べた書き文の分かち書きと仮名漢字変換†

——二文節最長一致法による分かち書き——

牧野 寛†† 木澤 誠††

べた書き文からの漢字変換を行ううえで必要となる自動分かち書きの方法を提案し，それらを用いた仮名漢字変換システムについて述べる．

本システムの実用上の目的は漢字入力を容易に行えるようにすることにあり，打鍵者に課する負担を小さくするために，入力仮名文として，分かち書き用の制御信号などを入れない，いわゆるべた書き文を採用した．

べた書き文から漢字仮名混り文への変換過程で必要となる分かち書き方法として，二文節最長一致法を用いている．二文節最長一致法とは連続する二文節と見なし得る文字列の長さを尺度として，文節分かち書きを行っていく方法である．また，付属語による分かち書きを用いて，未登録語（辞書中に存在しない語）の出現に対しても，分かち書きを可能としている．これらの分かち書き法を基本とするべた書き文の仮名漢字変換システムを構成し，実験を行った．

雑誌などからの214文（総文節数 2592 でその4%程度の文節に未登録語を含む）を対象とした実験では，97.2%の文節が正しく分かち書きされており，本論の提案する分かち書き方法が有効であることを示した．

1. まえがき

漢字仮名混り文で表記された日本語を取扱う場合，その入力方法は実行上の難点となっている[1]．現在は漢字鍵(けん)盤装置などを用いるのが一般的であるが，大量の情報を取扱い，かつ入力速度を向上させるためには，打鍵に相当の熟練を要することと打鍵に際して鍵盤を注視しなければならないところに実行上の制約がある．一方入力文をすべて表音表記した仮名書き文とする仮名入力方式は，漢字鍵盤装置よりも廉価でかつ一般性のある仮名鍵盤装置を用いて，鍵盤を注視する必要もなく打鍵することができる方法で，個々の入力装置の経費と操作員の習熟のしやすさにおける利点とともに打鍵の高速化が期待できる．その反面仮名書き文を計算機によって自動的に漢字仮名混り文に変換することが必要となり，いわゆる仮名漢字変換の問題が生じる．

仮名入力方式においても，計算機による漢字変換を容易にするために，入力する仮名文にあらかじめ分かち書きを施したり，特に制御信号を入れたりするシステムが提案されてきた[2)~6)]．しかしながら，このように入力文にあらかじめ処理を施すことは，入力操作員の負担となって入力速度の低下を招くばかりでなく，文法知識の深さや解釈の相違などによって，分かち書きのしかたに個人差を生じ易く，これが漢字への変換に影響を及ぼすという欠点を持つ．この欠点を解消し，入力操作員の負担を軽減するには，べた書き，すなわち分かち書きを行わずに記述された仮名入力文形式を採ることが望ましい．その代りに機械による分かち書き，いわゆる自動分かち書きの問題を解決しなければならない．

本稿では，べた書き文の自動文節分かち書きの方法を提案し，それらに基づく仮名漢字変換システムについて述べる．以下，2章では処理の基本となる文節について，3章では二文節最長一致法，付属語による分かち書きなどの分かち書きアルゴリズムについて，そして4章では3章で論じた処理を用いた実験システムについて，それぞれ述べる．

2. 文節と文節形

日本語文は文節* と呼ばれる単位が有限個連続したものと考えられ，さらに文節は通常漢字変換されうる自立語** と文節の切れ目情報を持つ付属語*** を含むことから，べた書き文の分かち書きおよび漢字変換の処理単位として，文節を用いることとする．したがって分かち書きの基本となる文節の持つ性質について以下に述べる．

文節の構造は形式的に図1に示すように表現できる．

いま文節が，単語 $w_1, w_2, \cdots, w_n$ からなるとき，

† Automatic Segmentation for Transformation of Kana into Kanji by HIROSHI MAKINO and MAKOTO KIZAWA (Faculty of Engineering Science, Osaka University).

†† 大阪大学基礎工学部情報工学科

* 文を実際の言語としてできるだけ多く句切った最も短い一句切[7]．

** それだけで一つの文節になることができる単語．

*** それだけでは文節を作ることができず，かならず自立語に結びついて用いられる単語．

〈文　節〉=〈自立部〉〈付属部〉
〈自立部〉=〈自立語〉|〈接頭語〉〈自立部〉|〈自立部〉〈接尾語〉|〈数字〉〈助数詞〉
〈付属部〉=〈付属語〉|〈付属部〉〈付属語〉
〈自立語〉=〈体言〉|〈用言〉|〈その他の詞〉
〈体言〉=〈名詞〉|〈代名詞〉
〈用言〉=〈形容詞〉|〈形容動詞〉|〈動詞〉
〈その他の詞〉=〈副詞〉|〈連体詞〉|〈接続詞〉|〈感動詞〉
〈付属語〉=〈助詞〉|〈助動詞〉|〈補助用言〉|〈形式名詞〉
〈名詞〉=花|木|…　，〈代名詞〉=これ|それ|…
〈副詞〉=きっと|たくさん|…，〈連体詞〉=この|ある|…
〈接続詞〉=だから|しかし|…　，〈感動詞〉=ああ|まあ|…
〈助詞〉=が|は|…　，〈助動詞〉=られる|れる|…
〈補助用言〉=ある|いる|…　，〈形式名詞〉=こと|もの|…

図 1 文節の定義

Fig. 1 The definition of a segment.

