解　説

かな漢字変換†

森　健一††　河田　勉††

1. まえがき

日本語は，漢字，平仮名，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，日本語情報を計算機処理する上で，入力の問題が最大の技術的障害とされてきた．邦文タイプライタや漢字タブレット，漢テレ鍵盤では，数千の漢字の配列の中より，一字ずつ文字を拾って文章を作成する方法がとられているため，入力速度，オペレータの習熟度，入力精度などの点で問題があり，英文用のワードプロセッサのように便利で，コンパクトな事務用文章処理装置は，日本語の場合は実現が難かしいとされてきた．

一方，専門オペレータ向の高速漢字入力方式としてはラインプット方式[1]をはじめとする，2打鍵コード入力方法[2]が工夫され実用化されているが，誰でもが気軽に習熟できる性質のものでないため，主として出版業，漢字ファイル作成などの大量漢字データの入力に用いられている．

カナタイプライタを使える人なら誰でも利用できる漢字入力方法として，カナ鍵盤から文章をカナ文で入力し，計算機を用いて漢字カナ混り文へ変換する方式は，日本において最も普及しやすい入力方法であるとして，多くの研究がなされてきた．

カナ漢字変換には，研究ベースのものと，ワードプロセッサなどの実用に供される装置やシステムとしてまとめた実用指向のものとがある．カナ漢字変換方式を用いた本格的な日本語ワードプロセッサ[3]は，東芝で開発したものが最初であるが，カナ漢字変換の研究は昭和39年の九州大学の栗原らの特許出願[4]に遡る．栗原らは，カナ文の文節分かち書き，単語辞書による照合法，構文解析法，意味解析などカナ漢字変換に関する基礎的な手法を提案[5]し，この方法をもとに沖電気はカナ漢字変換システムを試作した[6]．その後，この成果を基礎にしてNHKの相沢らは，ニュース文に限定して局所的文法処理，単純意味処理の手法を用いて実験システムを作成した[7]．同様に日本ソフトの藤井らは，共同通信社の海外特派員からの外電のローマ字電文を対象に限定した変換プログラムを作成した[8]．同音異義語は2個までに絞り，その選択は行わずに単に並記するなど，精度よりも，早く電文の内容を知ることに重点をおいたシステム作りとなっている．

カナ漢字変換の効率，精度を向上させる問題に関し，電電公社の木村らは同音異義語解析のための関連語情報や意味分類情報の利用を試み[9]，阪大の牧野らはカナ漢字変換における品詞判定を容易にするためのカナ文の分かち書き方法を提案している[10]．以上の実験システムはすべて中型機以上の電子計算機で構成されていたが，実用化のためにはミニコンピュータ，さらにはマイクロコンピュータ程度の小型機で効率および精度のよいカナ漢字変換を実現する必要がある．東芝の河田らの方法では，単語辞書の学習的構成法を提案し，変換効率，精度を向上させるとともに，ミニコンによって実験システムを作成し，制限のない一般文章を対象にしたカナ漢字変換方法を提案した[11]．その後も東工大[12]，電総研[13]など，大学，官公庁研究所，企業研究所においてカナ漢字変換の研究やそれを応用した日本語ワードプロセッサの研究開発が続けられている．

本稿では，2章でカナ漢字変換に関する種々の研究における入力鍵盤，分かち書き，変換方式，同音異義語処理，特殊分節処理，辞書，変換率，校正編集方式などを分類，比較し，問題点を考察し，3章でシステムとしてのカナ漢字変換を論じる．

2. カナ漢字変換における問題点

カナ漢字変換の研究においては，その効率，精度に影響をもついくつかの問題点がある．以下にこれらの問題点について考察する．

2.1 入力鍵盤

† Kana-Chinese Translation by Kenichi MORI and Tsutomu KAWATA (Toshiba Research and Development Center).

