資　料

カナ漢字変換の一方法*

牧野　寛**　勝部康人***　木澤　誠**

Abstract

In the computer handling of Japanese sentences described normally with *kanji* (ideographic characters) and *kana* (phonetic characters), transformation by computer from *kana* to *kanji* in the input sentences presented exclusively in *kana* is one of the hopeful input methods. For doing this the authors propose the "independent-dependent word segmentation", by which the input sentences are segmented with one space each between an independent word and a dependent word.

The transformation process is in three steps: (1) elimination of non-transformable segments referring to the specified list and modifying the results, (2) determination of the parts of speech of segments by analyzing their functions, and (3) consultation with the dictionary for transformation from *kana* to *kanji* with the aid of the parts of speech.

As a result of some experiments, about 92 per cent of the transformable segments are successfully trasformed into *kanji*.

1. まえがき

計算機によって日本語を扱う場合の最大の難点は，計算機への入力の問題であろう．この問題を解決する手法は，大きく2つに分けることができる．1つは漢字仮名文字を直接読み取る認識の問題として考え，これに対する有効な手法を開発する「直接方式」であり，他方は仮名タイプ，漢字タイプ等で入力する「間接方式」である．ここでいう「直接」，「間接」の意味は入力過程において原理的に人間の介在を許すか否かの相違である．前者は後者に比べて人間にとって，より強力な方式といえるが，一般に複雑な処理形態をとり専用の処理機器，読取装置等が必要であり，通常の計算機システムでの実行可能性という点において，汎用性に乏しい．一方後者はさらに2つに分けられ，その一方の漢字鍵盤タイプは高速に打鍵（または入力）するためにはかなりの習熟を要し，誰もが能率よく入力することは困難であるといえる．他方仮名タイプは欧文タイプと同程度の速度で打鍵することが可能であり，欧文タイプと同様の操作性を持っている．さらにカード穿孔機，紙テープさん孔等の入力装置を用いることができ，通常の計算機システムで入力可能という意味で，非常に汎用性が高いが，仮名漢字変換という日本語特有の難点を持つ．

本稿では，仮名漢字変換について，入力形式及びそれに適した仮名漢字変換アルゴリズムを考える．

2. 自立語付属語分かち書き

現在までに報告されている仮名漢字変換は[1),2)]，識別のための特殊文字（制御記号）を挿入しなければならないという不便さを伴い，使用者に少なからず負担を強いる結果となっている．

仮名漢字変換は仮名入力文の形式と大いに関係し，これは仮名漢字変換システムにおける打鍵速度の差，変換の正確さの差となって現われてくる．ここで，べた書きを含めて従来用いられてきた分かち書きの方法を以下の 1)～5) にまとめると，

1) 空白****を全く用いないいわゆるべた書き
2) 恣意的に空白を挿入する分かち書き

---

* Transformation of kana-input into kanji-presented sentences by Hiroshi MAKINO, Makoto KIZAWA (Faculty of Engineering Science, Osaka University) and Yasuto KATSUBE (Graduate School of Osaka University)

** 大阪大学基礎工学部情報工学科
*** 大阪大学大学院基礎工学研究科

**** 1字分の間隔（スペース）の意味に用いる．

…最小の区切り単位は文節*とし，複数個の文節が連続することを許した分かち書き

3) 文節分かち書き

…文節単位ごとに空白が挿入される分かち書き

4) 単語分かち書き

…単語ごとに空白が挿入される分かち書き

5) 文字種分かち書き

…文字種ごとに空白が挿入される分かち書き

などが挙げられる．1),2),3) 及び 5) のそれぞれは仮名漢字変換を必要とする部分の抽出に難点があり，4) は入力規則が複雑となる．このため筆者らは次のような分かち書き方法を提案する[3]．

6) 自立語付属語分かち書き

この規則は自立語**と付属語***との間に空白を挿入するという原則で示される．但し，2個以上の自立語が連続し，それらが1語となって意味をなす場合以外は，自立語間に空白を挿入するという規則を含んだ分かち書きの方法である．**Fig. 1** に上記の 1)～6) に対応した例を示す．

