多段階分割法によるべた書き日本語文のかな漢字変換†

荒木健治†† 栃内香次†† 永田邦一††

べた書き文のかな漢字変換方式においては，計算機で文を分割する際に一般に極めて多数の候補が出現し，それを一つに決定することが大きな問題となる．本論文は，この問題に対して，他の語と重なりあうことなく確実に分離できる部分（キーワード）から順次，段階的に単語のあてはめを行う手法を提案し，さらに本方式の前提となるキーワードの存在とその有効性を確認し，その上で本方式を用いた実験システムを開発し，その性能評価実験を特定専門分野の学術文献を対象に行った結果について述べたものである．本方式では，最初にキーワードを用いてべた書き文を分割し，ついでそのキーワードによるあてはめの拡張を行い，以後カタカナ語のあてはめ，文節端の助詞候補の検出，連接情報を用いた単語のあてはめ，助詞候補の評価，接辞の処理，最後に一字漢字語のあてはめを行う．実験により500語程度のキーワードで平均3～4文字程度にべた書き文が分割できることがわかり，本方式によるべた書き文のかな漢字変換システムを開発した．さらに，工学に関する3分野の学術文献を資料として，性能評価実験を行った結果，90%以上の変換精度が得られることがわかり，本方式の有効性を示すことができた．

1. はじめに

現在，計算機への日本語文入力法の主流はかな漢字変換方式である．中でも，かな文を分ち書きせずにべた書きのまま入力し，単語への分割もすべて計算機で処理するべた書き入力方式が望ましいとされている．しかし，この方式では文を単語に分割する際に極めて多数の候補が出現し，その中から正しい分割を決定する必要がある．この問題に対し，二文節最長一致法[1]，文節数最小法[2]等種々の方法が提案されているが，構文解析等に複雑な処理を要するという問題がある．

一方，人間が読む場合は，べた書き文であってもそれほど多数の読み方を考えることなく読み下すことができる．このことから，人間の場合は比較的少数の手掛りとなる部分を見いだして最初に分割し，前後のつながりに矛盾がなければ同様な操作を順次続けることによって分割を進めていると考えることができる．

これに基づき，本論文では他の語と重なりあうことなく一意に分離できる部分から，順次，段階的に単語分割を行う手法を提案し，この方式による特定専門分野の学術文献を対象とするべた書き文のかな漢字変換について述べる．

† Multi Stage Segmentation Method for Kana-Kanji Translation of Non-Segmented Japanese Kana Sentences by KENZI ARAKI, KOJI TOCHINAI and KUNIITI NAGATA (Department of Electronic Engineering, Faculty of Engineering, Hokkaido University).

†† 北海道大学工学部電子工学科

2. 多段階分割法

2.1 基本概念

本方式は，べた書き文を最初に少数の手掛り語によって一意に分割し，以後の分割候補を少なくしてからさらに分割を進めるものである．このような手掛り語を以後キーワードと呼び，KW と表す．KW には，一意な分割を実現するために他の部分からの高い分離性を有することと，少数の KW で効率よく分割するために高頻度であること，の二つの条件が必要である．また，KW によって分割された文をさらに分割してゆく際も，同様により確実性の高い部分から段階的に処理すべきであると考えられる．

2.2 予備実験

2.2.1 KW 抽出[3]

前述のような性質をもつ KW が実際の文献にどの程度含まれているかの調査を行った．用いた資料は表1に示す4種で，総文字数は146,336字である．

これらの資料をローマ字表記べた書き文に変換し，以下のアルゴリズムによって KW の抽出を行った．

（1） 資料中の任意の語 W について，W の読みが資料をべた書きで表した文字列中に出現する箇所を検出する．

（2） 検出された箇所にWをあてはめ，誤変換になるかどうかを検査する．ここで，誤変換とは字種が原文と異なっている場合，およびWが複数の漢字語にまたがるようにあてはまる場合である．

表 1 キーワード抽出実験に使用した資料
Table 1 Collected data for the key-word extraction experiment.

No.	著者名	題　名	文字数
1	穂鷹良介	「データベース要論」	63,325
2	中原啓一	「情報検索」	52,186
3	斉藤　康	「ローマ字漢字変換方式による研究者向き日本語処理システム」(北大工学部修士論文)	19,632
4	岡沢好高	「研究者向き日本語処理システムにおける新出語登録方式と性能評価」(北大工学部修士論文)	11,193
総文字数			146,336

表 2 キーワードによる分割実験に用いた資料
Table 2 Collected data for the key-words division experiment.

題　名
昭和54年信学会部門全国大会，443
昭和55年電気四学会北海道支連大，195
第21回情処全国大会，71-9
第22回情処全国大会，31-1
北大大型計算機センターテクニカルレポート No. 3, p. 2 (1980)
昭和48年電気四学会北海道支連大，168
昭和46年電気四学会北海道支連大，148
昭和49年信学会全国大会，1772
第14回情処全国大会，125
第15回情処全国大会，321
北大工学部研究報告，No. 101, p. 71 (1980)
第16回情処全国大会，128
第17回情処全国大会，102
第18回情処全国大会，13
第19回情処全国大会，4C-6
電気学会情報処理研究会資料，IP 77-24 (1977)
情報処理，Vol. 18, p. 135 (1977)
電気学会情報処理研究会資料，IP 78-82 (1978)
札幌医大情報処理学講義テキスト
北大大型計算機センターニュース，Vol. 14, No. 2, p. 23 (1982)