†† 東京芝浦電気(株)総合研究所

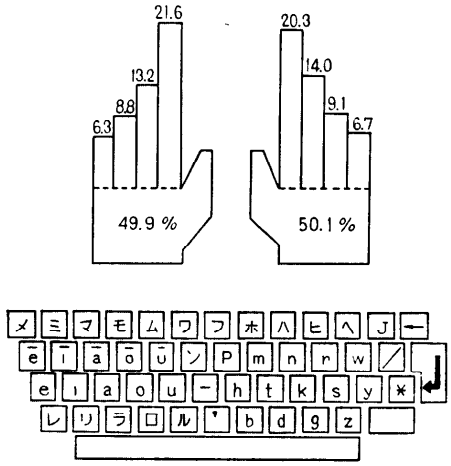

図-1 高速にカナ文の入力できる鍵盤例（ラインプット鍵盤，上は漢字かな混り文を入力したときの各指の負担率）

カナ漢字変換の入力鍵盤には，カナ鍵盤を用いる場合と英文タイプの鍵盤を用いることが出来る．前者は通常48鍵で，カナ小文字などはカナ記号シフト鍵を押して打鍵する方法がとられる．後者はローマ字表記により入力する方法で，海外からのニュースの送稿の場合などに用いられている．一般的にはローマ字表記の方が打鍵数が多くなるが，鍵盤配列を工夫するとカナ表記入力よりローマ字表記の場合の方が高速入力をすることができる[14]．**図-1** に示したラインプット鍵盤[1]では，右手で子音を，左手で母音（3段目が平仮名，2段目がカタカナ用）を交互に打鍵できるように配列されているため，ローマ字表記ながら JIS カナ鍵盤を用いる場合より2倍以上高速にカナ文を入力できる．

2.2 分かち書き

カナ文を入力する際にどのような方式で入力するかは，打鍵速度，品詞認定の精度，合成語の処理や変換率に影響をもっている．主な方式には次のようなものがある．

a）べた書き方式
　スペースを全く用いない方式で，入力速度は早いが，変換率は悪くなる．

b）文節分かち書き方式
　文節単位にスペースを挿入して分かち書きする方式

c）漢字部指定方式
　漢字部をカッコでくくり，平仮名部はべた書きする方式

d）単語分かち書き方式
　単語単位にスペースを挿入して分かち書きする方式

e）自立語付属語分かち書き方式[10]
　自立語と付属語との間にスペースを挿入して分かち書きする方式

制限をつければつけるほど，カナ漢字変換プログラムは，単語の品詞，付属語の活用や接続検定，合成語の処理が容易になるが，一方，入力オペレータは考えながら打鍵したり，リズムを乱されて入力速度が落ち，また誤入力率が高くなる．**図-2** には同一文を上記の5種類の入力方式によって表記したものである．単語分かち書きはローマ字表記の場合の標準的な方式である．漢字になるべきところが，正しく漢字に変換されるためには，当然ながら漢字部指定方式が適している．入力オペレータ自身が入力文をチェックでき，かつ入力速度が高いのは，文節単位分かち書き方式であるが，カナ漢字変換プログラムはそれだけ難かしくなる．逆に最も条件の厳しい自立語付属語分かち書き方式は，オペレータが正確に自立語と付属語の判定が出来れば，カナ漢字変換を要する部分の検出が比較的容易となる利点がある．

2.3 変換方式

カナ漢字変換でカナ文字列を対応する漢字列に変換するとき，完全に自動化をすることはできない．すなわち，著者がどのような意図で文章を書いているかによって同一のカナ文字列が異なる漢字列に対応することがあるからである．例えば「シリツ」は「私立，市立」があり，どちらが正しいかは著者のみが知っている．しかしながら，このような場合を除いては，カナ文字列を出来るだけ，正しい日本語の漢字列に変換することが必要である．

カナ漢字変換には漢字単位に辞書を用意する方法と熟語単位に辞書を用意する方法とがある．前者は漢字とその読みの対応表を辞書とする方法であり，取扱う漢字数により数千字分の辞書を用意すればよい．辞書の大きさに制限のある場合に適した方法であるが，同音異字数の出現度[6]は1字に対し 13.64 字もあり，正