自立語付属語分かち書きを用いる仮名漢字変換の利点は以下にまとめることができる．

a) 入力規則が簡単であるから，自立語か付属語かの区別さえつけばよいので，正確な文法知識を要求せずオペレータの負担が軽い．

b) 仮名漢字変換を要する部分の抽出が比較的容易である．

現在人類はすぐれた目と指先の感覚を持っている。

(1) ゲンザイジンルイハスグレタメトユビサキノカンカクヲモッテイル。

(2) ゲンザイ ジンルイハ スグレタメト ユビサキノカンカクヲ モッテイル。

(3) ゲンザイ ジンルイハ スグレタ メト ユビサキノ カンカクヲ モッテイル。

(4) ゲンザイ ジンルイ ハ スグレ タ メ ト ユビサキ ノ カンカク ヲ モッ テ イル。

(5) ゲンザイジンルイ ハスグレタ メ ト ユビサキ ノ カンカク ヲ モ ッテイル。

(6) ゲンザイ ジンルイ ハ スグレ タ メ ト ユビサキ ノ カンカク ヲ モッ テイル。

Fig. 1 Examples of segmentations in a Japanese sentence.
(1) without segmentation (2) segmented between phrases (3) segmented after auxiliary words (4) segmented by words (5) segmented between *kanji* and *kana* (6) segmented between independent and dependent words

* 実際の言語として，文を不自然でない程度に区切った最小の単位．
** それだけで1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，ここでは助詞，助動詞，補助用言をさす．
**** 未然，連用，終止，連体，仮定，命令の各形，
***** したがって，仮名入力文は複数個の区切語とそれらを区切る空白及び句読点からなるといえる．

c) 付属語の性質すなわち付属語によって与えられる自立語の品詞情報を用いることができる．

3. 仮名漢字変換と品詞判定

仮名漢字変換の方式としては，変換単位によって，漢字1文字を変換対象とする語単位変換と，意味をなす語列を変換対象とする熟語単位変換とが挙げられるが[1]，本システムでは，仮名入力文からの仮名漢字変換を必要とする部分の抽出が自立語を単位としたものになることから，後者の変換方式を採るものとする．

自立語は名詞(体言)，副詞，連体詞及び感動詞からなる活用しないすなわち語形変化のない自立語と，用言すなわち動詞，形容詞及び形容動詞からなる活用する自立語に分けられ，後者は語尾が規則的に変化し，それによって6つの活用形を持つ****,[4]．このことから終止形のみを辞書の見出しとして，変形規則を用いれば，全ての活用形を辞書の見出しとする場合に比べて 1/6 の規模の辞書とすることができる．ところで本変換システムでは簡単にしかも高い正解率で終止形から他の活用形の仮名漢字変換ができることから，辞書探索方法として，活用する自立語と活用しない自立語の場合に異なる探索方法を採っている．したがって効率よく仮名漢字変換を行うために，自立語が活用するか否かという簡単な品詞判定を行うだけでよい．

いま自立語付属語分かち書きを用いた仮名文において空白及び句読点で区切られた仮名文字列を区切語(くぎりご)と定義すると，品詞判定の前処理として自立語を抽出するために，区切語が自立語か付属語かをまず判定しなければならない*****．この判別方法として付属語リストを作成し，そのリスト中に入力された区切語がなければ，自立語とみなすという簡単なアルゴリズムを用いた．しかしながらこの方法は付属語と同音となる自立語は常に付属語とみなされるという欠点を持つ．**Table 1** (次頁参照) に付属語と同音となる自立語の例を示す．このように付属語リスト単独では，付属語と同音となる自立語が抽出できず，またそのような自立語が比較的高頻度で現われるため，全体としての誤りが多くなる．ところが自立語付属語分かち書きにおいては，連続する区切語が付属語として存在することはないという原則を利用して自立語抽出(付属語除去)の正解率を上げることができる．