* 総文字数69,099字.　** 著者はいずれも抄内，ほか

(3) 全箇所で誤変換でないとき，語WをKWとする．

べた書き文をKWによって分割する際，上記の操作により得られたKWをすべて使用することは処理効率の点で必ずしも得策ではなく，出現頻度の大きいものを使用する方がよい．そこで，取り出されたKWを出現頻度順に並べ，累積KW占有率を求めた．その結果を**図1**に示す．なお，KW占有率の定義は以下の式による．

KW占有率＝文献中のKWのべ語数/文献の総語数

図1より，KWの頻度上位500語をとると累積KW占有率は37.5%で以後ほぼ一定となっていることがわかる．

2.2.2 KWによるべた書き文の分割[4)]

上述のようにして抽出したKWを用い，**表2**に示す資料についてべた書き文の分割実験を行った．KWは前記抽出実験で得られたものの頻度上位より300語，500語，1,000語，1,500語，および2,000語をとり，各場合についてKW間の文字数の分布を求めた．KW間文字数の平均値を**表3**に示す．KWを500語より増してゆくとKW間文字数はかえって増加する．その理由としては，KWを多くするとKW同士が互いに中途で重複する場合が増すが，このような場合はその箇所の分割を行わないので，かえって分割数

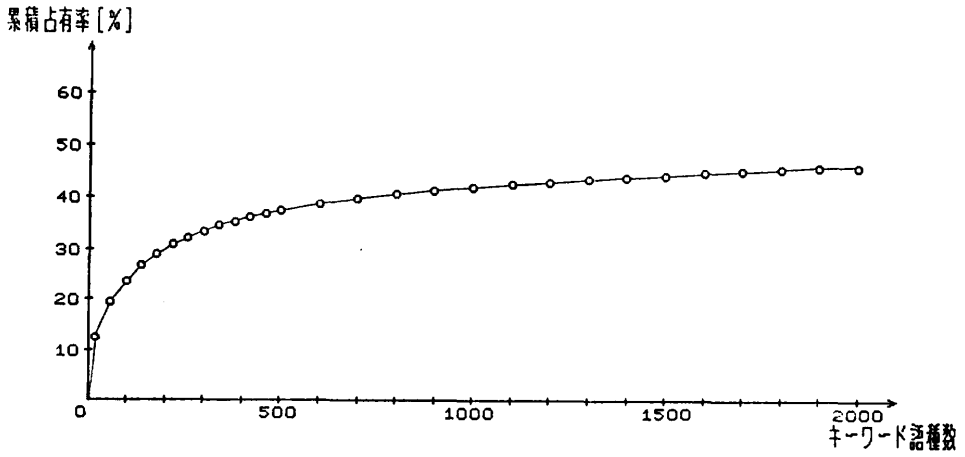


図 1 累積キーワード占有率（キーワード抽出文献における占有率）
Fig. 1 Accumulate rate of the key-words occupation (in the data extracted key-words).

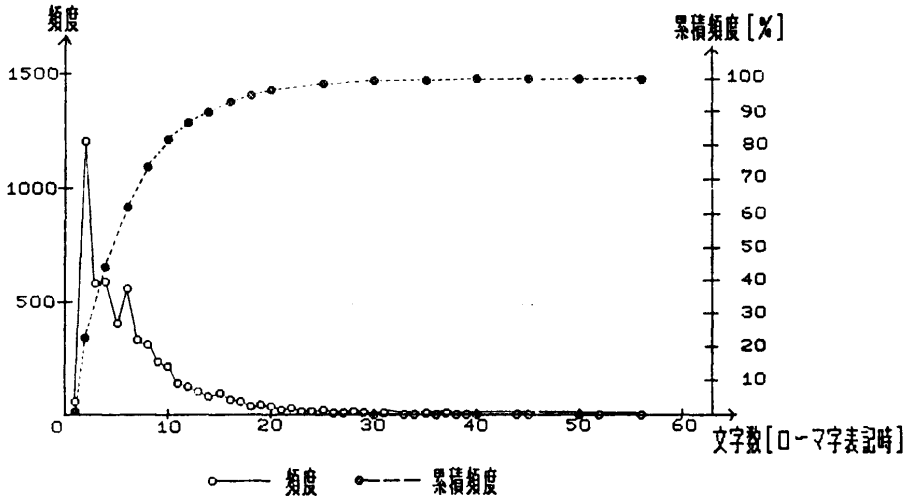


図 2 キーワード間文字数の頻度分布（キーワード語種数 500 語）
Fig. 2 Distribution of the characters between the key-words (Key-word's number: 500 words).

が少なくなるためである．以上により，KW を 500 語とした場合に最も細かく分割されることがわかる．図 2 に KW を 500 語とした場合の KW 間文字数の分布を示す．

以上の結果より，KW は頻度上位の 500 語をとればよいことがわかった．この場合べた書き文字列はかな文字で 3〜4 文字ごとに分割されることになる．

3. 実験システム

3.1 システム構成

前章の結果に基づき，多段階変換方式によるべた書き文かな漢字変換システムを作成した．システムは，北海道大学大型計算機センターの HITAC M-280 H 上に実現され，使用言語は PL/I である．

実験システムの構成を図 3 に示す．システムへのべた書き文の入力はローマ字で行う．これをべた書き処

表 3 キーワード間文字数の平均値
Table 3 Average character number between the key-words.

