

適応機能をもつ変換辞書を備えた日本語スクリーンエディタ

駒谷喜代俊 中西教子
(三菱電機 (株) 中央研究所)

1. はじめに

計算機による日本語処理は、汎用計算機からパーソナルコンピュータにまで広がり、ワードプロセッサの登場によって一般にも身近な存在となってきた¹⁾。市販のワードプロセッサはそれぞれに工夫をこらし、誰でも使えるようになっている反面、従来からの計算機利用者にとっては、キーの配置や操作方法に違いがあって必ずしも使いやすいとはいえない。また、文法情報に基づいた変換が主で、意味的に正しい変換を行うには最終的に利用者の選択が必要になる場合がある。しかし、今後は日本語の文書作成だけでなく、日本語による表示や入出力は計算機にとって欠かせないものになることは明らかである。

そこで、我々は従来からあるテキストエディタになれた人を対象として次のような特徴を持つ日本語スクリーンエディタ *K e d i t* を開発した²⁾。

- ① 利用者の使用状況に自動的に適応する
- ② 従来のエディタと同じイメージで使える
- ③ ハードウェアを特定しない

①は既存のワードプロセッサが普遍性のある文法情報を用いているのに対して、個人的な使用情報に基づいたワープロを目指している。これは、ある利用者が頻繁に使う言葉はあまり多くなく、その中には同音語も少ししかないという仮定をもとにしている³⁾。そして、利用者の用語や語法を自動的に登録して、それだけの情報でどの程度まで効率よい漢字変換ができるかを検討した。

②は従来からプログラムや英文の作成にスクリーンエディタを利用してきた人が違和感なく利用できることを目的としている。また、日本語と英語が同じイメージで作成・編集できることも目的の一つである。このため、*k e d i t* には日本語用と英語用のエディタがあり、ユーザはコマンドを任意の制御キーに割り当てることができる。この2つのエディタは同じ方針のもとに開発し、ソースコードも共通した部分が多い。現在は主に計算機関係の研究者が日常の文書作成や日本語を含んだプログラム開発に利用している。

③はパソコンから汎用機までの様々なハードウェア上で同じエディット環境を実現することを目的としている。このため、*K e d i t* はC言語で記述しており、現在はパソコン (Multi-16) とスーパーミニコン (MX/3000, VAX-11) 上で稼働している。その他、UNIX, CP/M, MS-DOSなどが動く機種へは移植可能である。

ここでは、上で述べた適応機能が漢字変換に与える効果について実験を行ったのでその結果を報告する。まず、2章でこの日本語スクリーンエディタの概要、3章で適応機能の詳細について説明し、4章、5章で適応機能の効果を調べるために行った実験方法とその結果について述べる。

2. *K e d i t* の概要

K e d i t には日本語用と英語用の2つのエディタがある。これらは、同じ方針のもとに開発されたもので、ファイルI/O・編集機能ではステップ数で約5割が共通している。コマンド数は日本語特有のものが17、英語特有のものが12で、その他55個が共通しており、

各コマンドは制御キー（CTRL,ESC）を用いた入力方法をとっている。そして、日本語用と英語用が同じイメージで使えるように、同じ内容のコマンドは同じキーに割り当てている。また、キーへの割当は利用者が自由に変更することができる。

日本語の入力には、新たにキー配置を覚える必要のないようにローマ字入力を採用している。そして、大文字と小文字の区別を利用して、大文字で漢字部を指定する方式をとっている。これは、利用者を従来のテキストエディタになれた人と想定していることと、ある言葉を漢字で書くか、かなで書くかは書き手の意志によるところが大きいと考えられるからである。また、漢字部指定方式は打鍵数が多くなるが、英文の入力に慣れていればシフト操作はあまり苦にならないと考えたためである。なお、漢字、ひらがな以外のカタカナ、数字、アルファベットは、表1に示した方法で区別している。

字種	ローマ字	例
漢字	大文字	KANJI -> 漢字
ひらがな	小文字	kanji -> かんじ
カタカナ ^④	対応する英単語	system -> システム
数字	そのまま	3.14 -> 3.14
アルファベット	先頭に¥をつける	¥Kanji -> Kanji

表1 ローマ字による字種指定方法

^④ただし、ひらがなをカタカナに変換するコマンドも用意している。

変換キーには次の3種類がある。英文と同じイメージで操作できるように、もっとも基本となる変換キーにSPACEキーを割り当てている。ただし、SPACEキーが単独で入力された場合は全画の空白に変換される。ESCAPEキーは英単語や英文で参考文献を書く場合に用いるが、特に日本語を含んだプログラム開発のためにすべて半画のまま入力するモードも用意している。

