適応型変換辞書を用いるかな漢字変換†

栃内香次†† 斉藤 康††

本論文は，研究報文の作成など学術文書処理において有用なかな漢字変換手法について述べたものである．学術文書では，事務文書その他と異なる特有の用語が多数使われる．このうち相当部分はいわゆる学術用語で，これはさらに専門分野ごとに大幅に異なっている．本論文では，専門分野を限定し，そこで頻用される語を抽出して構成される小容量の変換辞書を用いてかな漢字変換を行う手法を提案し，あわせてそれによる研究者個人用変換システムの構成について報告する．この変換システムでは，辞書は動的に構成され，使用を重ねるにつれて内容が変化し，使用者に適応するようになっている．辞書に収容する語の初期値としては，情報処理学会誌からとった30篇の論文において出現頻度が2以上の2,293語を用いている．また，主たる使用者が研究者であることを考慮し，入力は通常のTSS端末からローマ字で行い，大文字，小文字の使い分けにより漢字とかなの区別を行っている．このシステムを使用し，同一著者の文献を用いて入力実験を行った．その結果，約7,700漢字語の入力により変換率が93%から95%に増加し，変換辞書の適応により変換性能が向上することが認められ，本方式の有効性がたしかめられた．

1. はじめに

一つの研究を遂行する過程では，1)論文，報告書，2)学会などの講演予稿，3)会議，打合せ資料，等のさまざまの日本語文書が作成されるが，その作業は最終原稿に至るまで研究者自身の手によらなければならない場合が多い．かな漢字変換方式による日本語ワードプロセッサの利用は，この作業を支援する有力な手段である[1]．

上記諸文書を「学術文書」とよび，通常のオフィス活動において作成される事務文書と比較すると，

1) いわゆる学術用語が多数使われ，
2) 他にも学術文書特有の用語が多く，
3) 氏名，地名などの固有名詞はあまり多くない，

等の特徴をもつ．もちろん，専門分野ごとに，さらに各著者ごとによく使われる用語には相違がある．

われわれは，これらを考慮したかな漢字変換方式の実験を行っている．この方式の特徴は，

1) 専門分野に合わせて抽出された少数の語を収録し，
2) かつ収録語が個々の使用者に適応してゆくように構成された，

動的な変換辞書を用いる点にある．

以下，本論文では変換辞書の構成および試作した実験システムの概要について述べる．

2. 動的変換辞書方式

語単位のかな漢字変換方式において，一般に変換辞書は数万語が必要とされる[2]．しかし，対象をある分野の学術文書に限定すると，よく使われる語は少数であり，研究者個人にまで限定すればさらに減少すると予想される．そこで，以下に示すような変換辞書の構成が考えられる．

1) 対象とする専門分野の文献でよく使われる語を抽出し，原辞書を作る．
2) 各使用者は個別に変換辞書をもつ．この辞書は原辞書を初期値とし，使用につれて使用者に適応してゆくように構成される．

これにより，以下の諸点が特別な工夫をすることなく自然に実現される．

1) 変換辞書は個々の使用者に適応し，しかも必要な記憶容量は小さい．
2) その専門分野で使われない語は辞書に収録されないので，同音語の発生が少ない．
3) 同様の理由で，誤入力された語のうち入力時点でただちに検出されるものの割合が大きく，後から修正しなければならないものが減少する[3]．
4) 逆に，一般には使われない語でも特定の分野，個人に多用される語は辞書に収録される．
5) 研究対象の変化などによってよく使われる語が変動すると，辞書の内容もそれに追随する．

一方，問題点としては，

† Kana-Kanji Translation Using Adaptive Kanji-Word Dictionary by KOJI TOCHINAI and YASUSI SAITO (Department of Electronic Engineering, Faculty of Engineering, Hokkaido University).