文節内の連続する二つの単語 w_i, w_{i+1} の接続に関する規則は正規文法で表現できる．したがって，w_i と w_{i+1} の間の接続関係を示す関数 $C(w_i, w_{i+1})$ を次のように決めれば，

$C(w_i, w_{i+1})=1$　w_i と w_{i+1} の接続は可

$C(w_i, w_{i+1})=0$　w_i と w_{i+1} の接続は不可

文節内の各単語 w_i は，次の関係を満たす．

$C(w_i, w_{i+1})=1$　$1\leq i\leq n-1$

さらに各単語の性質として，文節の末尾になり得るか否かが決められる．例えば，用言の未然形などは文節の末尾とはならないといえる．したがって単語 w_i に対して関数 $T(w_i)$ を次のように決めれば，

$T(w_i)=1$　単語 w_i は文節の末尾となる

$T(w_i)=0$　単語 w_i は文節の末尾とならない

文節内の各単語 $w_i, \cdots, w_n$ の満たす性質は次のようにまとめることができる．

〈文節の性質〉

$C(w_i, w_{i+1})=1$　$1\leq i\leq n-1$

$T(w_n)=1$

逆にこの性質を満たす単語列を文節と呼ぶこともできるが，ここでは形式的に文節となり得る単語列（または文字列）という意味で，文節形と呼ぶ．

3. べた書き文の分かち書き

3.1 文節形抽出

べた書き文の分かち書きは，文字列の分離およびその部分列が文節の性質を満たすかどうかの認定を行わなければならない．すなわち上で述べた文節形を文字列から分離，抽出する必要がある．次に文節形の文字列からの抽出処理について述べる．

文節形の抽出には，

（1） 自立部の抽出

（2） 自立部に接続する付属語列の抽出

の二つの処理を必要とする．自立部には自立語および接頭・接尾語の付いた複合語* をも許しているが，ここでは簡単のために自立部は自立語のみからなるものとして話を進める．

文字列からの自立語部分の分離，抽出は自立語辞書**の見出しとの最長一致を考慮した方法を用いる．すなわち文字列と辞書見出しとの照合において，最長一致が採れる見出しを持つ単語（自立語）を文節形の自立語部として設定する．しかしながら活用語尾を持つ用言は不変化部分が辞書見出しとして採用されているので，最長一致のみの探索では，このような用言などの抽出が困難なことから，より短い見出しを持つ単語をも探索する***．このように，最長一致の単語およびその見出しに含まれる単語が，文節形の自立語部として記憶される．さらに探索された自立語によっては，その品詞に対応する活用語尾表と後続の文字列との照合を行うことにより，正確な自立部の分離，抽出をする．続いて，求められた自立部の後続の付属語の探索を付属語辞書，接続行列****を用いて行い，もし接続可能な付属語と認定されれば，文節の末尾となるかを調べる．順次残余の文字列に対して，この付属語の認定処理をくり返し，一つの自立部に対する最長の付属語列を求める．結局，ある文字位置からの最長の文節形とそれに含まれる文節形の分離，抽出が行われる*****．**図 2** に文節形抽出の概略の流れ図を，**図 3** に抽出例を示す．

3.2 二文節最長一致法

文節形を抽出した段階で分かち書きを行う方法として，最長の文節形によって文字列の分離を行うことが考えられるが[8]，そのようにするとその付属語部に続く文節の自立語の先頭の文字列の一部も求める文節形に含まれる誤った分かち書き，いわゆる「ぎなた読み」の分かち書きになるおそれを生じ，続く文字列の解析を必要とする場合が起る．したがって，より正確な分かち書きを行うには，二文節の範囲にわたって解析を行った後に，二文節の境界を定める二文節最長一致法を用いる必要がある．

二文節最長一致法とは，連続する二文節（形）の長

---

* 例えば，「日本人」では「日本」と「人」の結合した複合語として処理される．

** 以下で用いられる辞書，表については 4.1 節参照．

*** 自立語辞書項目にさらに短い見出しを含むか否かの情報，すなわち探索を継続するか否かの情報が記載されている．

**** 前章の関数 $C(w_i, w_j)$ を行列表現したもの．

***** 前章の文節の性質を満足する文字列のみが抽出される．

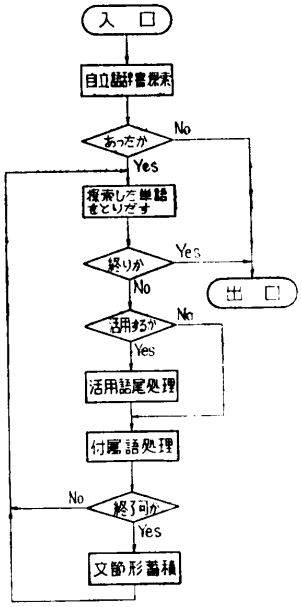

図 2 文節形抽出の流れ図

Fig. 2 The flow-chart of extraction process of segments.