いまリスト参照により，連続する区切語 w_i, w_{i+1} が付属語とみなされた場合に，

(i) w_{i-1} が自立語とみなされるならば，w_{i+1} は

Table 1 An example of hononymous words

同音語	付属語としての出現率* p_d(‰)**	自立語としての出現率* p_i(‰)**	自立語/同音語 $\frac{100\cdot p_i}{p_i+p_d}$ (%)
か	13.165	1.343	10.2
が	63.936	0.073	0.114
から	18.385	0.016	0.087
たて	70.814	0.524	0.740
とに	83.943	0.984	1.17
のは	62.954	0.255	0.041
ばへや	102.217	7.651	7.49
	151.109	0.039	0.026
	84.286	0.276	0.327
	6.984	0.135	1.93
	5.135	0.055	1.07
	4.332	0.472	10.9

* 文献 5) に基づく.
** ‰(パーミル) は千分率を表わす.

自立語,

(ii) w_{i-1} が付属語とみなされるならば, w_i は自立語,

(iii) w_{i-1}, w_i の区切り記号が読点ならば, w_i は自立語

と修正する. この修正アルゴリズムを採用することにより, 付属語リストを参照するという方法の自立語抽出率を大幅に改善することができる. 自立語抽出修正アルゴリズムの流れ図を **Fig. 2** に示す.

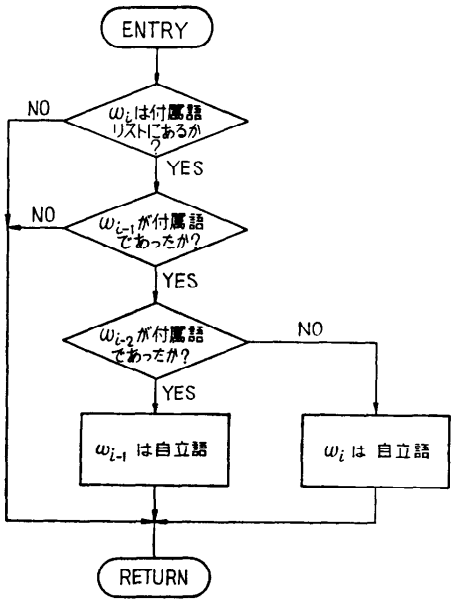

Fig. 2 The flow-chart of modified algorithm in extracting independent words.

4. 品詞判定処理

4.1 付属語の類別

付属語の持つ接続性についての統計的及び文法的性質を基準にして付属語の部分集合を作り, それらの被接続語 (すなわち自立語) の品詞判定を行うためにまず付属語の類別について述べる. 付属語の1つの性質すなわち活用する自立語及び活用しない自立語からの接続頻度を規準にして付属語集合Uを次のように類別する.

いま付属語を $x(\in U)$ とし, 活用しない自立語からの接続頻度を a_x, 活用する自立語からの接続頻度を b_x とすると, 活用しない自立語からの接続度 f_x は,

$$f_x=\frac{a_x}{a_x+b_x}$$

と定義でき(**Table 2** 参照), これを用いてUを次のように分割する.

$$\begin{cases} f_x>\theta_1 & \text{であれば} \quad x\in A \\ f_x<\theta_2 & \text{であれば} \quad x\in B \\ \theta_2\leqq f_x\leqq\theta_1 & \text{であれば} \quad x\in C \end{cases}$$

さらに集合Cを次のように分割する.

$$\begin{cases} \theta_3\leqq f_x\leqq\theta_1 & \text{であれば} \quad x\in C_1 \\ \theta_2\leqq f_x<\theta_3 & \text{であれば} \quad x\in C_2 \end{cases}$$

但し $\theta_1, \theta_2, \theta_3$ は $\theta_1>\theta_2>\theta_3$ なる関係を持つ分割パラメータである. 以上の分割の特性をまとめたものが **Table 3** (次頁参照) である.

