計算言語学 20－3
（1979. 11. 16）

仮名漢字自動変換方式による日本語ワード・プロセッサ

天野真家、河田勉、森健一　（東芝電気　総合研究所）

1．　概論

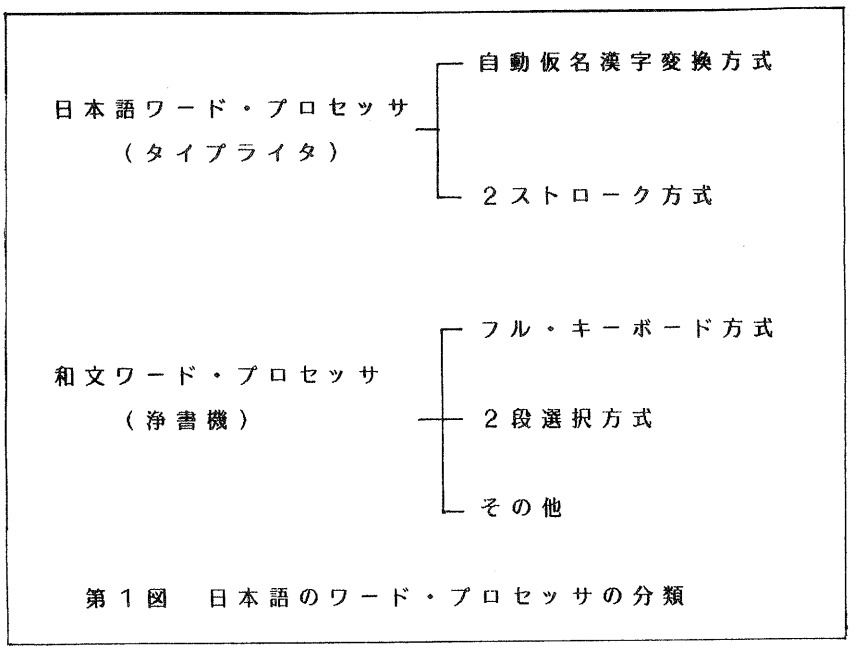
ワード・プロセッシングと言う言葉は周知の様にＩＢＭがデータ・プロセッシングに対する概念として提出したものである。それを受けてワード・プロセッサと言う機械が現われた。しかし、ワード・プロセッサとは一体どの様な機械を指すのであろうか。英文ワード・プロセッサの場合その定義はかなり明確でこの比較的新しいワード・プロセッサと言う言葉を従来の言葉を用いて言い変えれば、メモリー・タイプライタと言えよう。そして、この言葉がそのままワード・プロセッサの定義であると言って良いであろう。しかし日本語ワード・プロセッサの場合、その定義はそれ程簡単では無い。そもそも欧米の意味でのタイプライタは日本には存在しなかった。従来の和文タイプライタが英文タイプライタの意味でのタイプライタとは言えず、単なる浄書機に過ぎないことはここで述べるまでもなく夙に言われてきたことである。タイプライタは少なくとも次の３つの条件を満たさなければならない：

１）　手で書くより速く打てる。
２）　手で書くより楽に書ける。
３）　手で書くより奇麗に書ける。

又、上の１，２の条件を満たすにはタッチメッソドができる機械でなくてはならない。この条件に当てはめて考えれば和文タイプライタが何故単なる浄書機にすぎないかは自明であろう。英文ワード・プロセッサと同等の意味で日本語ワード・プロセッサを考えるなら、それはタイプライタを基礎として考えられなければならない。日本にも英文タイプライタに匹敵するものとして仮名タイプライタがある。これを入力機器として用い自動仮名漢字変換を行なえば、日本語でも英文ワード・プロセッサに劣らぬものをつくれるはずである。

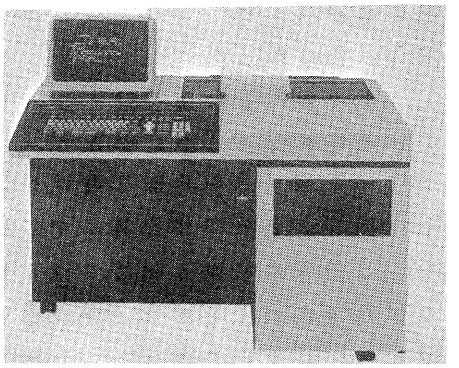
日本語ワード・プロセッサと呼ばれる物の現状は入力方式の違いによる多様性を見せている。従来、その分類は、ただ機械的に方式の違いとしてしか分類されていなかった。ここでは、漢字入力装置としてのコンセプトの違いによるワードプロセッサの分類を行なってみる。