(文字列) ソウイウ……

(文節形) ① ソウイ (名詞)
② ソウ・イ・ウ (副詞+補助用言語幹+活用語尾)
③ ソ・ウ (動詞語幹+活用語尾)

図 3 文節形抽出の例

Fig. 3 An example of extraction of segments.

さを分かち書きの尺度とし，二文節形として最長の解釈を与える区切りを二文節間の境界と認定する方法である．求める区切りとしては二文節間の境界のみを決定し，後半の文節そのものは決定せず，さらに後続の文節形によって，順次決定を行っていく．したがって二文節最長一致法は文節の決定に後続の文字列中に存在する文節形を用いる方法といえる．図4に二文節最長一致による文字列の解析例を示す．同図(a)では，二文節としての解析が成功する(a-1)と(a-3)の長さを比較することによって，(a-1)の分かち書きを採り，同図(b)では，(b-1)の分かち書きを採る．一方，同図(c)の例のように二文節形の長さが等しい場合には，後半の文節形の自立部に最長一致を適用し，(c-2)で示される分かち書きを採用する．しかし，後半の文節形「ハナレテ」はさらに「ハナレ△テ…」の可能性があるために，「ハナレテ…」以降の分かち書きは続く文節形を抽出することにより決定される．以上の二文節最長一致法を形式的に述べると次のようになる．

* 最長の二文節形が1通りの場合 (図4(a)参照)
** 最長の二文節形が複数個存在する場合 (図4(b)参照)

(a) ソウイウザッシヲ……
(a-1) ソウイウ△ザッシヲ△…
(a-2) ソウイ△…
(a-3) ソウ△イウ△…
(b) ソウイクフウヲ……
(b-1) ソウイ△クフウヲ△…
(b-2) ソウ△イク△…
(c) カゾクトハナレテ……
(c-1) カゾクトハ△ナレテ△…
(c-2) カゾクト△ハナレテ△…
(注) △は区切りを示す．

図 4 二文節最長一致による解析例

Fig. 4 Segmentation process by the longest-matching of two successive segments.

いまm番目の文節に対応するp番目の解釈可能な文節形を$B_m(p)$とし，文頭からの$B_m(p)$の開始文字位置，終了文字位置をそれぞれ$S_m(p)$，$E_m(p)$とすれば，$B_m(p)$の長さ$l(B_m(p))$は，

$$l(B_m(p))=E_m(p)-S_m(p)+1$$

で表わされる．同様に，$B_m(p)$に連続して見出される$(m+1)$番目の文節に対応するq番目の解釈可能な文節形を$B_{m+1}(q)$，その長さを$l(B_{m+1}(q))$とすると，連続する二文節形$B_m(p)$，$B_{m+1}(q)$の長さは，

$$\begin{aligned} l((B_m(p),\ B_{m+1}(q)))&=l(B_m(p))+l(B_{m+1}(q)) \\ &=E_{m+1}(q)-S_m(p)+1 \end{aligned}$$

で表わされる．m番目の文節の開始文字位置は一定であるとすると，二文節形の長さは後半の文節形の終了文字位置によって評価される．したがって，m番目の文節$B(m)$の二文節最長一致による決定規則は，

(i) $E_{m+1}(k)>\max_{j\neq k} E_{m+1}(j)$

なるkが存在する場合*，

$$B(m)=\{B_m(i)\mid E_m(i)+1=S_{m+1}(k)\} \qquad (1)$$

(ii) $E_{m+1}(j_1)=E_{m+1}(j_2)=\cdots=E_{m+1}(j_r)$

なる場合**，

$$B(m)=\{B_m(i)\mid \min_{1\leq i\leq r} E_m(i)\} \qquad (2)$$

と書ける．ただし，(ii)の場合，構文的により妥当だと考えられる二文節の境界を採用する．例えば，

(a)言語に△合った (名詞・助詞△動詞・助動詞)
(b)言語△似合った (名詞△動詞・助動詞)

の2通りの解釈がなされる場合，(2)式による分かち書きは(b)であるが，構文的に望ましいものとして(a)が求められる．このように構文情報の活用が二文節最長一致の枠内でなされる．

二文節最長一致によって求められた文節は，品詞列として決定されるため，この時点でも単語が一意に定まらない場合には，それらの単語の使用頻度によって

決定される．(4.4 節参照)

3.3 複合語処理

実際の文章では文節の自立部として，接頭語，接尾語，助数詞を含む語がしばしば出現する．これらの語は自立語の辞書見出しとして採用されていないことから，複合語*として処理する必要が生ずる．