†† 北海道大学工学部電子工学科

1) 収録語が限定されるので，未登録語が出現する可能性がいつでもあり，
2) 同音語がなく，1語だけ収録されている語の同音語があらたに出現すると誤変換される，

という2点がある．そこで，これらに関して，

1) 未登録語の出現頻度が十分小さく，
2) 上記2)に起因する誤変換が十分少ない，

ことが必要である．

3. 学術文献にあらわれる漢字語

ある専門分野でよく使われる漢字の語（以下，漢字語という）は限られるというわれわれの予想を確認するために，現実の文献について調査を行った．

3.1 資　　料

情報処理学会誌，同論文誌に掲載された論文，解説等から30篇をとり，その本文から漢字語を抽出して資料とした．そののべ語数，異なり語数などを**表1**に示す．

漢字が連続している場合は2～3字を基準として複数の語に分割している．なお，1字の語を少なくして同音語を減少させるために[4]，接辞もなるべく分離せずに語のなかに含め，さらに漢字1字の語では送りがなを1字付加した形を1語としたものがある．

3.2 新出語および同音語

漢字語を累積してゆくと，はじめはそれまで未登録の語（以下，新出語という）が頻繁にあらわれるが，累積が進むにつれて減少する．**図1**はこれを示すもので，横軸は累積ページ数，縦軸は資料1篇ごとの新出語出現率である．

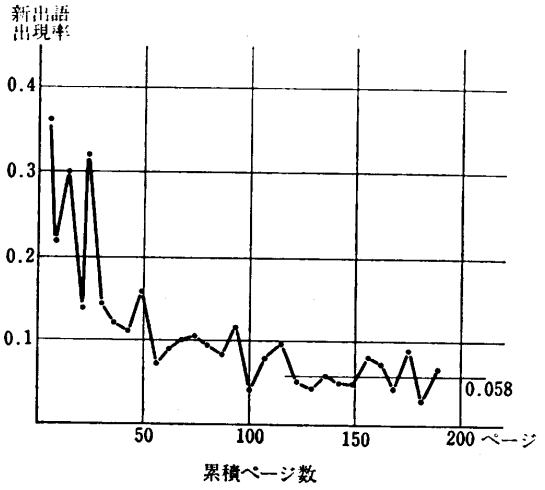


図1 新出語出現率の推移
Fig. 1 Change in the rate of first appeared words.

表1 漢字語データ
Table 1 Collected Kanji-word data.

資料数	30 篇*
総ページ数	189 ページ**
漢字語のべ語数	34,906 語
〃 （ページ当り）	平均 184.7 語/ページ
	(111.6～307.9 語/ページ)
漢字語異なり語数	3,671 語
〃 （頻度2以上）	2,293 語

* 資料内訳：論文26篇，解説2篇，講演1篇，談話室1篇
** ページ数分布：8ページのもの2篇，7ページのもの13篇，6ページのもの12篇，5ページのもの1篇，3ページのもの1篇，2ページのもの1篇

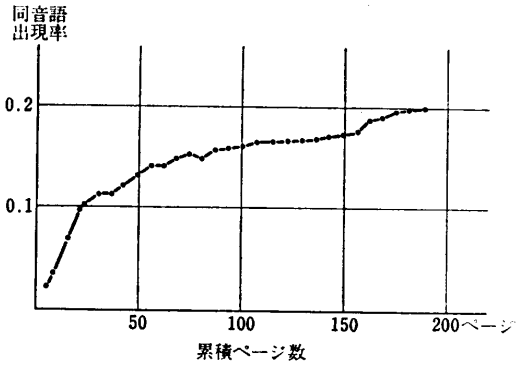


図2 同音語出現率の推移
Fig. 2 Change in the rate of homonyms.

この結果は以下のように解釈できる．

1) 資料中の漢字語は，各文献共通によく使われる語と，文献ごとに散発的に使われる語からなる．
2) 前者に属する語の新規出現率は累積とともに減少し，120ページ以降ではほぼ0となる．
3) すなわち，120ページ以降の新出語はほとんどすべて後者に属する語である．
4) 後者に属する語の出現率はほぼ一定で，その平均値は図1に記入してあるように5.8%である．

次に，**図2**にこの資料における同音語出現率を示す．この値は語の累積とともに増加し，3,671語を累積したところで約20%に達している．

3.3 変換性能の推定

以上の結果から，累積された約3,700語を収録した辞書を用いてかな漢字変換を行う場合の性能を以下のように推定できる．

1) 対象は情報処理に関する文献とする．
2) 1ページ中の漢字語の出現率は，表1にしたがい185語とする．
3) 1ページの総字数は1,400字とする*．また，

* 情報処理学会誌を例にとると，図，表，式などを除いた正味の字数はおよそこの程度となる．

1漢字語は2字からなるものとする.

4) 新出語，同音語出現率は，3.2節によりそれぞれ5.8%および20%とする.

