自然言語処理 31－1
（1982. 5. 21）

文節を最小単位とするわかち書き文のローマ字漢字変換

千葉 和彦　浅見 徹　博松 明
（KDD研究所）

1．まえがき

日本語文章の入力方式としてのカナ漢字変換は、オペレータの負担が少なく、容易に入力できる手段の一つとされている。カナ漢字変換ではシステムが変換単位を決定しやすいよう、何らかの形で入力文をわかち書きしておく場合が多い。この時のわかち書き方法は、単語単位のわかち書きからベタ書きまで種々考えられ、最小単位を単語とした自由入力形式のもの［4］、文節を若干拡張したもの［5］、全くのベタ書きのもの［3］等が提案されている。

通常の漢字カナ混じり文ではわかち書きの習慣がないため、一般にわかち書き単位の数が多いほど、分割する場所に関する知識や余分なキーの押下などオペレータの負担が大きい。比較的よく用いられている文節わかち書きにおいても、文節そのものが『それ以上切ることのできない語のまとまり』（日本文法大辞典）のような定義のされ方で、補助用言や連語の扱い等判断に迷う場合も多く、文法知識や解釈の相違によってわかち書きに個人差が生じる。逆にベタ書きに近づくほど文章解釈の際のあいまいさが増し、変換システムの処理は複雑になるであろう。

本稿では、最小単位を文節とする恣意的わかち書きローマ字文を入力とするローマ字漢字変換システムについて述べる。このシステムでは、変換後に会話型の編集・再処理段階を設けることにより、初期変換を比較的簡単な処理ですませている。

2．文節を最小単位とするわかち書き

一般に、変換システムにとっては詳細なわかち書きがなされたほうが、処理は楽になるであろうが、オペレータには

（1）区切り記号挿入位置の判断、
（2）余分な区切りキーの押下、

等の負担が加わることになる。

一方ベタ書き入力の場合はオペレータに対するこれらの負担は軽減されるが、ベタ書きすることによって発生するあいまいさのため、システムの処理は複雑にならざるを得ない。オペレータがあいまいさを意識している場合、例えば、『ここではきものを・・・』のように、文節分割のあいまいさが生じることが明白な場合は、オペレータが以後の処理を意識して積極的に区切りの指示を行なうことはそれほど抵抗がないと考えられる。また、入力文を見直す場合にも、適度な区切り記号の挿入は有効であろう。

日常漢字カナ混じり文を使用している我々には、わかち書きの習慣がないため、ローマ字文をわかち書きにより入力する場合には、そのわかち書き規則が

（1）規則が簡単で、覚えやすい。
（2）分割する場所が容易に判断できる。（自然な切れ目である）
（3）余分なキー操作が少ない。（回数が少ない）

という条件を満たしていることが望ましい。厳密な文法的規則でわかち書きを行なうことは、文法が提案者によって異なる場合も存在し、オペレータの訓練（学習）が必要となるため、避けるべきであろう。

本システムで採用したわかち書き法は、文節わかち書きを拡張し、最小単位が文節であることだけを要求する形式とした。入力に課される制限は文節内に区切り記号（スペースを採用）を挿入しないことのみであり、完全文節わかち書きから、ベタ書きまでの範囲で自由に区切り記号を挿入できる。すべての文節に区切り記号を挿入することは、前述のような負担がオペレータに課されるが、必ずしもすべての文節でなく、任意に挿入するようにすれば、それほどの負担ではないであろう。

また、ここでいう文節では、補助動詞や接尾語的に用いられる語も自立語として文節を構成しうるとしている。これにより、オペレータが文節であるかどうかを判断に迷う部分、例えば”～していく”、”～している”のような補助動詞、”～しなければならない”のような連語に関しては連続入力、分割入力を共に許している。

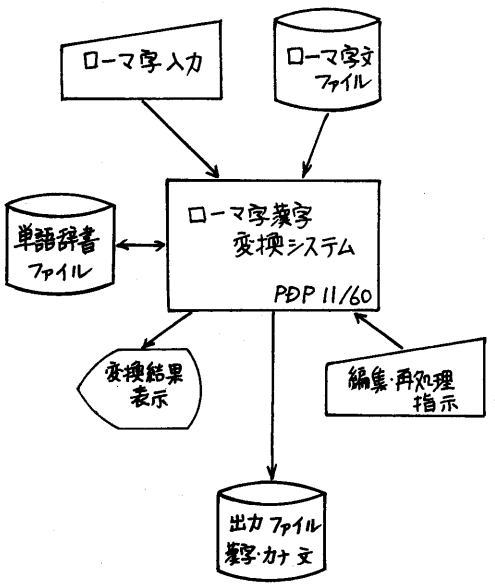

図１　　システム構成

３．ローマ字漢字変換システム

本システムは小型計算機ＰＤＰ１１／６０上にマクロアセンブリ言語によってインプリメントされている。本システムの構成を図１に、処理の流れを図２に示す。変換に使用する辞書は新明解国語辞典［２］に収録された約６万語、文法は西村らの日本語基本文法［１］をベースとした。