複合語はその文法的特徴が少なく，接頭語，自立語，接尾語の順序列という枠組のみでは，辞書探索の結果，多くの偽りの複合語を求めてしまう．とくに未登録語などによって，他の文節形が見出せない場合に，偽りの複合語が採用され，誤った分かち書きを行うことになる．このような複合語処理における偽りの複合語の発生を抑える方法として，漢字構成の面からみた複合語の次の性質を利用する．

「接頭語，接尾語はほとんどの場合，漢字2字以上の自立語と結合する．」

(a) 接頭語処理

自立語の抽出の際，まず接頭語辞書を探索し，接頭語と同音となる文字列があればその文字列で分離し，さらに自立語辞書探索を行って，漢字2字以上の自立語が存在すれば，一つの複合語として文節形の自立部候補とする．さもなければ，複合語としての文字列の分離，抽出は行わない．

(b) 接尾語処理

接尾語の処理は接頭語の場合とは逆に，漢字2字以上の自立語の分離が行われれば，接尾語辞書を探索して，複合語の分離，抽出を行う．大部分の接尾語は名詞の品詞属性を持つと考えられるが，状態を示す接尾語「～的」，「～風」などは形容動詞として扱い，活用形の処理を行う．なお一部の単語「私達」，「人達」などは，それぞれ「人」，「私」に接尾語「達」が結合した複合語と考えられるが，上述の性質を効果的に用いるために，それぞれ自立語として扱っている．

(c) 助数詞

入力文字列中に数字または数字列が出現すると，後続の文字列と助数詞辞書との照合を行う．

以上の各処理は文節形の抽出の段階でなされ，各辞書と文字列との一致が採れた時点ではあくまで解釈可能な自立部の候補として扱い，実際の文節は二文節最長一致によって決定を行う．

3.4 未登録語の処理

これまで述べてきた分かち書き方法は自立語辞書で探索された語を，いわゆる「核」として，それに接続する付属語を求めることによって，分かち書きを行うものであった．しかし文中に未登録語**が存在する場合には，分かち書きの前提となる文節形の解析および抽出ができず，したがって未登録語を含む文字列以降の分かち書きができなくなる．このように未登録語はべた書き文の分かち書きを行ううえで大きな障害となる．この問題を解決する一つの方法は自立語に依存しない分かち書き，すなわち付属語のみによる分かち書き方法を用いることである[8]．

付属語はその接続する語の文法的性質，すなわち体言か用言かによって大きく二つに分類され，用言に接続する付属語はその接続する語の活用形により，さらに細分類される．この分類クラスを

(i) 主に体言に接続する付属語……A

(ii) 主に用言に接続する付属語

被接続語の活用形 未然形……B

連用形……C

終止形……D

仮定形……E

とすると，文字列の付属語の判定は次のようになる．

〈付属語の判定〉

(i) クラスAの付属語と同音の文字列（以後付属語単位と呼ぶ）はすべて付属語と見なす．

(ii) クラス$B, \cdots, E$に属する付属語単位はその直前の1文字が**表1**中のそれぞれのクラスに対応する文字であれば，付属語と見なす．

これは用言の各活用語尾が表1に示す各文字で終了することを用いている．なお付属語の判定に用いた付属語とそのクラスおよびその付属語を先頭とする付属語列の頻度***を**表2**に示す．

結局，付属語による分かち書きアルゴリズムは以下のステップで示される．

表1 動詞の活用と活用語尾

Table 1 Conjugation and inflectional characters of verbs.

活用形	活用語尾
未然形	カガサタナハマラワイキシチニヒミリエケセテネヘメレオコソトノホモロ
連用形	イキッギシチニンヒミリエケセテネヘメレジゲゼデ
終止形	クグスツヌブムルウ
仮定形	ケゲセニネベレエメ

* ここで扱う複合語は接頭語，接尾語などが自立語と結合してできる語をさしており，二つ以上の自立語からなる語は複合語として含めていない．

** 自立語辞書に存在しない語．

*** 資料[9]による頻度．

表 2 付属語の接続による分類

Table 2 Classification on connectability of dependent-words.

先頭の付属語	接続	度数	割合* (‰)	先頭の付属語	接続	度数	割合* (‰)
の	A	12842	173.2	や	A	360	4.9
に	A	9411	126.9	う	B	345	4.7
て	C	7760	104.6	など	A	300	4.0
を	A	7077	95.4	だけ	A	271	3.7
は	A	5715	77.1	ず	C	219	3.0
た	C	5434	73.3	でも	A	208	2.8
が	A	4803	64.8	より	A	180	2.4
だ	A	3054	41.2	ながら	C	176	2.4
で	A	2176	29.3	たら	C	173	2.3
と	A	1826	24.6	ん	B	173	2.3
も	A	1549	20.9	たり	C	165	2.2
ない	B	1480	20.0	し	D	69	0.9
ます	C	1435	19.4	らしい	A	57	0.8
から	A	1168	15.7	べき	D	46	0.6
です	A	738	10.0	なく	C	41	0.6
へ	A	483	6.5	ばかり	A	32	0.4
か	A	428	5.8	しか	A	20	0.3
ば	E	413	5.6	たる	A	20	0.3
まで	A	403	5.4	計		71050	958.1

* (‰) は千分率を表わす.