いままで述べたことは, 付属語の後置詞としての性質をもとにした分割についてであったが, 次に付属語の前置詞的な性質について述べる. 付属語の中には, 被接続語となる場合に接続する自立語の品詞を指定す

Table 2 The frequencies of connections of independent words to dependent words

付属語 x	活用しない自立語からの接続頻度† a_x	活用する自立語からの接続頻度† b_x	接続度 $f_x=\frac{a_x}{a_x+b_x}$
か	321	142	0.693
が	4,635	219	0.955
から	946	114	0.892
たり	1	146	0.007
だ	533	2	0.996
だけ	55	15	0.786
だろう	47	63	0.427
である	469	1	0.998
です	409	19	0.956
と	1,032	590	0.636
ながら	17	159	0.097
など	107	2	0.987
にも	285	4	0.986
のが	2	112	0.018
のは	7	203	0.033

† 文献 6) による.

Table 3 Classification of dependent words by modes of connections

類別		性質	例
A		全てあるいはほとんど全ての場合活用しない自立語に接続する付属語	だ，です，など，にも，…
B		全てあるいはほとんど全ての場合活用する自立語に接続する付属語	たり，ながら，のが，のは，…
C	C_1	活用しない自立語に比較的多く接続する付属語	か，だけ，と，…
	C_2	活用する自立語に比較的多く接続する付属語	だろう，…

るものがある．

付属語「の」及び「～の」に接続する単語は名詞であることが多い．ここで「～の」は付属語「の」が語尾となる付属語を表わしている*．このことから集合Dを次のように定める．

$$D=\{\text{「の」},\text{「～の」}\}$$

また，付属語「を」，「に」，「をも」さらにそれらのそれぞれについての複合付属語「～を」，「～に」及び「～をも」** の次にくる単語は動詞であることが多い．したがってEなる付属語集合を次のように定める．

$$E=\left\{\begin{array}{l}\text{「を」},\text{「に」},\text{「をも」},\\ \text{「～を」},\text{「～に」},\text{「～をも」}\end{array}\right\}$$

以後仮名入力文の区切語総数をnとし，各区切語を文頭から $w_1, w_2, \cdots, w_n$ と表わすものとする．上述の付属語集合 A～E を用いた区切語 w_i の品詞判定手続きを以下に示す．但し w_i は付属語でないと判定されたものとする．

〈手続き I〉

step 1: $w_{i+1}\in A$　ならば　step 7 へ
step 2: $w_{i+1}\in B$　ならば　step 8 へ
step 3: $i\leqq 2$　ならば　step 6 へ
step 4: $w_{i-1}\in D$　ならば　step 7 へ
step 5: $w_{i-1}\in E$　ならば　step 8 へ
step 6: w_i は品詞判定不能　step 9 へ
step 7: w_i は活用しない自立語　step 9 へ
step 8: w_i は活用する自立語　step 9 へ
step 9: RETURN

4.2 付属語以外による品詞判定

付属語以外で出現する単語数が少なく，それらが高い出現頻度を持ち，かつ自立語の品詞情報を与える品詞は，連体詞*** 及び形式名詞**** である．連体詞は単語数が少なく 42 個であり[7]，名詞を修飾し，被修飾名詞はその直後に来ることが多い．形式名詞は意味上から言って名詞の実質を備えていない名詞で，これらが用いられる場合には，これらを限定する語が必ず前に存在しなければならない[8]．その限定語として，〈用言＋付属語〉及び連体詞が挙げられる．すなわち区切語 w_i が形式名詞であれば，w_{i-1} は用言，付属語及び連体詞のいずれかである．形式名詞の単語数も少なく 17 個程度である[8]．またサ行変格，上一段及び下一段に各活用する動詞，形容動詞並びに形容動詞はそれぞれ特有な活用語尾を持つ．そのうち上一段動詞は「イ段の音」に「ル」，「レ」，「ロ」などが接続された語尾を持ち，さらに形容詞は「カロ」，「カツ」，「シイ」，「ケレ」などの活用語尾を，形容動詞は「ダロ」，「ダ」，「デ」，「ナラ」などの活用語尾をそれぞれ持つ．したがってこれらの用言は活用語尾によって品詞判定をすることができる．これらをまとめて自立語の集合Fを次のように定める．

$$F=\left\{\begin{array}{l}\text{サ変動詞，上一段動詞，下一段動詞}\\ \text{形容詞，形容動詞}\end{array}\right\}$$