5) したがって，1ページに出現する新出語，同音語はそれぞれ11語(22字)および37語(74字)となる.

6) これら48語(96字)以外の漢字語は正しく変換され，他に誤変換その他の誤まりはないものとして変換率を推定すると，

$\{(1,400-96)/1,400\}\times 100 \fallingdotseq 93\%$

となり，一応の水準に達していると評価できる[2].

対象を特定個人の文献に限ると，よく使われる語が上記資料に対して変動することは当然ありうる．したがって，変換辞書を動的に構成して，

1) 各文献共通によく使われる語を初期値とし，

2) 使用中に出現する新出語を登録するとともに，

3) 長期間使用されない語は削除する，

ことにより新出語出現率はさらに減少すると予想される．また，不要な語の削除により収録語数は一定に保たれ，同音語出現率は増加しないと予想される.

4. 変換システムの試作

4.1 処理形態

われわれの手許で利用できる計算機環境にあわせて，1)文は漢字表示機能のない通常のTSS端末から入力し，2)大型計算機センターの計算機で変換処理を行ってから，3)結果をホスト計算機の漢字プリンタに出力する，という処理形態とした.

この場合漢字出力を端末で見ることができないので，漢字語を辞書に登録する際に識別用の補助情報を付加し，必要な際に表示するようにしている*.

入力をカナ，ローマ字のいずれで行うかについては，次の理由によりローマ字を採用した.

1) おもな使用者は研究者であって，TSS端末やタイプライタの使用経験があり，欧文鍵盤配列に慣れている者が多い.

2) 大文字，小文字の使いわけにより，後述のように漢字とかなの指定を自然に行うことができる.

3) 漢字プリンタには，漢字以外にも通常の鍵盤にない多数の文字，記号がある．これらは鍵盤上にある文字，記号の組合せで表現するが，これもローマ字を使うほうが研究者には自然に感じられる.

4.2 漢字とかなの区別

ローマ字入力を漢字かなまじり文に変換するとき，漢字とかなの区別を入力者の指定によるか，自動的に行うかの二つの方向がある．われわれは以下の理由により前者を採用した.

1) ある語を漢字，かなのいずれであらわすかは自由度が大きく，規則性に乏しい.

2) すなわち，ある語を漢字であらわすか否かは書き手の意思によって決定される.

3) それゆえ，変換システムは書き手の意思をうけ，それにしたがって処理を行うようにすべきである.

欧文鍵盤では，シフト操作によって大文字と小文字を使いわける．これを利用し，以下のようにして漢字，かなの指定を行うこととした.

1) 漢字語の先頭文字を大文字とする.

2) 漢字語にひらがなが続く場合，空白を1個おく.

3) 漢字語に数字や記号が続く場合は，空白をおかずそのまま続ける.

4) カタカナや英字は特定の記号でかこんであらわし，これらが続く場合も空白なしにそのまま続ける.

以下に二，三の例を示す.

例1 Kanzi␣tokananokubetu␣wo Okona␣u
漢字とかなの区別を行う

例2 Syouwa 57 Nendo Yosan
昭和57年度予算

例3 Nihongo Nyuuryoku X sisutemu X
日本語入力システム

例3で，大文字Xはカタカナ部分の指示子である．なお，ローマ字表記は日本式[5]に準じ，それに外来音表記のためのものを加えている.

4.3 変換辞書の構成

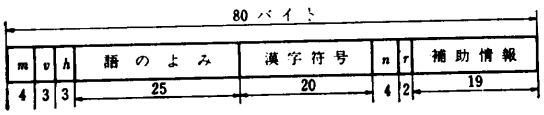
変換辞書の構成における主要な問題点は，辞書の容量および管理方式である．この2点に関し，試作システムでは下記のようにしている.

はじめに，容量は以下により2,500語とした.

1) 3章で述べたように，調査資料30篇にあらわれる漢字語の語種は3,671語(のべ34,906語)であるが，このうち1,378語は頻度1である.

2) 頻度1の語は辞書に収録する語の条件，「よく使われる語」を満たしているとはいえない．そこで，残り2,293語が収録する語の目安となる.