３．１　変換方式

入力ローマ字の解析は文節最長一致原則による。これは辞書引きによって選択された自立語候補に対して、その自立語が文節を構成するのに必要な後続付属語と入力文の後続文字列が一致しているかどうかを判定し、文節として認定される最も長いものを選択するものである。文節最長一致の概念を図３に示す。

活用語の辞書登録は取り扱いやすさから語幹のみとしているが、このため語長の短い活用語は自立語のみの最長一致基準では無視されることが多く、特別な配慮をする必要があるが、文節最長一致法ではその必要がない。このように、文節最長一致法では自立語のみの最長一致法に比べ、自立語部が短く、付属語部が長いような文節の誤変換が救済ができる。

本システムでは、この文節最長一致原則の例外による誤変換の発生については、編集・再処理を会話型で行なうこととし、最初の変換では辞書引きと付属語のチェックという比較的簡単な処理しか行っていない。誤変換の発生及びその再処理については次章で述べる。

入力ローマ字文を文節最長一致法で文節として認識したときに、その長さが同一となるものを同音語とするが、これには以下のものが考えられる。

（１）　自立語が同長・同品詞のもの
（２）　自立語が同長・異品詞のもの

（3）　自立語が異長のもの

これらの同音語については、自立語の出現頻度情報に基づいて順位付けられ、その１位のものに変換される。（1）に対しては構文的な選択基準は存在しない。（2）についても、最長となる後続付属語列が、どちらの品詞に対しても接続可能であれば自動的に選択ができない。（3）は自立語と付属語列の区切り部分に対して異なる解釈が可能な場合であり、文節としては同長であるが、自立語の長さが異なるものである。（例えば、”人は”と”火とは”）（3）は通常の意味の同音異義語ではないが、これらに対しても選択表示可能とすることで、最終的な誤り率を低下させることができる。

３．２　誤変換再処理

誤変換の原因としては、以下のようなものが考えられる。

（1）辞書未登録語の存在
（2）文節最長一致原則の例外
（3）入力誤り（タイプミス等）
（4）同音語の存在

これらに対応して、編集段階では以下の５つのコマンドを用意して、訂正・再処理を行っている。

（A）未定義語切り出しコマンド
（B）文節指定コマンド
（C）文字訂正コマンド
（D）同音語表示コマンド
（E）編集終了コマンド

編集段階では、ディスプレイ上に文節毎に表示された、変換結果を基にして上記コマンド用いて訂正を行う。各コマンドはパラメータとして、文節毎につけられた文節番号を持ち、対象となる文節を指定する。各コマンドの機能を以下に述べる。