原文　高性能で可搬性にも優れている
a）コウセイノウデカハンセイニモスグレテイル
b）コウセイノウデ　カハンセイニモ　スグレテイル
c）〔コウセイノウ〕デ〔カハンセイ〕ニモ〔スグ〕レテイル
d）コウセイノウ　デ　カハンセイ　ニモ　スグレテ　イル
e）コウセイノウ　デ　カハンセイ　ニモ　スグレ　テ　イル

図-2 カナ文入力の分かち書き方法

しい漢字の選択が大変になる．一方，熟語単位の場合には同音異義語の出現度は平均2.32語であり，後述する文法処理などと組合せるとさらに少なくすることができる．熟語単位の辞書は4～8万語が必要であり，その安価な実装法の実現とともに，高速な検索方法を工夫する必要がある．

辞書の照合方法としては，最長一致法と総当り法とがある．前者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検出して答の候補とする方法である．もし最長一致した語の次の文字列との接続条件（例えば動詞の後の活用語尾など）が不適当であれば，次に長い一致語を候補とする．総当り法では語長の長いものに優先度を与えずに，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方法で最も本格的な方法である．これは最長一致法のもつ欠点。例えば「ヒトハ」は常に「人は」に変換され，「火とは，非とは」には変換される可能性がないことを防いでいる．

2.4 同音異義語処理

同音異義語を合理的にプログラムでどのように選択するかが，カナ漢字変換の中心的課題であり，文法的処理，意味論的処理と頻度情報による処理がある．

2.4.1 文法的な同音異義語処理

これは自立語と付属語との接続の関係に規則性があることを利用する．例えば動詞，形容詞，形容動詞はその活用語尾が活用形で定まっており，それに続く語の品詞が特定のものであることを必要としている．

例：コウショウな→　高尚な　○
　　交渉な，高唱な　×

辞書の熟語にその語の品詞の情報を付加することによって，このような選択をプログラム化できる．名詞と形容動詞の語幹になる語（平和，平和な）や名詞とサ変動詞の語幹になる語（交渉，交渉する）は，辞書中にその情報を付加する必要がある．

語尾変化，音便変化および助動詞，助詞の接続関係などの文法辞書も表形式で表現されることが多い．

2.4.2 意味論的な同音異義語処理

動詞にはその前の主語，目的語に特定の意味範囲の単語を要求する場合が多い．例えば「行く」は

（動作主体）が（場所）へ行く

となり，助詞「が」の前には動作できる主体をあらわす語がくる．

例：センセイが行く（先生はよいが宣誓は不可）

また，単語間に慣用的なつながりや，類義語や反対語が文章中に並立の形であらわれることがあるのを利用することもできる．

例：イッセンを画す（一線はよいが一戦は不可）

さらに，一つの文中に，あるいは複数の文章中の意味的な概念のつながりや構造から正しい語を推定する方法が考えられている[9]が，文章の意味解析の研究が本格的に進まないと未解決な問題が残されている．

2.4.3 頻度情報による同音異義語処理

国語辞典のポケット版サイズのものには，5万ないし7万語の単語が採録されている．一方，日本人の成人は平均3万語前後の単語を理解でき，その共通語は1万3千語程度であることが知られている[15]．その差の1万7千語程度は個人によって異なっていることになる．同様に新聞の政治，経済，社会，学芸欄などによって，使用単語の出現頻度が異なっていることも知られている[16]．カナ漢字変換の入力文の対象分野により，単語の使用分布状態に偏りがあることを利用して，同音異義語の選択をより容易にすることが出来る[11]．極端な応用例[8]では，出現頻度の上位2単語までに限定してしまう場合も考えられる．