- ① SPACEキー（□で示す） 変換のみを行う
- ② RETURNキー（☐で示す） 変換と改行を行う
- ③ ESCAPEキー（☒で示す） 変換せずに半画のまま入力する

入力の単位に関しては、入力例1)のように漢字とひらがなが交互に並ぶような場合にはべた書きでよいが、それ以外の字種については入力例2)のように字種の区切りごとに変換キーを入力する必要がある。しかし、漢字とひらがなだけの場合にでも、入力の文字列が長くなると、利用者自身も読みづらくタイプミスが増えることや後述する同音語の処理のためにも文節単位の入力が望ましい。

入力例

- 1) KYOUwaAMEdesuga,ASUwaKAIFUKUshimasu.☐
->今日は雨ですが、明日は回復します。
- 2) Kedit☒wa☐¥C☐GENGodeKIJUTSUshiteiru.☐
->KeditはC言語で記述している。

さて、変換辞書には各利用者が共有するシステム辞書と個別に持つユーザ辞書があり、ユーザ辞書はさらに漢字用とかな用の2種類がある。入力ローマ字列の大文字部分つまり漢字はまず漢字用ユーザ辞書を検索し、変換できなければシステム辞書を検索する。小文字部分はかな用ユーザ辞書を検索し、なければ文字単位に分解して変換する。システム辞書は約21000語の見出し語を収録しており、見出し語（ローマ字表記）とJISコードを併記した構造をとっている。また、この辞書は追加、削除、変更等はできない。漢字用ユーザ辞書は、見出し語とコード以外に接続情報・頻度情報などを持っており、最大3000語の見出し語を登録できる。漢字用ユーザ辞書への登録は、利用者が画面に表示された文字を切り出して行う方法と、適応機能により自動的に行われる場合がある。また、かな用ユーザ辞書はかな文字と利用者が登録した英単語を収録しており、見出し語とコードだけを併記している。辞書への登録は漢字の場合と同様で、利用者が画面の文字を切り出して行う。なお、ユーザ辞書はシステム辞書と異なり、辞書内容を日本語文書と同じ方法での編集することができる。

3. 適応機能

かな漢字変換方式で効率良くかつ正確に日本語を入力するためには、複合語と同音語の処理が重要な問題となる。辞書に複合語をすべて登録していけば、辞書は際限なく大きくなり、検索に多大な時間を必要とする。逆に、複合語が登録されていなければ、利用者自身が分割して入力しなければならず、非常に使いにくくなる。また、文節入力方式では自立語の接続情報などに基づいて、自動的に分割して変換する方法も考案されているが⁴⁾、100%完全な変換は望めず検索回数がふえるため処理時間は多くなる。また、同音語の処理では品詞や活用などの文法情報を用いればかなりの程度まで正しく変換できるが、それでも利用者がまったく選択しなくてよいというわけではない。このように複合語・同音語の変換効率を高めようとするれば、辞書の容量は増大し、変換処理系も複雑なものになる。このため、処理時間は多くなり、利用者にとっては不透明な処理系となる恐れがある。しかし、複合語や同音語が適切に変換されないと、選択などのために入力を中断する時間が多くなり、非常に使いにくい処理系になる。

そこで、我々は各個人が頻繁に使う複合語や同音語の数はそれほど多くないと仮定して、利用者の使用状況に応じて辞書の内容を自動的に適応させる機能をもつ日本語スクリーンエディタを開発した。これは、一度使った言葉は次からは自動的に正しく変換できるように、その接続情報などを含めて漢字用ユーザ辞書に自動的に登録し、逆に長い間使われていないものは削除する機能に基づいている。このため、漢字用ユーザ辞書には見出し語とJISコード以外に、図1に示した頻度情報、最新の使用情報、登録情報を併記している。さらに、用言や同音語については利用者が用いた付属語を接続情報として登録している。

(頻度情報)(使用情報)(登録情報)見出し語 □ コード
 (頻度情報)(使用情報)(登録情報)見出し語 □ コード □ コード □ . . .
 (付属語 □ コード)

図1 漢字用ユーザ辞書のデータ構造

さて、複合語の処理については、二回連続して漢字変換が行われると、それらを連結し複合語として自動的に登録する。たとえば、利用者がまず、“JOUHOU”と入力し変換を行い、

続けて"SHORI"と入力し変換を行うと、「情報処理」と表示される。このとき、漢字用ユーザ辞書には

JOUHOUSHORI 情報処理

が自動的に登録され、登録情報としては複合語に対応した値が入る。これにより次からは、"JOUHOUSHORI"と続けて入力しても「情報処理」に変換できる。また、利用者は何が複合語として登録されているか知らないわけであるから、はじめから"JOUHOUSHORI"と入力し変換する場合がある。このときは変換に失敗するが、K e d i t ではテキスト中に"JOUHOUSHORI"の文字列がそのまま残されている。そこで、カーソルをもどし、"JOUHOUSHORI"として変換すれば正しく表示され、辞書にも登録される。