キーワード語数	キーワード間文字数の平均値
300 語	7.1
500 語	6.8
1,000 語	7.0
1,500 語	7.4
2,000 語	7.6

* 資料はローマ字表記で入力され，したがって文字数はローマ字の字数である．

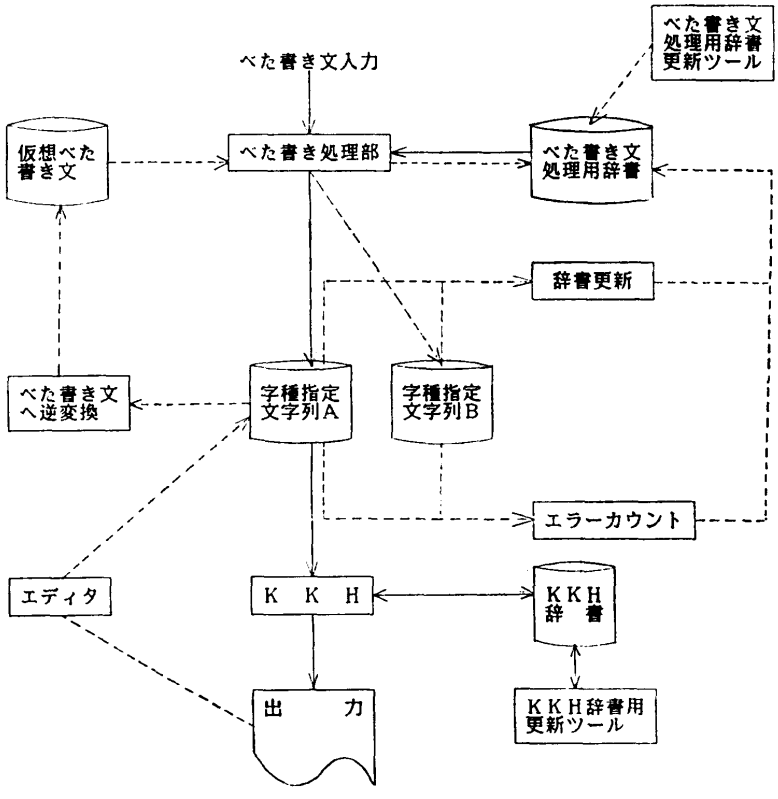


図 3 システム構成
Fig. 3 Construction of system.

理部で単語に分割し，漢字，かな，カタカナを示す情報を付加する．以下，これを字種指定文字列という．字種指定文字列は，既存のかな漢字変換システムKKHの入力様式に合せて作られ[5]，これをKKHに入力して，漢字かな混じり文を得る．さらに，辞書には語の出現頻度，履歴および変換誤りなどの情報が付加され，これをもとに辞書を更新し，使用につれて変換精度が向上するようにしている．図3で，実線は，べた書き処理部を示し，破線は辞書更新部を示す．

3.2 処理手順

3.2.1 変換辞書の構造

表4に辞書の一覧を示す．辞書中の各語にはそれぞれ頻度，履歴およびエラーが記録されている．ここで，頻度はべた書き処理部で参照された回数を，履歴は最後に参照されてからの期間を，また，エラーは変換誤りを引き起こした回数を表す．具体的には以下のような処理を行う．

1) べた書き処理部で参照された語
　頻度←頻度＋1　　履歴←0
2) 一つの文書の処理が終了したとき，一度も参照されなかった語
　履歴←履歴＋1
3) 変換結果と校正後の文書と比較して異なる語
　エラー←エラー＋1

辞書中の語は次式に示されるCの値の小さい順に配列されている．

$$C=\alpha E+\beta R-H \qquad (1)$$

ここで，E：誤変換度数，R：履歴，H：頻度，α，β：係数である．

表4　変換辞書の種類
Table 4　Table of the translation dictionaries.

No.	辞書名	連接情報*の有無	機能
1	キーワード辞書	無	べた書き文を字種指定文字列に変換する際に用いる．
2	カタカナ語辞書	無	
3	助詞辞書	無	
4	単語辞書	有	
5	接辞辞書	無	
6	一字漢字語辞書	有	
7	単語補助辞書	有	辞書更新の際，頻度，履歴の情報により4,6から削除された連接情報を蓄える．
8	一字漢字語補助辞書	有	

* 連接情報とは，単語Wの前後の文字（または記号，数字，空白等を含む）a，bの組をいう．

また，連接情報を持つ辞書には連接情報ごとに頻度と履歴が登録されている．ここで，連接情報とは，単語Wの前後の文字（記号，数字，空白等を含む）a，bの組をいう．

3.2.2 変換アルゴリズム

図4にべた書き文の変換処理過程を示す．

a．KWによる分割

入力されたべた書き文に対し，最初にKWによる分割を行う．**図5**に分割の例を示す．ここで，**図6**の例のように複数個のKWが重なってあてはまることがある．このような重なりは定義から明らかなようにKWを抽出した文献では出現しないが，他の文献では出現する可能性がある．このような場合は，以下の処理を行う．