また日本語文における文末は，用言または〈用言＋付属語〉である場合が多いという事実も品詞判定に利用する．よって連体詞，形式名詞，自立語集合Fなどを利用した区切語 w_i の品詞判定は次に示す手続きとなる．但し w_i は付属語でないと判定されたものとする．

〈手続き II〉

step 1: $i=n$ ならば　step 7 へ
step 2: w_{i-1} が連体詞ならば　step 6 へ
step 3: w_{i+1} が形式名詞であり，w_{i-1} が連体詞でなければ　step 7 へ
step 4: w_1 が F に属する自立語の品詞判定に有効な語尾と同じ語尾を持てば　step 7 へ
step 5: w_i は品詞判定不能　step 8 へ
step 6: w_i は活用しない自立語　step 8 へ
step 7: w_i は活用する自立語　step 8 へ
step 8: RETURN

5. カナ漢字変換

5.1 変換手続き

品詞判定手続き I 及び II を用いたカナ漢字変換を次に示す手続きで行う．

step 1: 日本語文を自立語付属語分かち書き規則によりカナ入力に変換する．
step 2: $i=0$

* 例として「との」，「までの」などが挙げられる．
** 例として「かを」，「などに」，「かをも」などが挙げられる．
*** 連体詞には「この」，「例の」，「大きな」，「大した」等が属する．
**** 形式名詞には「ところ」，「はず」，「由」，「わけ」等が属する．

step 3: $i \leftarrow i+1$

step 4: $i=n+1$ であれば step 15 へ

step 5: w_i が付属語または連体詞，接続詞，代名詞及び形式名詞*のうちカナ漢字変換を必要としない語であれば step 3 へ，カナ漢字変換を必要とする語であれば step 11 へ

step 6: $i=n$ であれば step 13 へ

step 7: 品詞判定手続き I

$\begin{cases} w_i \text{ が活用する自立語であれば} & \text{step 13 へ} \\ w_i \text{ が活用しない自立語であれば} & \text{step 11 へ} \end{cases}$

step 8: 品詞判定手続き II

step 9: w_{i+1} が C_1 に属する付属語であれば step 11 へ

step 10: w_{i+1} が C_2 に属する付属語であれば step 13 へ

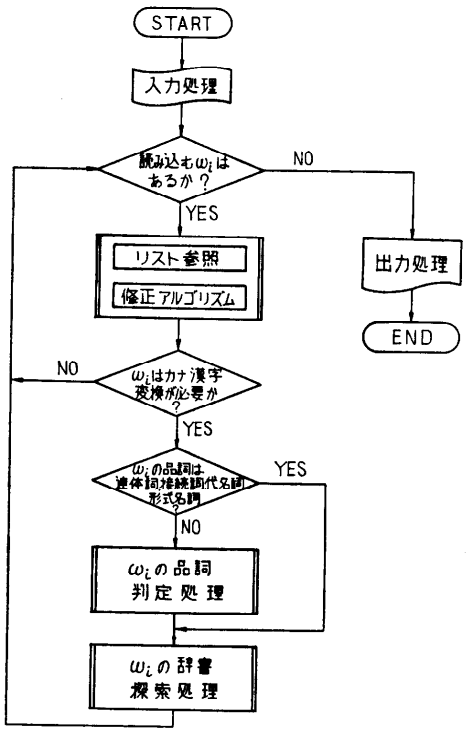

Fig. 3 The flow-chart of the transformation process from *kana* to *kanji*.

step 11: 活用しない自立語のカナ漢字変換処理*

step 12: w_i がカナ漢字変換されず，step 11 に至る前の手続きが step 9 または step 10 であれば step 13 へ，それ以外は step 3 へ

step 13: 活用する自立語のカナ漢字変換処理**

step 14: w_i がカナ漢字変換されず，step 13 へ至る前の手続きが step 10 であれば，step 12 へ，それ以外は step 3 へ

step 15: カナ入力文に対する漢字カナ混じり文の出力後，停止する．

上記の手続きの大まかな流れ図を **Fig. 3** に示す．

5.2 実 験

実験に用いた計算機は FACOM 230-45 S であり，実験のハードウェア構成を **Fig. 4** に示す．またカナ漢字変換プログラムの使用言語は PL/I である．