日本語ワード・プロセッサを英文ワード・プロセッサの思想を継ぐものとし、和文ワード・プロセッサを和文タイプライタの後を継ぐものとするのである。この様に見てみると、現状は第１図の様に分類できよう。



第１図　日本語のワード・プロセッサの分類

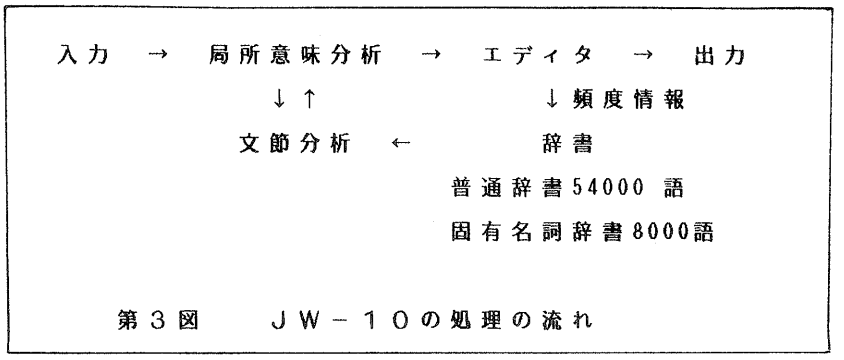
日本語ワード・プロセッサＪＷ－１０は以上の様な思想的背景を持って実現された物である。



第2図 JW-10

２．　システム構成

第２図にＪＷ－１０の外観を示す。第３図には、ＪＷ－１０の処理の流れを示した。タイプされたかな文は局所意味分析部よって自動的に文節に分かち書きされ、更に文節の解釈をうける。それに基ずいて文節分析部が文節単位でかな漢字変換を行なう。漢字に変換された文節はエディタに送られ表示、印刷される。編集・校正中に得られた同音語の頻度情報は辞書に送られ学習される。



第３図　　ＪＷ－１０の処理の流れ

３．　入力方式

日本語をかな文で入力するには、二つの方法がある。

（イ）　文節分ち書き方式

（ロ）　シフト方式

前者は文節と文節の間で文節キーを叩いて、かな文を表す方式である。後者は漢字の部分をシフト・キーで表す方法である

たとえば『衆議院の選挙が実施された。』という文の入力は、次のようになる。

（文節分ち書き方式）

しゅうぎいんの＿せんきょが＿じっしされた。

（シフト方式）

［しゅうぎいん］の［せんきょ］が［じっし］された。

但し＿は文節キーを、［は漢字シフト・イン、］は漢字シフト・アウトを示す。

入力されたかな文は、「局所意味分析部」により全て文節分ち書き文に変換され、文節分析部に送られる。

４．　局所意味分析

４－１　局所意味分析概論

局所意味分析（以後　ＬＭＡ；Local Meaning Analysis）とは日本語が持つ多数の字種が各々それら独得の意味を担っている事実に着目し、それらの字種の近辺で局所的な意味のまとまりをみい出そうとするものである。

例えば、「だい３４かいの・・・」という文字列は「第３４回の」と解され「台３４会の」とは解されない。これは「３４」と言う数字の近辺で局所的意味のまとまりがあるからである。同様に「カントてき」と言う場合、「カント敵」とは考えられない。この場合の「てき」は「カント」と言う人名の勢力範囲にあるからであり、独立に意味の存在する「敵」がカントと言う人名の直後に来る可能性は、「的」よりも小さいと予測した結果である。これがＬＭＡの取る立場である。

４－２　局所的意味のまとまり

４－２－１　片仮名／英字　の局所的意味

片仮名／英字　文節の構造は次のようになっている。

｛接頭辞／名詞｝片仮名／英字｛接尾辞／名詞｝｛付属語｝”

但し、｛｝はその内容を省略できることを示す。又、／は二者択一を、”は繰り返しを許す事を示す。

例：　こくさいＢＵＳＩＮＥＳＳしゃ；

国際ＢＵＳＩＮＥＳＳ社　（ＬＭＡを行なった場合）

国債ＢＵＳＩＮＥＳＳ紗　（ＬＭＡを行なわなかった場合）

４－２－２　数字の局所的意味

数字文節は次の構造を持つ。

｛前置助数詞｝数字［｛後置助数詞｝］”｛数詞接辞｝｛付属語｝”

前置助数詞とは数字の前に置かれ数字を修飾する接辞である；　第、約、数、昭和……

後置助数詞は数字の後に置かれ数字を修飾する接辞である；　台、回、階、巻、月、……

数詞接辞とは後置助数詞の直後に置かれ、数字を修飾する接辞である；以上、程度、程……

数詞文節は次の３種に分けられる。

（１）単純数詞

１２３４は……

（２）　助数詞付き数詞

昭和５４年には……

（３）　助数詞、数詞接辞付き数詞

２台程度の……

（１）についてはＬＭＡは何も行う必要がない。

（２）以下は、「・・・数字・・・」と言う構造を持っている。只、構造上だけでは、「・・・」の所が助数詞であるかどうかを決定することができない。例えば、「歯車８は」のような［名詞－数字］のような構造のものも含まれるからである。このチェックはＬＭＡで行なわれる。