2.5 特殊文節処理

日本語は接頭語，接尾語による造語や，二つ以上の熟語をつなげて漢字列の長い文節を自由に作る能力に富んでいる．これらの造語を入力オペレータに正確に分かち書きさせることは，かなりむつかしいことから，カナ漢字変換プログラムは特殊処理として，辞書に格納した単語との照合機能以外に，次のような特殊文節処理機能をもたなければならない．

（a）複合語処理　　例：機械翻訳　等
（b）接辞処理　　例：大型化，新法案　等
（c）数詞処理　　例：第15回　等
（d）固有名詞処理　　例：地名，人名，企業名等

複合語や接辞つきの語をすべて単語辞書に登録したのでは，辞書の容量は際限のないものになるであろう．固有名詞は普通名詞とは異なる特殊な読み方をす

結合の強いもの ↑

		例
数詞	助数詞	百・円
数詞	数詞	百・十三
人名	人名接辞	佐藤・様
人名	人名	徳川・家康
地名	地名接辞	東京・都
地名	地名	港区・三田
普通名詞	普通名詞	国際・会議

図-3　接辞結合強度表[9]

るので，別個の辞書として取扱われるのが普通である．固有名詞や接辞の結合には特有のものが多い．したがって，特殊文節処理では，図-3のように接辞をいくつかのグループ（数詞接辞，人名接辞，地名接辞など）に分け，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試みられている[6]．

2.6 辞書ファイル

辞書ファイルには，単語辞書ファイル，文法辞書ファイル，意味情報ファイル，接辞ファイル，固有名詞ファイル，使用頻度ファイルなどが考えられるが，その収録語彙数をどのようにとるかは前述したように，対象分野によって異なる．一種類の辞書ファイルから頻度情報を学習的に対象分野に適応させ，汎用辞書から個別辞書を自動的に作成する方法[17]が実際の日本語ワードプロセッサに適用され効果を上げている．

文法辞書ファイルには，自立語の動詞，形容詞，形容動詞などの用言の活用型，助動詞の活用表，助動詞の接続表などを含んでいる．意味情報ファイルに何を含むかは研究者により一定していないが，意味的関連語，分類語彙表[18]の意味分類，動詞の格支配表などが用いられている．

固有名詞のうち，人名は日本人で姓が11万種以上[19]，名が24万種以上あることが知られており，全国の市町村大字名は約12万程度である[20]．これをすべて固有名詞辞書に収録することは，カナ漢字変換の効率の面からは得策ではなく，普通は1～2万語を上限として収録し，出現頻度の小さい固有名詞は漢字単位の入力で行う折衷方式がとられている．

辞書ファイルは固定式（システム設計者のみが変えられる方式）と可変式（ユーザが新しい語の登録や削除ができる方式）とが考えられ，実験システムでは多くは前者の方式がとられ，日本語ワードプロセッサ[17]では後者の方法が採用されている．さらに，文章入力中に定義した新しい語や，同音異義語中の選択語を優先的に処理する暫定辞書を設けることは，カナ漢字変換の効率を高めるのに効果的である．

2.7 変 換 率

カナ漢字変換の変換率の定義は，カナ文のどこが漢字に変換されるべきは著者の意志によって決るという意味で客観的に定めることが難かしいが，変換結果からみて日本語として意味が正確に理解できるものは正しい変換と認めると考えてもよい．例えば，最近では「または」「および」などの接続詞は平仮名書きするが，辞書に「又は」「及び」が登録されていれば，漢字となって出力される．この場合，カナ漢字変換としては正しい変換とみなすことになる．2.2 節のC）で述べた漢字部指定方式の場合には，どこが漢字になるべきかは指定されているので，その部分が漢字変換されなければ正しい変換とはみなされないことになる．