（1） 一方が他方を含む時（図6の1）は，文字数の多い方（この例では「形式」）を採用．
（2） 中途で重複する時（図6の2）は，分割を行わない．

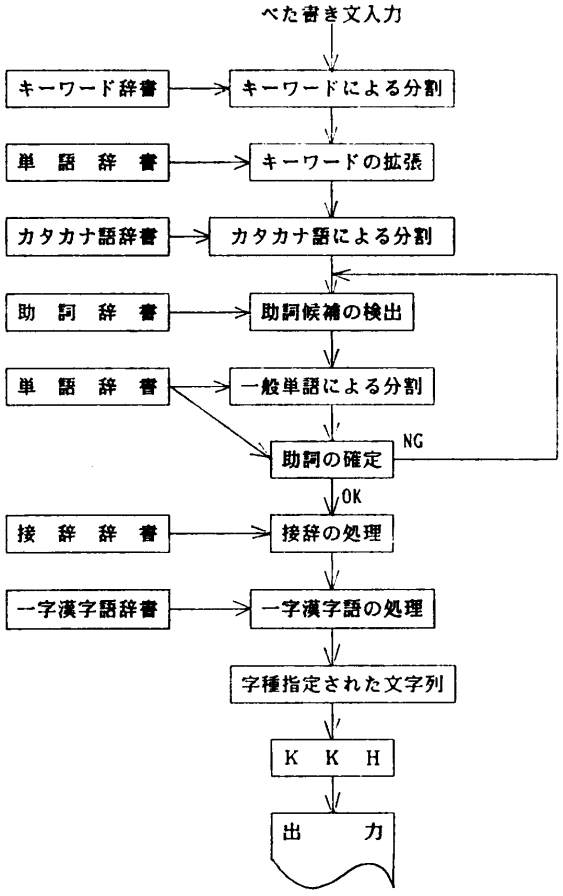


図4　処理の流れ
Fig. 4　Process.

1. [けいさん] き [しすてむ] の [ないぶ] の [じょうたい] の [きろく] [を] くわ [しく] [ちょうさ], ぶんせき [する] こと [によって] [せいのう] ひょうか [を] おこなおうと [する] もの [である].
2. [ていそく] にゅう [しゅつりょく] [そうち] とふぁいる [そうち] かんの [しすてむ] にゅう [しゅつりょく] は [こうりょ] [しない].
3. [この] てんにも [ちゅうい] [を] はらい, [つぎ] [のような] [かてい] [について] のべる.

[ ]: キーワード

図 5 キーワードによる分割の例
Fig. 5 Examples of the division by key-words.

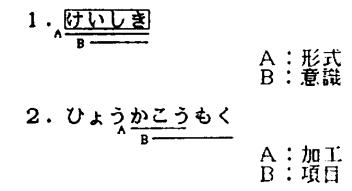


図 6 キーワードによる分割方法
Fig. 6 Key-word duplication.

b．KW の拡張

あてはまった KW のうちの漢字語について，その前後に存在する文字列を付加した形の語が単語辞書に存在する場合，KW の代りにその語をあてはめる．これを KW の拡張という．**図 7** に例を示す．

c．カタカナ語による分割

KW で分割されたべた書き文について，次にカタカナ語による分割を行う．学術文献では一般にカタカナで表記される外来語が頻出するが，これらには他の部分と重なることなく，一意に分割可能なものが多い．そこで，これらをカタカナ語辞書に蓄積しておき，これを用いて分割を行う．分割方法は，KW の場合と同様である．

d．文節端助詞候補の検出

以後の処理を行う前に，文節端の助詞である確率が 60% 以上に達する[6]「の，は，に，が」を検出する．これらを文節端助詞候補と称する．この処理により分割候補が少なくなり以後の選択が容易になる．なお，「を」は通常の文では必ず文節端助詞で他から独立している．それゆえこれは KW となり，すでにあてはめられている．

e．一般単語による分割

単語辞書を用いて一般の単語による分割を以下の順序で行う．

（1） 連接情報の頻度の大きいもの．

（2） （1）が同じ時……式（1）に示される C の値が小さいもの．

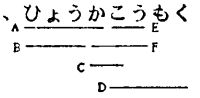
（3） （2）も同じ時……文字数の多いもの．

（4） （3）も同じ時……前方よりあてはまるもの．

一般の単語については，種々の分割の可能性がある．そこで，単語と前後の文字との連接情報を用いて一意に決定している．連接情報による単語の分割を**図 8** を例にして説明する．この文字列から得られる候補は，A～F までの 6 通りである．このなかで，例えば E の「項」の前後は「か」と「も」であるので，文字列中より得られるこの語の連接情報は「か，も」となる．一方，単語辞書の「項」の欄の連接情報を見ると「か，も」が頻度 1 と登録されているので，この連接情報の頻度は 1 であることがわかる．同様のことを A，B，C，D，F について行い図中の表のような頻度を得る．初めに，連接情報の頻度の一番高い A の「評価」を採用する．この結果，E と D が残る．このうち連接情報の頻度の高い D が採用され，この文は「，評

○［計算］き → 〈計算機〉
○にゅう［出力］ → 〈入出力〉
［ ］: キーワード
〈 〉: 拡張キーワード

図 7 キーワードの拡張の例
Fig. 7 Examples of the key-words extension.



・文字列中より得られる連接情報

A： P 評価 こ
B： P 表 か
C： う 過去 う
D： か 項目 が
E： か 項 も
F： う 加工 も

・辞書から得られる連接情報とその頻度

		連接情報	頻度
◎	A評 価	P こ	6
	B 表	P か	1
	C過 去	う う	0
◎	D項 目	か が	3
	E 項	か も	1
	F加 工	う も	2

P：句読点

図 8 連接情報による単語の分割
Fig. 8 Word division using the triplet of the word.