ＬＭＡにより次のような結果が得られる。

だい２の：　第２の　（ＬＭＡあり）

台２の　（ＬＭＡなし）

２だいの：　２台の　（ＬＭＡあり）

２題の　（ＬＭＡなし）

（３）　助数詞、数詞接辞付き数詞の処理

この処理により次のような文節が処理される。

４５かいいじょう：　４５回以上　（ＬＭＡあり）

４５会異常　（ＬＭＡなし）

以上、ＬＭＡがここで行っている仕事は

（１）　分ちがきされていない文字列の連続体の中から局所的意味のまとまりを推測し、それを取り出す。

（２）　辞書を引く事によりその推測を検定する。

である。第４図にＬＭＡの結果の例を示す。

	ＬＭＡあり	ＬＭＡなし
だい２３かいの：	第２３回の	台２３会の
７どいじょうの；	７度以上の	７度委譲の
やく３４㎜；	約３４㎜の	役３４㎜の
どうホテルは；	同ホテルは	胴ホテルは
トロントはつ；	トロント発	トロントはつ

第４図　ＬＭＡの結果

５．　仮名漢字変換

５－１　変換方式

変換は局所意味分析部の制御に基ずき文節単位で行なわれる。辞書による形態素の分析と、文法による付属語結合分析がその主体であるが、日本語は複合語を作る能力に富んでいるので複合語分析も同時に行なっている。

辞書の照合方法には、最長一致法と総当り法とがある。前者はかな文と辞書に収録した単語を比較するときに、かな文字列が一致する限り最も長く一致する単語を検出し、もし最長一致した語の次の文字列との接続条件（例えば動詞の活用語尾と助動詞の接続は限られている）が満されればその語を答とし、もし条件が満されなければ次に長く一致する語を候補語とする。この方法は照合アルゴリズムが比較的簡単なため多くの研究者によって採用されているが、次のような誤変換をする可能性が残されている。例えば「ひとは」は常に「人は」に変換され、「火とは」には変換されることがない。このような点からＪＷ－１０ではあらゆる可能性を調べた上で選択する総当り法を採用している。

５－２　同音異義語処理

熟語単位の場合、同音異義語の出現度は辞書の収録語数、入力文章の性質によっても異なるが　１語当り1.7 ～2.5 語程度出現する。同音異義語をどのように合理的に軽減するアルゴリズムを開発するかがかな漢字変換の中心的課題のひとつである。

ａ）文法的処理

これは自立語と付属語との接続関係に規則性があることを利用する。例えば動詞、形容詞、形容動詞などはその活用語尾が活用形によって定まっており、活用形に続く助動詞は特定のいくつかに限られている事実を用いる。

例：きょうりょくな　→　強力な　○
　　協力な　×
　　きょうりょくを　→　協力を　○
　　強力を　×

辞書の単語に品詞情報を付加することにより、このような分析を行なう。

b）局所意味処理による方法については第４章参照

c）頻度情報による学習

国語辞典のポケット版には、５～７万語の単語が収録されている。一方、日本人の成人は平均約３万語の単語を理解し、誰にも共通している語は約１万３千語程度である。その差の約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、学芸の各欄によって、使用単語の出現分布が異なっていることも知られている。かな漢字変換の入力文の対象分野により、単語の頻度分布に偏りがあることを利用し、対象分野がきまると出現頻度の小さい単語を無視すれば、見かけ上、同音異義語の出現度を小さくすることができる。ＪＷ－１０は汎用辞書から個人別あるいは分野別の辞書を学習的に作成し、使っている内に個人、又は分野に適応した辞書になって行く。

５－３　特殊文節処理

日本語は接頭語や接尾語、あるいは、２つ以上の単語をつなげて長い複合語を自由に作る能力に豊んでいる。これらの造語をタイピストに正確に分ち書きさせることは面倒なことになるため、かな漢字変換では次のような特殊文節処理の機能が必要となる。

a）複合語処理

例：きかいほんやく→機械　翻訳

b）接辞処理

例：しんぎじゅつ　→新技術

c）固有名詞処理　　地名、人名、企業名等

固有名詞文節は普通名詞文節のとは異なる構造を持っている。従って、普通名詞と同じ方法で処理する事は得策ではない。

例；　さとうし　→　佐藤氏　○
　　佐藤市　×

の様に、佐藤市と言う市が存在しない事をプログラムは知っているので

又、　　例；　さかいし　→　坂井氏、酒井氏、堺市

おおさかふ　さかいし　→　大阪府堺市

の様にさかいしが単独である場合と他の文節が伴なっている場合とでは処理が異なる。単独の場合、上記のどれも正しいが大阪府が伴なうと堺市に限られるのである。即ち、この場合、大阪府と言う情報でさかいしは市であるべきと判断され曖昧性が解消されるのである。

