音声タイプライタなどが実現した場合には，ここで取り扱った，いわゆる読み，あるいは発音を漢字，あるいは spelling に置き換える問題が大きく浮び上がってくるであろう．

より現実的な技術という立場から考えても，ここで扱った問題は，単に和文入力装置という狭い立場ではなく，いわゆる日本語を含む情報処理システムという高い立場から，考える必要があると思われる．

参考文献

R. Casey, et al: "Recognition of printed Chinese characters", IEEE Trans. EC-15, No. 1, pp. 91～101 (FEB 1966).

(昭和44年7月21日)