カナ漢字変換の精度を評価するには，次のような数値を使うことができる．

$$\text{正変換率}(C)=\frac{\text{正しく変換できた字数}}{\text{投入文章の総字数}}\times 100$$

$$\text{多変換率}(M)=\frac{\text{同音異義語となった字数}}{\text{投入文章の総字数}}\times 100$$

$$\text{誤変換率}(E)=\frac{\text{誤まって変換された字数}}{\text{投入文章の総字数}}\times 100$$

$$\text{無変換率}(N)=\frac{\text{未登録語のため仮名となった字数}}{\text{投入文章の総字数}}\times 100$$

$C+M+E+N=100$ となる．M は同音異義語解析の結果によっても，なお2個以上の同音異義語が残っている場合で，かつ，その中に正しい変換単語が含まれている場合であって，候補語のいずれもが正しい語でなければ誤変換率に数える．無変換率は辞書に単語が登録されていないために，漢字に変換されずにカナ文字のまま残った文字の率である．これまで報告された研究では，辞書の収録語数，対象文章，実験条件がそれぞれ異なるが，85～97% の正変換率を示している．

新語登録機能や暫定辞書機能は，一度発生した誤変換や無変換と同じものが2回目に出現したときには正しい変換をすることが出来るという意味で，カナ漢字変換の精度を向上させる上で効果的である．

意味論的な同音異義語処理はこれまでのところ，十分に効果的に働いたものは少ない．意味処理の内容が比較的単純のものが多く，単語のもつ連想的な結合関係のネットワーク，格助詞一動詞による意味構造，文脈における単語間の意味関係などの本格的な意味表現の処理が行われていないためと考えられ，今後に残された大きな問題となっている．

2.8 校正編集機能

カナ漢字変換においては，前述したようにカナ文の曖昧性による同音異義語の発生から，入力したい語を選択する機能や誤変換や入力タイプミスを訂正する機能が必要である．これらの校正機能はハードウェアの構成法によって，漢字ディスプレイ装置を用いた対話型のものと，漢字プリンタに変換結果を一度出力プリントして赤字訂正した上で，修正個所のデータを入力す

るバッチ型のものがある．校正の容易さからは対話型のものの方が優れている．

編集機能には，文字や文章の削除機能，挿入機能，置換機能などの基本的なものから，漢字検索機能，文章単位の入れ換え，転写，挿入，切貼，合成などの機能，外字処理機能，作表機能や出力印刷様式作成機能，編集結果のファイル管理機能，漢字ディスプレイや漢字プリンタへの出力印刷表示機能など多くの機能が必要である．特に対話型のワードプロセッサに用いるときには，これらの校正編集機能の使い易さによって入力効率は著しく左右される．

3. システムとしてのカナ漢字変換

現在までに実用を目的としてカナ漢字変換システムをまとめたものとしては，九大・沖電気による「漢字かな混り文変換システム」[6]，電電公社通研によるIROHA-1 システム[9]，NHK 技研による「カナ・漢字変換システム」[7]，日本ソフトウェアによる共同通信社向けシステム[8]，東芝による「日本語ワードプロセッサ」[17] などが報告されている．

カナ漢字変換はシステムとしてみるときは，あくまで日本語情報の入力の一手段であり，オペレータにとって多くの訓練を必要とすることなく，効率的に正しく日本語入力ができ，その入力された情報を容易に利用できることが必要である．この意味で **2.8** に述べた豊富なテキスト・エディティング機能とカナ漢字変換機能とは不可分の関係にあり，**図-4** に示した日本語ワードプロセッサはその具体的な一例である．

カナ漢字変換システムを実用的にさらに高めるには，

a）意味処理などの導入による変換率の向上
b）辞書検索方式，編集機能，辞書設計方式の研究による変換の効率，速度の向上
c）漢字プリンタ，漢字パターン発生器，ファイルメモリなどのハードウェア技術の改善による価格の引下げ

など，多くの研究努力の積重ねが今後も必要である．

辞書の設計法の優劣によっても変換率は大きく影響をうけるが，変換率の向上のためには入力しつつある文章が「何について述べているか」の文脈情報から意味処理を行う方法の研究が最も重要と思われる．単純な意味処理では変換率の向上はあまり期待できないからである．