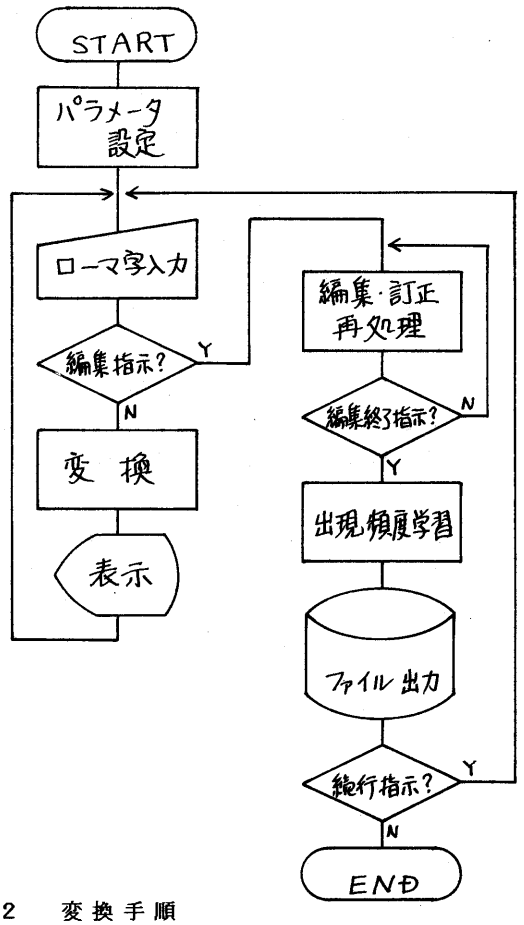

図２　変換手順

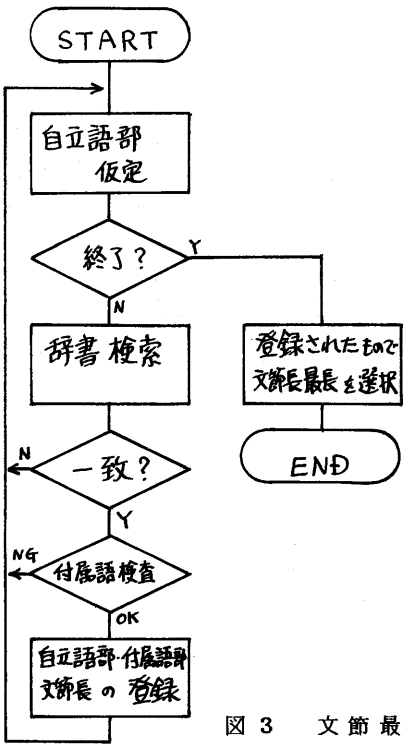

図３　文節最長一致

(A) 未定義語切り出しコマンド

指定されたローマ字列を未定義語とみなして切り出しかなで表示する。

(B) 文節指定コマンド

誤変換の一つの原因として、文節分割に関する情報が入力文から欠落していた場合に、最長一致原則の例外によって、文節分割点を誤る場合がある。この文節分割に関する情報を編集時に付加するためのコマンドとして、文節再指定コマンドを導入した。このコマンドでは、分割誤りを発生した文節と、区切るべき位置として後続文節の自立語の先頭を指示する。システムは指定されたストリングを自立語の先頭と解釈し、そこから再処理を開始し、最初の変換結果と再処理結果の文節分割点が一致するところまで再処理を行なう。また必要であれば、指定された直前のストリングに対しても再処理を行う。これは、その文節の先頭から指定された点までのストリングを一文節とみなすため、付属語の相違のみで生起した区切り誤り以外の場合に必要となる。

(C) 入力訂正コマンド

タイプミス等誤って入力されたローマ字列を変更し、その文節の先頭から再処理を行う。再処理の範囲は文節指定コマンドと同じ。

(D) 同音語表示コマンド

同音語を持つ文節は最初、自立語の出現頻度情報に従って付された順位の1位のものが表示される。同音語表示コマンドは、文節を指定することにより、その文節の同音語の次候補を表示する。

(E) 編集終了コマンド

編集・再処理の終了を意味し、出現頻度等の学習を行う。

文節最長一致法では、この原則の例外に対しては誤変換が避けられないが、その多くの場合は、付属語の誤認識による。ある自立語に対して、入力文と一致しかつ接続可能な付属語が複数存在する場合、一方は他方のプレフィックスとなっているため、最長一致原則では長いほうがとられ、”～とは”、”～とも”、”～には”等のような文字列は助詞＋助詞と解釈され、後続文節の自立語の先頭文字を助詞として文節に取り込んでしまう。これを自動的に救済するためには、以後の変換に何らかの誤変換が生じたことを検出し、バックトラックを行ない、再処理開始点を検出するという手順が必要である。辞書登録単語数が数万語のオーダーになると必ずしも変換不能な状態に陥ることがなく誤変換の検出は困難である。また、バックトラックの手法では再処理開始点の決定も困難であり、処理が複雑となり、繰り返し回数の増加が避けられない。

本質的にあいまいな文章（例えば、”ここではきものを・・・”）に対しては、自動的に変換することは構文的な処理のみでは不可能であり、オペレータからの文節情報の付加が必要となる。また後述の実験の結果、誤変換の収束は速く起こる（1～3文節）ため、誤変換部分の再処理は比較的短時間ですむ。

以上の理由から必ずしも完全な自動変換を指向せず、同音語の選択と同程度の労力で再処理を指示できるようなコマンドを用意し、会話型の編集指示を行なうという形態をとることにした。

4．変換実験

前章で述べたシステムを用いて、コンピュータ関係雑誌の解説記事から抽出した 170文章、約3500文節をローマ字表記したものを入力として変換実験を行なった。入力ローマ字文は完全文節わかち書きを施したものと、ベタ書きのものの両者について行なった。この結果を表１に示す。