* たとえば，"Zyouhou"には「情報」,「乗法」などの同音語があるが，前者には"information"，後者には"kakezan"という補助情報がつけられていて，これを端末に表示して選択を行う.



m：登録番号
v：版番号
k：使用履歴
語のよみ：最大ローマ字5文字/漢字×最大5漢字/語
漢字符号：4桁*/漢字×最大5漢字/語
n：使用頻度
r：同音語数
補助情報：任意の文字列
* 漢字符号を16進文字0, 1, …, 9, A, B, …, Fであらわす

図3 漢字語辞書
Fig. 3 Structure of the Kanji-word dictionary.

3) 「よく使われる語」は個々の使用者により当然変化し，語数も変動すると予想される．そこで，上記2,293語に約200語の余裕をとり，2,500語とした．

次に，辞書の管理は語の参照頻度と間隔とにより，以下のようにしている．すなわち，辞書中の各漢字語には**図3**に示すように，その語の参照回数を示す頻度カウンタnと，前回参照されてからの経過期間を示す履歴カウンタhを設けてある．これら2種のカウンタを用い，以下の手順により辞書管理を行う．

1) システム起動時の初期化処理に際し，辞書中の全漢字語について $h+1\rightarrow h$ とする．
2) ある語が参照されたとき，その語について，$n+1\rightarrow n$, $0\rightarrow h$ とする．
3) 未登録語があらわれたとき，辞書が一杯であれば h/n が最大の語のなかから1語を削除し，そこに登録する．

以上により，使用を続けるにつれて，参照頻度が小さくかつ長期間参照されない語はしだいに追出され，よく使われる語が残って辞書は使用者に適応してゆく．

4.4 システム構成

試作システムは北大大型計算機センターの HITAC M-200H 上に作成されている．同計算機にはレーザビーム漢字プリンタが接続され，フロッピディスク等の媒体上に作成した漢字符号を入力して漢字かなまじり文を出力するソフトウェア (KHP) が提供されている[6]．そこで，試作システムは**図4**に示すように，入力文を漢字符号に変換して KHP の仕様に適合するファイルに出力する形態とした．

なお，同音語の選択は前述のように補助情報によって行う．現実の文書では同音語のうち1種のみが多用される場合が多いので，最初の選択の際に固定し，以後は自動的に選択する機能を組み込んである．

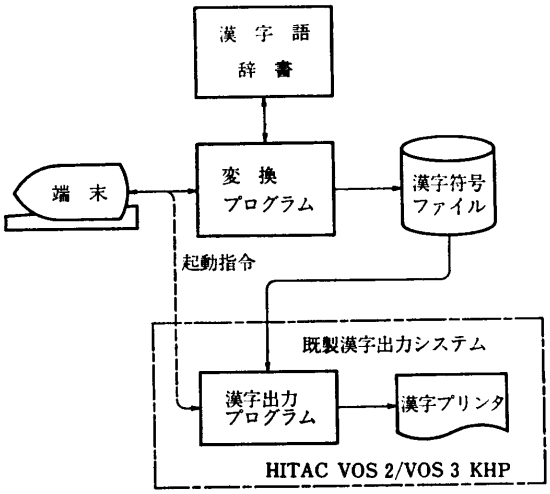


図4 試作システムの構成
Fig. 4 Schematic of the experimental system.

表2 実験結果の例
Table 2 Results of experiments.

		資料 1	資料 2	資料 3	資料 4	資料 5
入力字数, C		5,786	11,194	8,381	11,365	7,672
漢字語のべ語数, K		1,109	2,096	1,414	1,807	1,339
新出語	語数, W_n	71	114	65	61	48
	W_n/K	6.40%	5.44%	4.60%	3.38%	3.58%
同音語	語数, W_h	258	414	311	382	285
	W_h/K	23.26%	19.75%	21.99%	21.14%	21.28%
	W_h*	110	212	140	212	157
誤変換	語数, W_e	13	19	7	9	4
	W_e/K	1.17%	0.91%	0.50%	0.50%	0.30%
入力ミス	字数, C_e	10	23	20	13	7
	C_e/C	0.17%	0.21%	0.24%	0.11%	0.09%
正変換率		88.0%	90.0%	90.6%	91.9%	91.1%
正変換率*		93.1%	93.6%	94.7%	94.9%	94.5%

* 同音語選択の固定化処理を行った場合

4.5 実験結果

以下，試作システムで行った入力実験について述べる．入力資料は情報処理に関する学術文書で，したがって変換辞書の初期値には前述の調査で収集された頻度2以上の語，2,293語を用いている．