５－４　辞書

辞書には、単語辞書、固有名詞辞書の２種がある。辞書には、かな、漢字、文法情報、頻度情報などが含まれている。固有名詞は姓名、地名ともに１０万語以上となるため、普通は対象分野によって１～２万語を収録し、出現頻度の小さい固有名詞は漢字単位で入力する折衷案を用いた。又、次に述べる辞書への新語登録の機能で良く用いる物は登録しても良い。

辞書に新語（辞書に収録されていない単語）を自由に登録する機能をもたせることは、かな漢字変換の無変換率の向上のためには不可欠であり、さらに文書入力中に定義した新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、かな漢字変換の効率を高めるのに効果的である。

６．　テキスト・エディタ

エディタが豊富な機能を持つ事は不可欠な条件であるが、同時にその操作性が優れていなければ意味が無い。例えば、削除と行っても１字ずつしか出来ないようでは能率が非常に悪くなる。現実の編集・校正では過去の文書がたまるにつれ大量の削除・訂正が生ずるのである。ＪＷ－１０ではこのような事を考慮に入れた設計がなされている。

ＪＷ－１０のテキスト・エディタは、色々と便利な文書編集機能を備えている。例えば、商業文、公用文などの定形文書のファイルと、顧客の住所ファイルとを組み合わせて、多数の宛先きの異なる手紙文を自動的に作り出す機能は、挨拶状や通知書を作る上で大変便利である。さらに、一度作成した文書をファイルに保管しておき、任意のときに呼び出し、コピーや新しい文書作成の素材にすることができるファイル管理機能を完備している。ファイル機能はワード・プロセッサには不可欠の機能であり、これにより文書処理がこれまでにない形で実現される様になった。

この他、作表機能、自動ページング、文書の合成、枠空け、タブ、インデント処理、自動センタリングなど多くの便利な機能が、キーを打つだけで作業できるようになっており、文書の作成作業を大幅に時間短縮できる。

7．　結言

かな漢字変換による日本語ワード・プロセッサの出現は単にワード・プロセッサの出現に留まらず本来の意味でのタイプライタが日本語でも可能になった事を意味する。本来の意味とはタイプライタは速く、楽に、奇麗に、そして誰にでもタイプできなければならないと言う意味である。タイプライタが、書くと言う事に果たす貢献は大きなものである。そのためにカナモジカイの様な運動が存在するのである。仮名だけに限れば、タイプライタはできるのであるが、やはり、漢字を放逐する事はできないのが現状である。仮名でタイプすると言うメリットを持ったまま漢字で印刷できればそれにこした事はない。その上ワード・プロセッサとしての機能を用いれば文書の保存・加工機能により、定型文など容易に出来る。かな漢字変換日本語ワード・プロセッサの意義は、それによってこの様な“書く機械”が実現された所にある。

参　考　文　献

（１）　栗原，稲永：カナ漢字変換、九州大学工学集報、（1970年）

（２）　相沢，江原：計算機によるカナ漢字変換、ＮＨＫ技術報告、Ｖｏｌ．２５，５（1973年）

（３）　松下，山崎：漢字カナ混り文変換システム、情報処理、Ｖｏｌ．１１５，（1975年）

（４）　勝部，牧野：カナ漢字変換の一方法、電子通信学会研究会資料、ＰＲＬ７６－９，（1976年）

（５）　河田，天野：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、ＰＲＬ７６－４７，（1976年）

（６）　天野、河田、他：カナ漢字変換機能を備えたワード・プロセッサ　電子通信学会情報部門全国大会　90，（1977年）

（７）　森，河田，他：計算機への日本語情報入力、電子通信学会研究会資料、ＥＣ７８－２３，（1978年）

（８）　森、児玉、他：日本語ワード・プロセッサ（ＪＷ－１０）　電子通信学会総合全国大会　1211　（1979年）

（９）　河田、天野、他：日本語ワード・プロセッサ（ＪＷ－１０）のテキスト・エディタ

（１０）　K．Kawada，S．Amano：Japanese Word Processor the Sixth International Joint Conference on Artificial Intelligence－７９（p466～468）（1979年）
（１１）　K．Kawada，S．Amono，et al．：Japanese Word Processor ＪＷ－１０，ＩＥＥＥ Compcon７９（p238～242）（1979年）