1. 文節端助詞候補の前後に未変換文字列がない場合
   [システム](の)[内部]　　→「の」は助詞
2. 文節端助詞候補の前後に未変換文字列がある場合
   (a) 文節端助詞候補のみを解除してあてはめをしなおす.
   [なる]も(の)と[する].
   [なる]|もの|と[する].　　→「の」は助詞でない.
   (b) 単語のあてはめも解除してあてはめをしなおす.
   [時間]よりも(は)|約|できた.
   [時間]よりもは|約|できた.
   [時間]よりも|早く|できた.　→「は」は助詞でない.
   []: キーワード　(): 助詞　||: 単語

図 9 助詞の確定方法

Fig. 9 Method of the end-of-paragraph particles determination.

価/項目/が」と分割される.

また，記号，句読点，空白，KW，カタカナ語，助詞候補およびすでに分割済みの単語に完全に挟まれている語については連接情報に無関係に分割を行っている．これは，前後の分割が決定しているので，その間に挟まれた部分と同じ読みの語が単語辞書に存在すればそのまま確定してもよいと考えられるからである.

f．助詞の確定

一般単語による分割が終了した時点で，d．で仮に定めた文節端助詞候補が真に文節端の助詞であるか否か再検討を行い，助詞でないと判定された場合は分割をやり直す[7]．この例を**図 9** に示す.

g．接辞の処理

未変換部分に接辞の処理を行う．これは漢字，カタカナ，および数字の前後の文字に，あらかじめ接辞辞書に登録された接辞をあてはめるものである．接辞辞書には，**図 10** に示す 37 語が登録されている．なお，漢字一字でかつ読みも一字の接辞(「化(か)」,「非(ひ)」,「未(み)」など)は一字漢字語辞書に登録する.

h．一字漢字語の処理

最後に一字漢字語の処理を行う．一字漢字語とは，漢字一字でかつ読みも一字の語および「謝(しゃ)」,「感(かん)」,「発(はっ)」など，そのうしろに拗音，促音，撥音を伴う語をいう．一字漢字語は多数の語と複合語を作り，他の部分からの独立性が最も低いので，連接情報の一致したもののみあてはめる.

i．同音語の処理

同音語の選択については，3.2.2 の e. で述べた連接情報を用いる方法とほぼ同じ手法を用い，単語Wとその前後の文字（または記号，数字，空白等を含む）a，bとの三つ組 aWb の出現頻度により同音語を選択している．なお，本システムにおいては字種をべた書き処理部で確定した後，同音語についてのみこの処理を行っている．例えば，入力文字列が「～をしようとする」で「しよう」という読みが，べた書き処理部により漢字と確定した時，「仕様」と「使用」という同音語が存在する．ここで，二つの語を比較した場合「～を使用する」とは言うが，「～を仕様する」とは言わないので，「を，す」が前後に出現するのは，「使用」の方である．このような情報があらかじめ辞書に蓄えられており，「しよう」という読みは，「使用」であることがわかる．このような手法は，先に KKH において試みその有効性を確認しているので，本システムではそれを利用し，字種決定後，あらためて同音語選択のみを別フェーズで行うこととしている[8].

図 11 に変換結果を示す．図中の記号はどの段階であてはめられたかを示すものである．最後まで未変換の部分はかなで出力されるが，そのほとんどは原文でもかなで表される部分である.

かく	しょう	たい	ほん	だい	ない	きゅう	さい
ちょう	てい	どう	ふく	もと	りょう	かい	じょう
しき	せい	さく	てき	ほう	もう	よう	りょく
がた	つい	ぎゃく	そう	しつ	ちゅう	ろん	けい
こう	じゅん	かん	けん	とう			

図 10 接辞辞書の見出し語

Fig. 10 Catchwords of the affix dictionary.

1. けいさんきしすてむのないぶじょうたいのきろくをくわしくちょうさ，ぶんせきすることによってせいのうひょうかをおこなおうとするものである.
2. ていそくにゅうしゅつりょくそうちとふぁいるそうちかんのしすてむにゅうしゅつりょくはこうりょしない.
3. このてんにもちゅういをはらい，つぎのようなかていについてのべる.

(a) 入力べた書き文

1. 〈計算機〉[システム](の)[内部](の)[状態](の)[記録][を]|詳|[しく][調査],|分析|[する]|こと|[によって][性能]|評価|[を]|行|おうと[する]|もの|[である].
2. [低速]〈入出力〉[装置]と{ファイル}[装置]##間##(の)[システム]〈入出力〉(は)[考慮][しない].
3. [この]/点/(に)も[注意][を]|払|い,[次][のような][過程][について]|述べ|る.

[]: キーワード　〈〉: 拡張キーワード　{}: カタカナ語
(): 助詞　||: 単語　##: 接辞
//: 一字漢字語

(b) 出力結果

図 11 変換結果

Fig. 11 Result of the translation.

表 5 変換辞書の更新手順
Table 5 Flow of the renewing translation dictionaries.