結果の一例を**表2**に示す．ここで，入力資料は著者の1人による論文，講演予稿で，内容は漢字処理，性能評価，言語処理，計算機設置計画などである．これから以下の諸点が示される．

1) 入力漢字語に対する新出語出現率は実験の経過につれて6.4%から3.5%前後へと低下し，辞書

の内容が使用者に適応してゆくことを示している．

2) 同音語出現率は入力漢字語の21%強で，3.2節で得られた結果とほぼ一致する．

3) 誤変換率は入力漢字語の1.2〜0.3%で，新出語と同様に減少傾向が認められる．1語を2字に換算すると，これは入力字数の0.5〜0.1%に相当する．

4) 出力にあらわれる誤まりは誤変換と打鍵の誤まりとの和であり，入力字数の0.6〜0.2%である．

この結果は1人の使用者によるもので入力量も十分ではなく，これだけで本方式の性能を断定することはできない．しかし，以下に示すようにこの結果は一般性をもつと推定できる．

1) 資料1の入力は，変換辞書が初期状態で特定の使用者に適応していない状態で行われている．このとき，新出語出現率は6.4%で，3章で述べた値5.8%と同程度である．

2) 各資料の同音語出現率は20〜23%で，3章で述べた値にほぼ一致する．なお，4.4節で述べた同音語選択固定化機能により，このうち42〜55%が自動選択される．

5. おわりに

欧文鍵盤をもつ通常のTSS端末から日本語文をローマ字表記で入力し，小容量の動的変換辞書を用いて漢字かなまじり文に変換して出力する研究者個人使用向きシステムを試作し，実験を行った．

本システムの性能は以下のように要約できる．

1) 入力文中の漢字語のうち，辞書に未登録のものは当初5〜6%であり，使用につれて辞書が使用者に適応して3〜4%に減少する．

2) 同音語は入力漢字語の20〜23%存在するが，最初の選択に固定する機能によりその42〜55%は自動選択可能である．

3) 出力にあらわれる誤まりは入力字数の0.6〜0.2%である．

4) 以上から文字単位の正変換率を求めると，同音語選択固定化機能を用いた場合93〜95%となり，文書処理用かな漢字変換システムとしてほぼ実用水準に達していると評価できる[7]．

一方，問題点としては，以下の諸点がある．

1) 新出語登録の際，漢字符号を求めるために符号表をひく必要があり，手間がかかる．

2) 同音語はかなりの頻度で出現する．したがって選択操作の改善，自動選択処理の高度化など，種々の面で改良をはかる必要がある．

3) 慣れた使用者にとっては，端末に表示されるメッセージが長すぎる感があり，省略形を用意することにより高速化をはかる必要がある．

4) 本方式の有効性を一般的に検証するため，より広範な使用者，分野における入力実験が必要である．

これらの問題点について現在引続き検討を進めており，その結果についてはあらためて報告したい．

謝辞 種々ご討論，ご示唆をいただいた本学部電子工学科電子機器工学講座，永田邦一教授ならびに講座各位に厚く御礼申し上げます．また，卒業研究として資料調査，プログラム作成をお手伝いいただいた高平敏男，吉田裕司，岡沢好高各氏にあわせて感謝します．

参考文献

1) 神田泰典：日本語情報処理の新しい展開，事務管理，Vol. 19, No. 8, pp. 20-23 (1980).

2) 森 健一，河田 勉：かな漢字変換，情報処理，Vol. 20, No. 10, pp. 911-916 (1979).

3) Peterson, J.L.: Computer Programs for Detecting and Correcting Spelling Errors, *Comm. ACM*, Vol. 23, No. 12, pp. 676-687 (1980).

4) 栃内香次，高平敏男，斉藤 康：研究室向き日本語文入力方式の検討，情報処理学会第21回大会講演論文集，pp. 1045-1046 (1980).

5) 武部良明：日本語の表記，pp. 257-283，角川書店，東京 (1979).

6) 日立製作所：HITAC漢字編集プログラムKHP機能編，日立製作所ソフトウェア工場，横浜 (1980).

7) 森 健一，天野真家：日本語ワードプロセッサとテキストエディタ，電子通信学会誌，Vol. 63, No. 7, pp. 729-733 (1980).

(昭和57年1月7日受付)
(昭和57年10月4日採録)