(i) 解析不能となった文字列以降から，表2で示される付属語単位を捜す.

(ii) 上記の付属語の判定を行う.

(iii) 付属語と判定されれば，その付属語を付属語列の先頭または一部として最長の付属語列を求め，続く文節形の抽出を行う.

4. 実験システム

4.1 辞 書

分かち書きおよび漢字変換に用いる辞書を**表3**に示す．各辞書についての概略を以下に記す.

(a) 自立語辞書

(i) 語数…一般的な文章を扱うには最低限数千語は必要であることから，資料[10]の語い表の見出しより固有名詞，人名を除いた約 5500 語と従来の実験システムで使用していた辞書からの単語，約 2300 語を付加した 7800 語余りが収録されている.

表 3 分かち書きおよび仮名漢字変換用辞書

Table 3 The dictionaries for segmentation and transformation from Kana to Kanji.

(a)	自立語辞書
(b)	接続行列
(c)	用言の活用語尾表
(d)	付属語辞書
(e)	接頭語，接尾語，助数詞辞書
(f)	付属語分かち書き用辞書

カナ見出し	漢字コード	漢字数	品詞	活用	頻度	意味分類	継続情報

図 5 自立語辞書の形式

Fig. 5 A record format of an entry in the independent-word dictionary.

(ii) 形式…各単語は**図5**に示す情報を持ち，見出しの五十音順に配列されている．単語情報の内容について以下に示す.

- 見出し…いわゆる仮名見出しで，通常の国語辞典と同様なものであるが，用言については不変化部分(語幹)のみを見出しとしている.
- 漢字コード…漢字印刷機用のコードで，漢字コードは ISO 符号2字の組合せで表現されている.
- 漢字数…漢字コードの個数を示している．この情報は複合語処理における漢字の字数としても用いられる.
- 品詞…付属語の接続，活用の有無によって以下の分類を行っている.
  1. 接続詞，感動詞
  2. 連体詞
  3. 副詞
  4. 名詞
  5. サ変名詞
  6. 形容動詞
  7. 形容詞
  8. 動詞
- 活用…動詞の活用形処理に必要な分類を用いている.
  1. 五段活用
  2. 一段活用（下一，上一段）
  3. サ行およびカ行変格活用
- 頻度…資料[10]による頻度の順序づけを行ったもの.
- 意味分類コード…資料[10]の意味分類コードをそのまま採用している．ただし，ここでは未使用.
- 継続情報…見出しに含まれるさらに短い見出しの長さ（文字数）を示している．これは最長一致探索の欠点を補うために設けられたもので，直接見出しのポインタを採用しなかった理由は，辞書の追加，修正によっても変更する必要がないからである.

(iii) 辞書の統計…辞書の内訳として，品詞分類，同音語数（見出しの一致のみ）をそれぞれ**表4**，**表5**に示す.

(b) 接続行列

自立語，付属語および付属語間の接続規則を行列形

表 4 自立語辞書の品詞分類

Table 4 Classification of parts of speech in the independent-word dictionary.

品詞	語数	比率 (%)
接続詞，感動詞	80	1.0
連体詞	19	0.2
副詞	302	3.7
名詞	4468	54.3
サ変名詞	1546	18.8
形容詞	153	1.9
形容動詞	418	5.9
動詞	1240	15.1
計	8226†	100.1

† 多品詞語を含む

表 5 自立語辞書の同音語数

Table 5 Numbers of hononyms in the independent-word dictionary.

重複度	見出し語数	比率†† (%)
1	4017	51.4
2	798	20.4
3	289	11.1
4	105	5.4
5	53	3.4
6	23	1.8
7	13	1.2
8	7	0.7
9	11	1.3
10	5	0.6
11	5	0.7
12	2	0.3
13	4	0.7
14	0	0.0
15～	5	1.0
	5337	100.0

†† 比率＝(見出し語数×重複度/単語数)×100

式で表現したもので，被接続単位を行に，接続単位を列に対応させている．ここで用いた行列の大きさは 154×108 である．

(c) 用言の活用語尾表

動詞，形容詞，形容動詞およびサ変名詞における活用語尾とそれらの接続行列における行番号の対応表である．

(d) 付属語辞書

助詞，助動詞，形式名詞，補助用言および形式助詞(～おける，～おいてなど)とそれらの接続行列との対応を示す辞書である．

(e) 接頭語，接尾語，助数詞辞書

各辞書は見出しと漢字コードからなり，収録語数は

* 付属語と同音の文字列．

** 直前の文節の付属語との接続性が満足される必要がある．

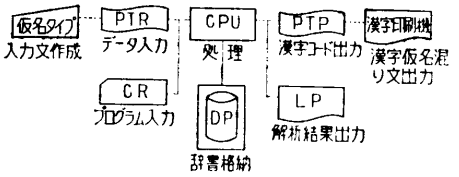

図 6 ハードウェア構成

Fig. 6 Hardware configuration.