手順	内容
1. 仮想べた書き文へ変換	出力結果を検査して誤りその他の校正処理を行いそれに基づいて字種指定文字列Aを訂正した後，再度べた書き文に変換する．これを仮想べた書き文という．
2. 頻度，履歴の更新	仮想べた書き文をべた書き文処理部で変換し，字種指定文字列Bを得る．この際，3.2.1で述べたように頻度と履歴を更新する．
3. 誤変換数の更新	AとBを比較して，誤変換数をカウントし各辞書に登録する．
4. 新語登録	Aに存在する語で，すべての辞書にないものを単語辞書または一字漢字語辞書に登録する．
5. 連接情報の更新	単語辞書中の語について，Aより単語の前後の文字情報を取り出し，単語辞書，単語補助辞書に連接情報を登録する．一字漢字語辞書，一字漢字語補助辞書についても同様である．
6. キーワード辞書，カタカナ語辞書の更新	(1) Aより2.2.1に示した条件を満たすものをキーワード候補として取り出す． (2) キーワード候補とキーワード辞書中の語を3.2.1で示したCの値の小さい順に並べ，上位500語をキーワード辞書に登録する． (3) カタカナ語のキーワード候補中頻度が小さくキーワード辞書に登録されなかったものをカタカナ語辞書に登録する．

3.3 変換辞書の更新

上述の処理によって得られた結果には，一般にいくつかの変換誤りが存在する．この情報によって変換辞書を更新し，使用につれて辞書が適応するようにしている．変換辞書の更新手順を**表5**に示す．なお，前述のように図3の破線部分がこの手順を示している．

なお，PL/I のソース・プログラムは約 3,000 行であり，処理時間は 400 文字当り（原稿用紙1枚相当）で5秒661ミリ秒，CPU 時間は3.1秒であった（なお，使用計算機は HITAC M-280 H である）．

4. 変換実験結果ならびに考察

4.1 実験方法

情報工学，機械工学，応用化学に関する論文，計34編（総文字数244,985文字）を用い，システムの性能評価実験を行った．**表6**に資料の一覧を，また，実験方法を**図12**に示す．なお，式(1)のパラメータ α, β は表6の文献 No. 1～14 を用いた実験で良好な変換率が得られた $\alpha=10$, $\beta=1$ とした[9]．

4.2 実験結果

図13に正変換率，未変換率，誤変換率の推移を示す．図からわかるように，正変換率は情報，機械，化学と分野が変わるごとに一時的に若干低下するがすぐ回復し，平均 90% 程度に達している．これは辞書の内容が一時的に分野に適応しなくなるためであるが，この結果からわかるように短期間で回復する．未変換率も同様に分野が変わる度に一時的に増加するが直ちに低下する．また，誤変換率はほぼ一定である．したがって，分野が変わる際に正変換率が低下するのは未

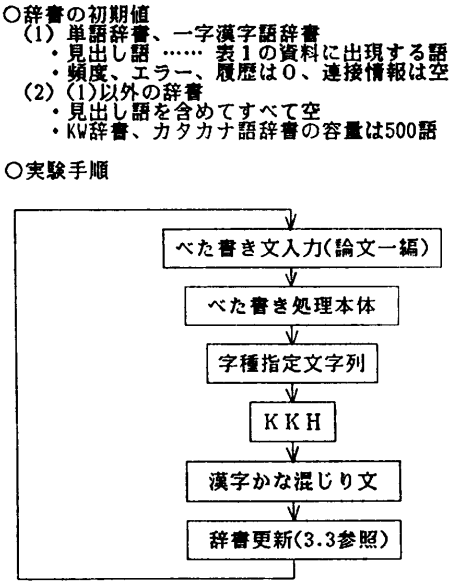


図12 実験方法
Fig. 12 Method of the experiment.

変換が増大するためであることがわかる．

各分野で辞書が安定したと考えられる文献 No. 6～14, 18～24, 29～34 の変換結果を**表7**に示す．

4.3 考察

表7より，本システムの変換アルゴリズムの主要な問題点として次の3点があげられる．

(1) KW の拡張と一般単語による分割で誤変換率が高い．
(2) 未変換が誤変換の2倍以上存在する．
(3) KW の誤変換率が 1.3% である．

これらについて，以下に示す対策が考えられる．

表 6 変換実験に用いた資料
Table 6 Collected data for the translation experiment.