実験において品詞判定処理に用いた付属語は出現率の高い 120 語であり，それらを **Table 4** に示す．Table 4 に含まれる付属語のみで付属語の全出現率の約 90% をおおう[6]．またここでの A, B, C_1 及び C_2 の類別は **4.1** で述べたものに対応し，類別のパラメー

Table 4 The dependent word list for decision of parts of speech

類別	付 属 語
A	からの，だ，だから，だが，だった，たる，で，であった，であって，であり，である，であるが，であれ，でし，でした，です，ですか，ですが，ですから，では，ではない，でも，という，として，とする，との，とも，など，に，には，にも，の，は，ばかり，へ，への，まで，も，や，を
B	あり，ある，いた，いる，う，けれど，けれども，ざる，し，ず，た，たら，たり，て，ていた，ている，てから，ての，ては，ても，ない，ないか，ないから，ないが，ないで，なかった，ながら，なく，なければ，ぬ，ねば，のか，のが，のだ，のだから，のだった，ので，のであった，のであり，のである，のです，のに，のは，のも，のを，ば，べから，べき，べく，べし，まし，ました，ましょう，ます，ますから，ますが，ません，よう，ようだ，ような，ように，られ，られる，れ，れる，ん
C_1	か，が，から，だけ，であろう，でしょう，と，とか，とは，までの，より，よりも
C_2	かも，だろう

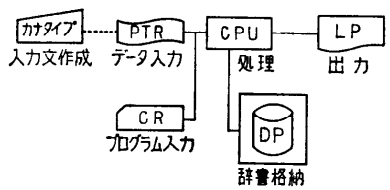

Fig. 4 Hardware configuration.

* 活用しない自立語のうち連体詞，接続詞及び名詞の一部（代名詞と形式名詞）は単語数が少なく，全ての語を書き並べてそれぞれリストを作成し，付属語と同様に処理する．なお接続詞及び代名詞の単語数はそれぞれ 79 及び 120 である[7]．

** 辞書探索における同音異字語の処理は文献 5) を利用して出現率による方法を用いた．すなわち，同音異字語のうち比較的出現率の高いものを処理結果として出力し，また同音異字語の全てが出現率の比較的低いものであれば，どの語も処理結果として出力しないという方法を用いた．

タ $\theta_1, \theta_2, \theta_3$ はそれぞれ 0.85, 0.5, 0.15 を用いたが，頻度による補正を行っている．またカナ漢字変換を必要としない語の除去処理では Table 4 に示した付属語のみならず，それらの付属語の組み合わせによってできる複合付属語をも使用した．ここで用いられた付属語は付属語の全出現率の約 99% をおおっている[6]．

Fig. 5 にカナ漢字変換例を示す．Fig 5 のカナ漢字混じり文出力の中で，下線の付された語は同音異字語が存在していることを表わしているが，本システムでは同音異字語の処理に出現率による方法*を用いているために，漢字変換されない場合が生じる．

5.3 実験結果及び検討

本実験では特別な選択を行っていない雑誌**などから採った日本語文 33 文を対象とし，その処理結果を **Table 5**(次頁参照) に示す．この表を参照して，カナ漢字変換を必要としない語の除去処理，品詞判定処理，辞書探索処理について検討する．

まずカナ漢字変換を必要としない語の除去処理について，その処理の正解率は次の 3 点に関係すると思われる．すなわち，

i) 全区切語に対する，除去されなかった区切語の割合…R_1

ii) 除去されなかった区切語とカナ漢字変換を必要とするにもかかわらず除去された語との和に対する除去されなかった区切語の割合…R_2

iii) 除去されなかった区切語に対し，それらのうちカナ漢字変換を必要とする区切語の割合…R_3

である．R_1 はカナ漢字変換を必要とする語の抽出能力の 1 つの指標と考えられ，この値が真値(変換を必要とする区切語の全区切語に対する真の割合) よりも小さければ，抽出能力が低く，大きければ抽出しすぎるということを表わしているが，正解率のみを考えれば大きい値をとる方が良いといえる．R_2 は抽出のもれの度合を表わし，R_3 は抽出された自立語のなかでの漢字変換率を表わしている．