それぞれ 35 語，147 語，141 語である．なお，これらの語は資料[12]の接辞の表より採録した．

(f) 付属語分かち書き用辞書

3.5 節の表 2 に掲げた付属語を採用している．この辞書と接続行列によって，文中のほとんどの付属語列を検出することが可能である．

4.2 システム構成

実験に用いたハードウェア構成を図 6 に示す．出力としての漢字仮名混り文は漢字コード列として紙テープに一度出力され，オフラインで漢字印刷機を動作させる．なお，使用計算機は FACOM 230-45S，使用言語は PL/I である．

4.3 辞書探索回数の制限

文字列から取り出される最長の文節形が正しい文節と必ずしも一致しない理由は，続く文節の自立語の一部を含むことによる．したがって，この原因となる自立語，すなわちその先頭の一部が付属語単位* から構成される自立語についてまとめたものが表 6 である．付属語単位 4 以上に分解される自立語数は 3 であるが，実際の文章中にこれらの語が出現しても分かち書きに影響を与えないことから**，最長の文節形と正しい文節とのずれは付属語単位数にして 3 以内と考えられる．以上のことから一つの最長の文節形に含まれる切れ目に対し，最大 4 通りの異なる文字位置に対して辞書探索を行えばよいことになる．具体例を図 7 に示す．図中，①～④で示される切れ目に対し，続く文節形の抽出，すなわち自立語辞書探索を試みる．このように実際の二文節最長一致法の適用にあたっては，辞書探索の回数を制限することによって処理時間の短縮をはかっている．

表 6 自立語の付属語による分解

Table 6 The decomposition by dependent-word strings.

付属語単位数	2	3	4	5
分解可能な見出し語数	457	44	3	0

(a) ホウソクニシタガウコトハ…
ホウソク・ニ・シ・タ・ガ△ウ…
④ ③ ② ①

(b) オコナワナケレバナラナイトシテモソレハ…
オコナワ ナケレ バ・ナラ ナイ・ト・シ・テ・モ△ソ…
④ ③ ② ①

(注) 記号△は最長の文節形による区切り，
・は文節形内の切れ目，下線は単語を示す．
番号①,…,④は自立語辞書探索を行う切れ目を示す．

図 7 最長の文節形と辞書探索の具体例

Fig. 7 Segments for the independent-dictionary search.

4.4 分かち書きおよび仮名漢字変換

入力文を読み込み，まず句読点によって分割し，その分割された文字列（以後区分と呼ぶ）に対し，文節形抽出，二文節最長一致法を適用して，順次分かち書きを行っていく．同時に決定された文節に対応する漢字化がなされる．同音異字が存在する場合にはそれらの頻度最高のものが出力されるが，いずれも頻度が低い場合には，辞書中の最初の単語がさし当り出力されている．すでに決定された区切りから連続して二文節が見出されない場合には，区切りからの最長の文節形を文節として，付属語分かち書きによって後続の文字列の区切りを求め，処理を続行する．**図 8** にべた書き文の仮名漢字変換の流れ図を，**図 9** に変換例を示す．

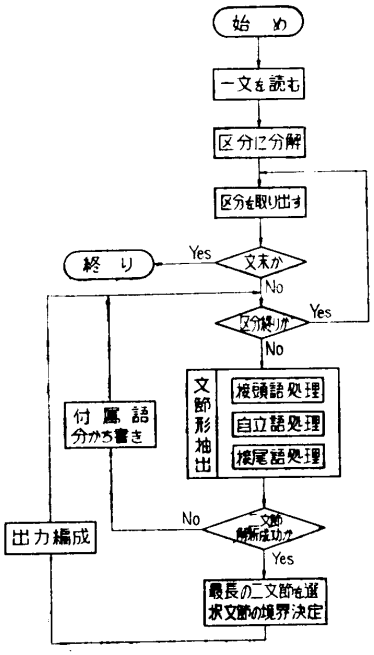

図 8 分かち書きおよび漢字変換の流れ図

Fig. 8 The flow-chart of segmentation and transformation.