分野	No.	論文名	著者名	巻名	文字数
分野（情報）出典（情報処理学会論文誌）	1	高集積マイクロコンピュータに適したマイクロプログラム制御方式	前島ほか	Vol. 23 No. 1	10,594
	2	COBOL マシンとその設計思想—ハードウェア構成について—	山本ほか	Vol. 23 No. 1	8,451
	3	フーリエ変換を用いたテクスチュアの構造解析	松山ほか	Vol. 23 No. 2	7,723
	4	日本語文入力用カタカナ語検出規則とオンライン国語辞典の一分析	木村	Vol. 23 No. 2	8,468
	5	インテリジェント・コンソール—OS の機能拡張の一方法—	有田	Vol. 23 No. 3	9,343
	6	ポータブル画像処理ソフトウェア・パッケージ SPIDER の開発	田村ほか	Vol. 23 No. 3	9,624
	7	グラフィック・ディスプレイ・ターミナルのための端末作画システム	高藤ほか	Vol. 23 No. 4	6,423
	8	オペレーティング・システムのファームウェア化対象選定法	長岡ほか	Vol. 23 No. 4	7,594
	9	プログラム階層構造の生成，処理，文書化能力を有するテキスト・エディタ	酒井ほか	Vol. 23 No. 5	7,916
	10	パステストに本質的な分岐に着目した網ら率尺度の提案	中所ほか	Vol. 23 No. 5	10,970
	11	計算機システムにおける性能管理の一方式とそれを用いた実験	吉住ほか	Vol. 23 No. 6	9,123
	12	高速パケット伝送路用前置処理装置の一構成法	寺田ほか	Vol. 23 No. 6	8,401
	13	ソフトウェア生産過程の評価実験に関する考察	有澤ほか	Vol. 23 No. 3	6,453
	14	文節数最小法を用いたべた書き日本語文の形態素解析	吉村ほか	Vol. 24 No. 1	9,658
		情報工学分野の文字数の合計			120,741
出典分野（北海道大学工学部研究報告（機械））	15	格子乱流中における垂直平板の流力特性	有江ほか	第 106 号	4,905
	16	疲れ強さにおよぼす加工硬化および残留応力の影響—低温焼鈍の場合—	秦ほか	第 97 号	8,890
	17	暖房用ストーブの燃焼性能に関する研究（第 1 報）—ポット式灯油ストーブの燃焼実験及び流動解析—	園田ほか	第 106 号	7,227
	18	暖房用ストーブの燃焼性能に関する研究（第 2 報）—温風，FF およびポータブル型各種ストーブの燃焼実験—	園田ほか	第 106 号	5,369
	19	任意に調節可能な座標線密度をもつ流れ場内格子点網の創成法	飯田ほか	第 107 号	7,084
	20	円弧切欠きと荷重端の干渉について（第 3 報　平面弾塑性応力問題）	岸田ほか	第 108 号	2,279
	21	境界層剝離の近似的な推定法	知名ほか	第 98 号	5,108
	22	換気回数の低い室内における都市ガスの燃焼と一酸化炭素の発生	伊藤ほか	第 105 号	7,931
	23	チェーンソーの振動におよぼす切削条件の影響	金内ほか	第 104 号	7,149
	24	任意形状を有する四辺形膜の自由振動	入江ほか	第 96 号	4,970
		機械工学分野の文字数の合計			60,912
出典分野（北海道大学工学部研究報告（応用化学）	25	ニオブ酸カリの焼結に関する研究	小平ほか	第 112 号	3,451
	26	金属酸化物と銅からなる混練触媒のキャラクタリゼーション及びメタノールリフォーミング反応	小林ほか	第 102 号	5,621
	27	混合粒子から成る気固系流動層の粒子混合と分級	千葉ほか	第 102 号	7,225
	28	有機アルミニウム化合物存在下での環化重合	横田ほか	第 102 号	5,240
	29	高分子のメカノケミストリー（I）（綜報）—メカノラジカルと結晶構造変化—	相馬	第 102 号	10,296
	30	高分子のメカノケミストリー（II）（綜報）—メカノケミカル反応—	相馬	第 102 号	10,968
	31	混合粒子系噴流層の流動特性	上牧ほか	第 109 号	4,640
	32	炭酸ガスの有効利用に関する研究—炭酸ガスと二硫化炭素を原料とする硫化カルボニルおよび一酸化炭素の合成—	杉岡ほか	第 93 号	5,160
	33	3-ジシクロヘキシルポリルアクリル酸エチルのアルコール中での光化学反応	徳田ほか	第 97 号	5,847
	34	2-Alkylbenzimidazol の改良合成法—*p*-Toluolsulfonsäure を媒体とする脂肪酸と *o*-Phenylendiamin との縮合—	高田ほか	第 97 号	4,884
		応用化学分野の文字数の合計			63,332
		総文字数			244,985

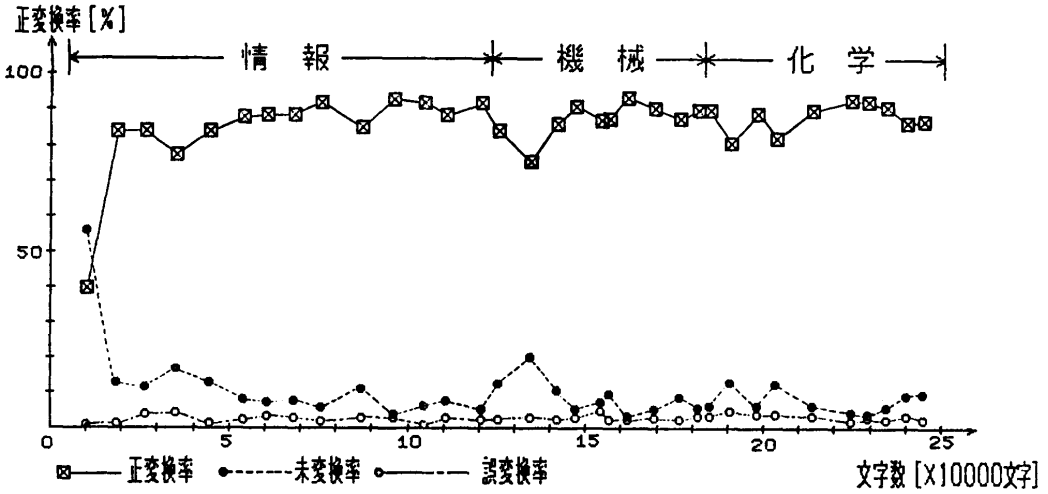


図 13 変換実験結果
Fig. 13 Result of the experiment.

表 7 あてはめ段階別エラー率
Table 7 Rate of the error in each stage.