次に品詞判定処理における評価値として，

iv) 判定された区切語に対する正しく判定された区切語の割合…R_4

を用いる．

さらに辞書探索処理における評価として，

v) 辞書探索された区切語のなかで，カナ漢字変換を必要とする区切語に対する正しく変換された区切語の割合…R_5

vi) 同音異字処理された区切語のなかで，正しく処

(例 1)

日本語文：

「人類はすぐれた目と指先の感覚を持っているが、それは、今ここ一例を示したように、人類の思考法形成の一助をになっていると考える。」

カナ入力文：

「ジンルイ ハ スグレ タ メ ト ユビサキ ノ カンカク ヲ モッ テイルガ、ソレ ハ、イマ ココ イチレイ ヲ シメシ タヨウニ、ジンルイ ノ シコウホウケイセイ ノ イチジョ ヲ ニナッ テイルト イエル。」

カナ漢字変換処理：

(1) カナ漢字変換を必要としない語の除去処理で、除去されなかった区切語

ジンルイ、スグレ、メ、ユビサキ、カンカク、モッ、イマ、ココ、イチレイ、シメシ、ジンルイ、シコウホウケイセイ、イチジョ、ニナッ、イエル

(2) 品詞判定処理

(2・1) 活用しない自立語と判定された区切語

ジンルイ、メ、ユビサキ、カンカク、イチレイ、ジンルイ、シコウホウケイセイ、イチジョ

(2・2) 活用する自立語と判定された区切語

スグレ、モッ、シメシ、ニナッ、イエル

(2・3) 判定できなかった区切語

イマ、ココ

(3) 漢字カナ混り文の出力処理

「人類ハスグレタ目ト指先ノ感覚ヲ持ッテイルガ、ソレハ、今<u>ココ</u>イチレイヲ示シタヨウニ、人類ノ思考法形成ノ<u>イチジョ</u>ヲニナッテイルト考エル。」

(例 2)

日本語文：

「このような状態では、人類的な思考法、ひいては思想などは成立するはずが無い。」

カナ入力文：

「コノ ヨウナ ジョウタイ デハ、ジンルイテキナ シコウホウ、ヒイテハ シソウ ナドハ セイリツスル ハズ ガ ナイ。」

カナ漢字変換処理：

(1) ジョウタイ、ジンルイテキナ、シコウ、ヒイテハ、シソウ、セイリツスル、ハズ、ナイ

(2)

(2・1) ジョウタイ

(2・2) セイリツスル

(2・3) ジンルイテキナ、シコウホウ、ヒイテハ、シソウ

(3) 「コノヨウナ<u>状態</u>デハ、人類的ナ思考法、ヒイテハ<u>思想</u>ナドハ成立スルハズガ無イ。」

Fig. 5 Examples of transformation.

* 前記の脚注参照.

** 例えば，リーダーズダイジェスト.

Table 5 Results of the experiment

入力文数					33
全区切語数					501
I	除去処理	入力語数			275
		抽出	カナ漢字変換候補語	268	
			連体詞・接続詞，形式名等	7	
		除去	誤って除去された自立語	0	
II	品詞判定処理	入力語数（連体詞等は除く）			268
		正しく判定された語数			196
		誤って判定された語数			10
		判定できなかった語数			62
III	辞書探索処理	入力語数			275
		カナ漢字変換	正	同音異字語 有り	42	201
				〃 無し	159	
			誤	〃 有り	5	6
				〃 無し	1	
			不定	〃 有り	11	11
				〃 無し	0	
			カナ漢字変換されず	正（漢字変換不要語）	59	60
				誤（漢字変換必要語）	1	
		辞書収録語数*				1047