表１において、誤りが発生した文節を、その原因によって２種に分類している。辞書未登録語は、”ディスク”、”アクセス”等の専門語がほとんどである。

完全文節わかち書きの場合に発生する誤変換は、接頭語・接尾語によるものがほとんどである。本システムでは、接辞については特別な処理を施していないため、誤りの発生率が高い。また、本質的にこの手法では変換できないもの、付属語検査処理の変更を要するものが 1.6％程度あった。

一方、ベタ書きで入力したものは完全文節わかち書き入力と比較して、変換率の劣化は 7.5％となっている。

この誤変換部分に対して、文節分割コマンドにより編集を行った結果、正解率は96.2％まであげることができた。このときのコマンドの使用回数は全文節数に対して、完全文節わかち書きの場合 2.8％、ベタ書きの場合 8.9％である。

文節分割コマンドでも訂正できなかった誤変換は５５文節あり、その内訳を以下に示す。

①品詞カテゴリによるもの

（例　独立な、等価な、見渡せる）

新明解国語辞典では、名詞・副詞のうちサ変動詞や形容動詞としての用法をもつものには『－する』『－な・－に』等の表示がされているが、必ずしも網羅されていないので、変換辞書作成時の品詞分類の際に異なるカテゴリに分類され、後続付属語が接続不能とみなされる場合がある。また動詞のうち可能動詞形をもつものについても同様である。

②動詞連用形の名詞転化によるもの

（例　取り扱い、問い合わせ）

新明解国語辞典において、名詞の見出しとなっていないものは、格助詞の接続を認めないため、誤りとなる。

③文語的表現によるもの

（例　～すべき、尽きざる）

本システムでは文語的表現には対処していない。

④接辞の濁りによるもの

（例　型（がた）、箱（ばこ））

本システムでは造語成分に対する特別な処理を行っていないので、その接続による濁音の発生には対処できない。

表１　　変換実験結果

		正解文節	誤り文節		文節分割コマンド投入率
			辞書未登録語	最長一致例外等	
編集前	完全文節わかち書き	９３．４	２．２	４．４	２．８
	ベタ書き	８５．９	２．２	１１．９	８．９
編　集　後		９６．２	２．２	１．６	

（単位　％）

⑤付属語検査処理の不備によるもの
（～するかが、～するのかは）
付属語検査処理は全ての場合を網羅するのは困難であり、例外が必ず発生する。これについては、さらに検討が必要であろう。

5．まとめ

文節を最小単位とするわかち書きローマ字文を漢字カナ混じり文に変換するシステムについて述べた。このシステムでは、入力を一旦漢字カナ混じり文に変換した後、編集段階において再処理を行なえる方式としたので、最初の変換の際の処理は比較的単純なものとなっている。

このシステムに対して完全文節わかち書きで入力を行なった場合と、ベタ書きで入力を行なった場合の比較では、変換の正解率に7.5%の差がでている。しかし、両者の中間的なわかち書きで入力を行なった場合には、区切り記号の挿入率が高くなるほど正解率は高くなるであろうが、これは線形ではなく、比較的少数の区切り記号の挿入で完全文節わかち書きを行なった場合に近い正解率を得ることが可能である。これは、誤変換の発生の原因がある程度限定でき、例えば、”に”や”も”などの助詞の直後に”は”や”も”で始まる自立語が続く場合は区切り記号を挿入する、というような簡単なルールが設定できると考えるからである。そのようなルールはオペレータが常識的に知っているものも多いであろうし、システム利用を重ねるうちに経験的に習得しうるであろう。この点については今後実験によって評価を行っていく予定である。

謝辞

日頃御指導を戴く当研究所の鎗治所長、寺村副所長、中井次長並びに端末装置研究室の各位に感謝する。また、新明解国語辞典の磁気テープを提供して下さった三省堂の関係者各位に感謝致します。

参考文献

[1] 西村他：日本語基本文法－複文編－、電総研報告第 784号

[2] 金田一他編：新明解国語辞典 第二版、三省堂

[3] 牧野他：べた書き文の仮名漢字変換システムとその同音語処理、情報処理学会論文誌 Vol.22,No1,Jan.1981

[4] 内田他：自由入力形式のカナ漢字変換、情報処理学会自然言語処理研究会資料27-3,Sep.1981

[5] 藤崎他：「ことだま」文書処理システムの文節わかち書き仮名漢字変換、情報処理学会論文誌 Vol.23,No1,Jan.1982

[6] 吉村他：日本語文の形態素解析における最長一致法と文節数最小法について、情報処理学会自然言語処理研究会資料30-7,Mar,1982

[7] 千葉他：最長一致型カナ漢字変換における文節分割と誤変換の発生について、昭和57年度電子通信学会総合全国大会1262、Mar,1982