変換効率や変換速度の向上は，単語辞書や文法辞書の構成法や検索法と関連して重要な研究課題となる．またミニコンピュータ以上の計算機を用いてスタンドアロン型のシステムを構成したのでは，システム・コストが高すぎることから，マイクロコンピュータでシステム構成することや，大型機を TSS 方式で使用するシステムが考えられている．

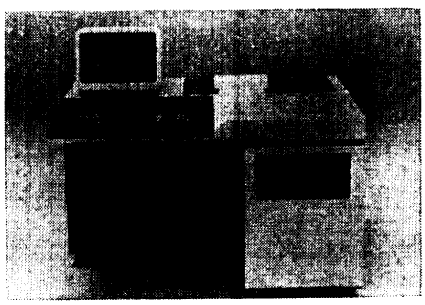

図-4　かな漢字変換を用いた日本語ワードプロセッサ (東芝 JW-10)

4. むすび

カナ漢字変換は日本語情報の入力方法の最も有力な方法の一つとして，今後も発達してゆくものと考えられる．さらに，音声認識装置と組合されて，音声タイプライタや口述筆記装置の実現も将来は可能である．カナ漢字変換の研究は 15 年近い歴史があり，その研究成果の上に立って実用的な日本語ワードプロセッサも市販されるようになってきた．本稿では，カナ漢字変換の研究上の問題点について，いくつかの面から解説を試みた．今後のカナ漢字変換の進歩と実用化により，日本語情報処理が誰にでも身近かに利用できるようになることを期待してやまない．

参考文献

1) 川上他：タッチ打法による漢字入力，情報処理 Vol. 15, No. 11, pp. 863-867 (1974).
2) 山田：日本語テキスト入力法の人間工学的比較，東大情報科学報告，Vol. 78, No. 6 (1978).
3) 森他：計算機への日本語情報入力，電通学会研究会資料，EC 78—23 (1978).
4) 大野他：仮名漢字変換方式，特許公報昭 42—3241 (1964).
5) 栗原，黒崎：仮名文の漢字混り文への変換について，九大工学集報，Vol. 39, No. 4, pp. 659-664 (1967).
6) 松下他：漢字かな混り文変換システム，情報処理，Vol. 15, No. 1, pp. 2-9 (1974).
7) 相沢他：計算機によるカナ漢字変換，NHK 技術研究，Vol. 25, No. 5, pp. 261-298 (1973).

8) 藤井他：カナ漢字への変換，電気四学会連合大会予稿集，177，pp. 619-622 (1971).
9) 木村他：日本語入力用カナ漢字変換システムの試作，情報処理，Vol. 17, No. 11, pp. 1009-1016 (1976).
10) 牧野他：カナ漢字変換の一方法，情報処理，Vol. 18, No. 7, pp. 656-663 (1977).
11) 河田他：ミニコンピュータを用いたカナ漢字変換システム，電子通信学会研究会資料 PRL 76-47 (1976).
12) 高木他：技術文書作成用カナ漢字変換システムの設計と試作，情報処理学会全国大会予稿，p. 103 (1977).
13) 横山他：構文解析を用いたカナ漢字変換，情報処理学会全国大会予稿，p. 102 (1977).
14) 渡辺他：日本語情報処理技術の動向，日本電子工業振興協会，53-C-349, p. 13 (1978).
15) 林巨樹：日本の言の葉，東京書籍 (1979).
16) 国立国語研究所：電子計算機による新聞の語彙調査(I)，秀英出版 (1970).
17) 天野他：カナ漢字変換機能を備えたワードプロセッサ，電子通信学会全国大会予稿集90 (1977).
18) 国立国語研究所：分類語彙表，秀英出版 (1964).
19) 丹羽：日本の苗字，日本経済新聞社 (1978).
20) 国土地理協会：国土行政区画コード総覧 (1972).

(昭和54年6月29日受付)