段　階	あてはめ数	誤変換数	誤変換率[%]
キーワード	26,831	340	1.3
キーワードの拡張	3,666	282	7.7
カタカナ	959	16	1.7
助　詞	10,248	422	4.1
単　語	10,689	672	6.3
接　辞	128	3	2.3
一字漢字語	654	22	3.4
小　計	53,175	1,757	3.3
未変換	3,966	3,966	—
合　計	57,141	5,723	10.0

* 表中の数値のうちあてはめ数，誤変換数は，辞書が安定したと考えられる文献 No. 6～14, 18～24, 29～34 の 22 編の合計で，誤変換率は，誤変換数/あてはめ数である．
平均値
未変換率＝(未変換数/総語数)×100＝6.9 [%]
誤変換率＝(総誤変換数/総語数)×100＝3.1 [%]
正変換率＝{(総あてはめ数－総誤変換数)/総語数}×100＝90.0 [%]
* 総語数＝未変換数＋総あてはめ数

(1) KW の拡張については，読みが一致した部分を拡張するので，誤変換率は辞書にどのような語があるかによって大きく左右される．そこで，この段階にも連接情報を用い，誤変換を少なくすることが考えられる．

(2) 一般単語による分割については，連接情報によってもなお一意に決定できないものが存在するので，各単語について KW との関連性を示す情報を付加しておき，その情報をもとにあてはめる方法が考えられる．

(3) 未変換については，辞書の不備（未登録語，連接情報の未登録など）によるところが多い．今回の実験では単語辞書の収録語数は約 6,000 語であり，一方研究テーマは論文 1 編ごとに変わるのでこのような結果になったと考えられる．

(4) KW 段階での誤変換については，KW は一度あてはまると固定され，誤変換が回復することがない上に，最初に決定するので以後の分割に影響する．そこで，KW を誤変換の比較的少ないものと多いものの二つに分類し，誤変換の比較的多いものについては，抽出文献によりいくつかのグループに分類しておき，同じグループの KW が同時に複数あてはまる時のみ分割するという条件を付け加えることが考えられる[10]．

5. おわりに

べた書き文のかな漢字変換では，単語への分割の際に多数の候補が出現し，その中から正しい分割を一意に決定することが大きな問題である．本論文では，比較的少数の手掛り語（KW）を用いてあらかじめべた書き文を分割し，以下確実性の高い部分から段階的に処理する方法を提案した．さらに，KW の存在を確認するために実験を行い，500 語程度の KW でかな文字で 3～4 文字程度にべた書き文を分割できることを見いだした．この結果から，多段階分割法によるべた書き文のかな漢字変換システムを開発し，性能評価実験を行った．その結果 KW による分割から段階的

に分割を行ってゆくアルゴリズムにより 90% 以上の変換精度が得られることがわかった．これは本方式の有効性を十分示していると考えられる．

今後さらに，未登録語の問題，KW との関連性の考慮，KW の階層化，辞書更新法の改良等について，検討を進める予定である．

謝辞　本研究に際し，種々御討論いただき，適切な御示唆をいただいた本学部電子機器工学講座各位に感謝します．なお，本研究の一部は科学研究費補助金（昭和 60 年度一般研究(C)第 60580015 号）の補助により行われた．

参 考 文 献

1) 牧野　寛，木澤　誠：べた書き文の分かち書きと仮名漢字変換，情報処理学会論文誌，Vol. 20, No. 4, pp. 337-345 (1979).
2) 吉村賢治，日高　達，吉田　将：文節数最小法を用いた日本語文の形態素解析，情報処理学会論文誌，Vol. 24, No. 1, pp. 40-46 (1983).
3) 荒木，鈴木，伊藤，栃内，永田：べた書き文入力におけるキーワード抽出，電気四学会北海道支部大会講演論文集，p. 270 (1983).
4) 荒木，鈴木，伊藤，栃内，永田：キーワードによるべた書き文の分割，電子通信学会総合全国大会講演論文集，7-49 (1984).
5) 栃内，伊藤，荒木，鈴木，永田：研究者向き日本語ワードプロセッサ KKH II の開発，北海道大学工学部研究報告，第 19 号，pp. 119-126 (1984).
6) 館林，中馬，杉村，小林，向井，滝口：自由文入力・仮名漢字変換方式，第 26 回情報処理学会全国大会講演論文集，pp. 1165-1166 (1983).
7) 荒木：キーワード方式によるべた書き日本語文のかな漢字変換，北海道大学工学部修士論文 (1985).
8) 栃内，伊藤，鈴木：前後連接文字を利用した同音語選択機能を有するかな漢字変換システム，情報処理学会論文誌，Vol. 27, No. 3, pp. 313-320 (1986).
9) 山田：キーワード方式べた書き文かな漢字変換におけるキーワード抽出と更新，北海道大学工学部卒業論文 (1985).
10) 荒木，内田，山田，栃内，永田：キーワード方式べた書き文かな漢字変換システムの変換性能の向上，電子通信学会情報・システム部門全国大会講演論文集，3-169 (1985).

（昭和 61 年 8 月 14 日受付）
（昭和 62 年 1 月 14 日採録）

荒木　健治（正会員）

昭和 34 年生．昭和 57 年北海道大学工学部電子工学科卒業．同大学院工学研究科修士課程を経て，現在同大学院博士後期課程在学中．語の階層化による日本語情報処理の研究に従事．電子情報通信学会会員．

栃内　香次（正会員）

昭和 14 年生．昭和 37 年北海道大学工学部電気工学科卒業．昭和 39 年同大学院工学研究科修士課程修了．現在同工学部電子工学科助教授．計算機応用，ことに日本語文書処理に興味をもつ．電子情報通信学会，日本音響学会各会員．

永田　邦一（正会員）

大正 15 年生．昭和 22 年東京帝国大学第二工学部電気工学科卒業．昭和 22 年日本電気(株)入社．昭和 37 年工学博士．昭和 48 年北海道大学教授，現在に至る．その間電気音響変換器・通信方式・音声情報処理の研究に従事．著書「有線通信工学」など，電子情報通信学会，IEEE，日本音響学会など各会員．