(1) べた書き文の分かち書き例

ニッポン コクミンハ，セイトウニ センキョサレタ コッカイニオケル ダイヒョウシャヲ ツウジテ コウドウシ，ワレラトワレラノ シソンノタメニ，ショコクミントノ キョウワニヨルセイカト，ワガクニ ゼンドニ ワタッテ ジユウノ モタラスケイタクヲ カクホシ，セイフノ コウイニヨッテ フタタビセンソウノ サンカガ オコルコトノ ナイヨウニスルコトヲ ケツイシ，ココニ シュケンガ コクミンニ ソンスルコトヲ センゲンシ，コノ ケンポウヲ カクテイスル．
ソモソモ コクセイハ，コクミンノ ゲンシュクナ シンタクニヨルモノデアッテ，ソノ ケンイハ コクミンニ ユライシ，ソノケンリョクハ コクミンノ ダイヒョウシャガ コレヲ コウシシ，ソノ フクリハ コクミンガ コレヲ キョウジュスル，コレハ ジンルイ フヘンノ ゲンリデアリ，コノ ケンポウハ，カカル ゲンリニ モトヅクモノデアル．
ワレラハ コレニ ハンスル イッサイノ ケンポウ，ホウレイオヨビ ショウチョクヲ ハイジョスル．
ニッポン コクミンハ，コウキュウノ ヘイワヲ ネンガンシ，ニンゲン ソウゴノ カンケイヲ シハイスル スウコウナ リソウヲ フカク ジカクスルノデアッテ，ヘイワヲ アイスル ショコクミンノ コウセイト シンギニ シンライシテ，ワレラノアンゼント セイゾンヲ ホジショウト ケツイシタ．
ワレラハ ヘイワヲ イジシ，センセイト レイジュウ，アッパクト ヘンキョウヲ チジョウカヲ エイエンニ ジョキョショウト ツトメテイル コクサイ シャカイニオイテ，メイヨ アル チイヲ シメタイト オモウ．
ワレラハ ゼンセカイノ コクミンガ ヒトシク キョウフトケツボウカラ マヌガレ，ヘイワノウチニ セイゾンスル ケンリヲ ユウスルコトヲ カクニンスル．
ワレラハ，イズレノ コッカモ，ジコクノコトノミニ センネンシテ タコクヲ ムシシテハ ナラナイノデアッテ，セイジ ドウトクノ ホウソクハ，フヘンテキナモノデアリ，コノ ホウソクニ シタガウコトハ，ジコクノ シュケンヲ イジシ，タコクトタイトウ カンケイニ タトウトスル カッコクノ セキムデアルト シンズル．
ニッポン コクミンハ コッカノ メイヨニ カケ，ゼンリョクヲ アゲテ コノ スウコウナ リソウト モクテキヲ タッセイスルコトヲ チカウ．

(2) 対応する漢字仮名混り文出力

日本国民は，政党に選挙された国会における代表者を通じて行動し，我等と我等の子孫のために，諸国民との共和による成果と，我国全土に渡って自由のもたらす恵沢を確保し，政府の行為によって再び戦争の参加が起こることの内容にすることを決意し，ここに主権が国民に存することを宣言し，この憲法を確定する．そもそも国政は国民の厳粛な信託によるものであって，その権威は国民に由来し，その権力は国民の代表者がこれを行使し，その福利は国民がこれを教授する．これは人類普遍の原理であり，この憲法は，かかる原理に基づくものである．我等はこれに反する一切の憲法，法令及び詔勅を排除する．日本国民は，高級の平和を念願し，人間相互の関係を支配する崇高な理想を不覚自覚するのであって，平和を愛する諸国民の諸国民の構成と審議に信頼して，我等の安全と生存を保持しようと決意した．我等は平和を維持し，先生と隷従，圧迫と辺境を地上から永遠に除去しようと勤めている国際社会において，名誉ある地位を占めたいと思う．我等は全世界の国民が等しく恐怖と欠乏から免れ，平和のうちに生存する権利を有することを確認する．我等は，いずれの国家も，時刻のことのみに専念して他国を無視してはならないのであって，政治道徳の法則は，普遍的なものであり，この法則に従うことは，時刻の主権を維持し，他国と対当関係に立とうとする各国の責務であるとシンズル．日本国民は国家の名誉に掛け，全力を上げてこの崇高な理想と目的を達成することを誓う．

(注) 片仮名は未登録語として付属語による分かち書きによって処理されたことを示す．

図 9 出力例

Fig. 9 Output examples.

表 7 実験結果
Table 7 Results of experiments.

	分かち書き†		漢字変換††	
	文節数	割合 (%)	文節数	割合 (%)
正	2519	97.2	2147	92.5
誤	73	2.8	175	7.5

† 未登録語を含むすべての文節を対象.
†† 正しく分かち書きされた文節より未登録語，固有名詞などを含む文節を除いた文節を対象.

表 8 分かち書きの誤り例
Table 8 Examples of segmentation errors.

原因		誤り例 (正解)
二文節最長一致	(1)	勝れ為と (勝れた目と)
	(2)	～のか低労働 (～の家庭労働)
	(3)	～をか基底酒ツスル (～を書き提出する)
未登録語の出現	(4)	タメンテ奇な (タメンテキナ)
	(5)	党よ牛の (トウヨウシノ)
	(6)	ポツポツミエテイテ (ポツポツ見えていて)
	(7)	様式とし手のキハンガ (様式としてのキハンガ)

(注) 片仮名は未登録語または付属語分かち書きで処理したことを示す.