* 該当する当用漢字を収録し，同音異字語については文献 9) より採録した．

理された区切語の割合…R_h

が挙げられる．さらに変換システム全体に対しては，処理速度等が考えられるが，入力形式，辞書構成などの違いにより他のシステムとの相対評価は困難である．したがってここでは次に挙げる評価基準しか用いない．

vi) カナ漢字変換を必要とする区切語に対して，正しく変換された区切語の割合…R_t

以上述べてきた評価基準に対する実験値をまとめて **Table 6** に示す．

実験におけるデータ量の少なさは否めないが，R_t で示される漢字変換率が 0.92 であったことで一応の成果が得られたと考える．しかしながらいくつかの問題が未だ残されている．その1つは漢字変換不要語の除去処理において，本実験ではそのような場合が生じなかったが，付属語と同音となる自立語が除去される場合が考えられることである．また1つは品詞判定処理における誤りによって漢字変換できない場合がある．本実験では一例のみであったが，その入力の一部を次に示す．

「…ウツリカワリ ヲ カンサツスル ニ オヨンデ…」

Table 6 System evaluation

		評価基準	理想値	実験値
I	除去処理	R_h	0.43	0.55
		R_d	1.00	1.00
		R_n	1.00	1.00
II	判定処理	R_f	1.00	0.73
III	探索処理	R_d	1.00	0.93
		R_h	1.00	0.72
全システム		R_t	1.00	0.92

ここで区切語「ニ」は付属語集合 A に属し，「カンサツスル」は活用しない自立語とみなされた結果である．さらに1つは同音異字処理の問題である．本実験では，先に述べたように出現率による方法を用い，その変換率は 0.74 であったが，この処理に対する正解率の高いアルゴリズムを考える必要がある．最後の1つは，原文と変換された漢字カナ混じり文との比較において1対1の変換がなされているかという問題である．

例えば，

「今ごく一例を示したように，…」

と言う日本語文が，処理の結果

「今極一例ヲ示シタヨウニ，…」

となった．このことは原文を忠実に再生するという観点から考えると，本システムにかぎらず，カナ漢字変換の持つ本質的な問題点であるかもしれない．

以上，本カナ漢字変換システムについていくつかの問題点を列挙したが，そのいくらかは意味を考慮に入れることにより改善されると考えられる．

6. あとがき

簡単な入力規則すなわち自立語付属語分かち書き規則を用いたカナ漢字変換について述べたが，付属語を手掛り語とする簡単な品詞判定のみで高い漢字変換率が得られることを実験的に示した．

本研究の動機が意味を考慮に入れずにどの程度の漢字変換がなされるかという疑問からであったことを考えると，予想以上の好結果であると思える．しかしながら誤変換の大半（誤りの 89%）が同音異字処理において生じていることから，その処理アルゴリズムを開発する必要がある．

今回なされた実験においては，分かち書きに誤りがないと仮定して行われたものであったが，実際には人間の区切り方に一貫性がなく，比較的多くの誤りが見受けられた．そのためそのような区切りの誤りに対応

できるように部分的な修正を加え，より入力しやすいシステムとなっている．

終りに，本研究を行うにあたり，御鞭撻を頂いた志村正道助教授，豊田順一助教授に深謝します．

参考文献

1) 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)
2) 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（昭 47）
3) 勝部，牧野，木澤：カナ漢字変換の一方法，信学会パターン認識と学習研資料，PRL76-9 (1976)
4) 今泉忠義：国文法の研究，旺文社，東京(昭 48)
5) 国立国語研究所：現代雑誌 90 種の用語用字(1) 総記および語彙集，秀英出版，東京（昭 44）
6) 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（昭 44）
7) 国立国語研究所：電子計算機による新聞の語彙調査(II)，秀英出版，東京（昭 46）
8) 時枝誠記：日本国文法口語編，岩波書店，東京(昭 49)
9) 武田，久松編：角川国語辞典，角川書店，東京(昭 42)

(昭和 51 年 10 月 1 日受付)
(昭和 51 年 10 月 23 日再受付)