4.5 実験結果

入力データは任意の分野から選んだ214文で，総文節数は 2592 文節である．なお固有名詞，外来語などは辞書に収録されていないことから，一応区切り記号を付加しているが，それ以外は原文の表音表記の仮名文となっている．

分かち書きおよび漢字変換の処理結果をまとめて表7に示す．なお処理時間は一文節当たり平均989 msであった．

分かち書きの誤りは，二文節最長一致による誤りと未登録語による誤りに分けられ，二文節最長一致の誤りとしては，二文節として同じ長さの解釈がなされる場合に，自立語部の長さの優先という規則 (3.2 節(2)式参照) によるもの(表8(1))と複合語処理 (接頭語，接尾語) での誤りが複合したもの(表8(2))が挙げられる．なお辞書中に該当する単語が存在するにもかかわらず，正しく処理できずに付属語分かち書きによって処理した例は，表8(3)の1例のみであった．

未登録語を含む文節が総文節の 4.4% 存在したが，そのうち正しく分かち書きされた文節*は 49.6% で，結果的に正しく分かち書きされた文節**が27.0%であった．

* 未登録語として処理された文節で，図8の例 (シンズル) のように片仮名で示された文節.
** 異なる解釈を行ったものを含む．例えば，語って→方って，などが挙げられる.

一方，誤りは，一文節を二つの文節に分割したものが 11.3% (表8(4)，(5))，区切り誤り，付属語が存在しないために次の文節で区切ったものなど未登録語を含む文節以外に影響を及ぼしたものが 12.2% であった (表8(6)，(7))．

以上分かち書きの誤りの多くが未登録語の影響を受けたもので，二文節最長一致による誤りは，総文節数の 1.2% であった．

漢字変換については，正しく分かち書きされた文節中で，未登録語，固有名詞などを除く文節を対象とし，その 71.4% が品詞列分解で一意に定まり，残りの 21.1% が頻度を用いた同音異字選択によるものであった．

5. あとがき

べた書き文の仮名漢字変換の基本となる分かち書き方法として，二文節最長一致法を，また未登録語の処理として付属語分かち書きを提案し，これら二つの処理を用いたべた書き文の仮名漢字変換システムについて報告した．

二文節最長一致法の特徴は，二文節によって区切りを確定していくために，区切り誤りの少ないこと，未登録語の影響によって細分割されすぎないことが挙げられる．一方付属語分かち書きは，網羅的な文字の切り出しによる処理に比べて，付属語リストによって処理するので，処理速度が速く，さらに実際の文節では4文字以上の見出しをもつ自立語が多いことから効果的に区切りを求めることができる．

実験例によれば，べた書き仮名文をこの方法によって 97.2% の成功率によって文節に分けることができた．さらに正しく分けられた文節の 92.5% が正しい漢字に変換される結果となったが，これに関しては同音異字に対する処理がまだあまり施されていないので，その良否の評価は必ずしも本稿の意図した対象ではない．しかしながら，分かち書き誤りの大半が二文節にわたる同音異義の問題と考えられることから，変換過程における同音語選択の問題とともに，分かち書きにおける曖昧さの解消には意味を考慮に入れた処理の導入が必要となろう．

本実験システムを基礎にして，人手による字句の修正変更機能を持つ実用的なべた書き文の仮名漢字変換システムはすでに試作されており[13]，その修正方法，未登録語の漢字化の問題などについて検討を行っている．

本研究に当たり，辞書作成など多くの尽力を頂いた大学院生岡田真和君に感謝するとともに，日頃御鞭撻を頂いております豊田順一助教授に深謝します．

参 考 文 献

1) 木澤：漢字の入力，電気学会誌，Vol. 97, No. 2, pp. 90-92 (1977).
2) 栗原，黒崎：仮名文の漢字混り文への変換について，九州大工学集報，Vol. 39, pp. 659-664 (1967).
3) 松下，山崎，佐藤：漢字カナ混り文変換システム，情報処理，Vol. 15, No. 1, pp. 2-9 (1974).
4) 情報処理振興事業協会資料：漢字かな混り文変換プログラム (1972).
5) 相澤，江原：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).
6) 牧野，勝部，木澤：カナ漢字変換の一方法，情報処理，Vol. 18, No. 7, pp. 656-663 (1977).
7) 岩波講座日本語 6，岩波書店，東京 (1976).
8) 牧野，木澤：べた書き文のカナ漢字変換，電子通信学会技報，PRL 77-27 (1977).
9) 国立国語研究所：現代雑誌 90 種の用語用字 (3)，分析，秀英出版，東京 (1969).
10) 国立国語研究所：現代雑誌 90 種の用語用字 (1)，総記および語彙表，秀英出版，東京 (1969)
11) 国立国語研究所：分類語彙表，秀英出版，東京 (1973).
12) 国立国語研究所：計算機による新聞の語彙調査 (II)，秀英出版，東京 (1971).
13) 岡田，牧野，木澤：べた書き文のカナ漢字変換システム，情報処理学会第 19 回全国大会講演論文集，pp. 445-446 (1978).

(昭和 53 年 10 月 17 日受付)
(昭和 54 年 3 月 15 日採録)