また、同音語については、できるだけ利用者が選択する割合を減らすために、次のような処理を行っている。まず、システム辞書で同音語があり利用者の選択により変換できた場合には、それが漢字用ユーザ辞書に自動的に登録される。ただし、その見出し語がユーザ辞書にすでにあれば、利用者の確認をうけてから登録する。また、用言の場合には語幹だけを登録し、送りがなは付属語として登録する。この場合、登録情報としては同音語に対応した値が入る。次に、ユーザ辞書に同音語があり利用者の選択により変換できた場合には、入力文字列に付属語があればユーザ辞書に登録する。このとき、選択された言葉に対してすでに付属語が登録されておれば、共通する部分だけを抽出して登録する。例えばユーザ辞書に"00"に対して(kii 大)と付属語が登録されており、"00kikunaru"が「大きくなる」と変換された場合には、"kii"と"kikunaru"から

(ki* 大)

と登録しなおされる。

そして、ユーザ辞書を検索して同音語があれば、まず付属語を調べ完全に一致すれば、自動的に変換する。また、完全ではないが付属語の先頭からある字数が一致すれば、それを優先して表示し利用者の確認をうける。それ以外の場合または付属語が登録されていないければ、最終使用語から順次表示していき、利用者の確認によって決定する。

このようにして、漢字用ユーザ辞書に次々と登録していくと、特に複合語において非常に長いものや無意味なものが合成されることがある。そこで、ある字数以上の複合語は登録しないように指定することができる。また、ユーザ辞書の見出し語数が一定値を越えると、利用者に整理をするようメッセージを出している。整理には頻度情報、使用情報、登録情報を用いて、ある期間を指定してその期間中に使用していないものを利用者の確認をうけて削除する方法がある。ただし、登録情報としては複合語、同音語、利用者登録の3つがあり、複合語だけを削除の対象としている。また、利用者が辞書自体を文書と同じように編集し削除することもできる。

以上のような処理を続けていくと、漢字用ユーザ辞書はしだいに各利用者の用いる用語や語法に適應していき、その結果、漢字変換の状況は向上していくことが期待される。

4. 実験方法

前章で述べた適応機能の効果を評価するために、K e d i t が正常に終了することに、図2に示した形式のファイルを作成し、必要なデータを自動集計している。

```

***  USER-NAME      ***  <( file-name )>(0)
5-OCT-1984 13:16:34.76 --- 13:28:01.56   0:12
[114   2] [(74, 0) (6, 4)]   [45 0, 0, 0] [19 4]
<(0,-0) (9,2,7,-0)>         {49,307,133}

```

図 2 エディタの使用結果例

この図の第3行目は左から順に、かなへの全変換回数(h1)と失敗回数(h2)、漢字への変換回数(k0)と直接システム辞書により変換した回数(k1)、漢字変換の全失敗回数(k2)および辞書になかったことによる失敗回数(k3)、漢字用ユーザ辞書で同音語がなく変換できた回数(k4)・付属語から判断して変換できた回数(k5)・優先表示した言葉が正しかった回数(k6)・利用者の選択により変換できた回数(k7)、さいごにシステム辞書で同音語がなく変換できた回数(k8)と選択により変換できた回数(k9)をあらわしている。つまり、

```

[h1   h2] [(k0, k1) (k2, k3)]   [k4 k5, k6, k7] [k8 k9]

```

また、4行目はかぎっこ内がかな用と漢字用のユーザ辞書における語数の増減を、中かぎっこ内が変換された記号・かな・漢字の文字数をあらわしている。ただし、漢字用ユーザ辞書の増加については複合語および同音語として登録された語数を別に示している。

さて、我々は2つの実験を行い、変換率および選択回数の変化やユーザ辞書の登録状況などを集計し、適応機能の効果を検討した。まず、第一の実験では漢字用ユーザ辞書を空の状態からはじめて同一分野の文献を入力した。漢字変換の回数が約300回ごとにデータをとり、2つの文献で文字数にして約11000字の入力を行った。

第二の実験として、K e d i t を59年5月から5ヶ月間にわたり約30名の研究者に日常的な文書作成に自由に利用してもらい、その使用結果を集計した。K e d i t 自体はそれ以前から利用されていたが、適応機能は実現されていなかった。そして、今回の実験では利用者に適応機能の内容は説明しておかなかった。また、利用者には全く自由な使い方をしてもらっているので、一回の使用結果は変換回数がまちまちである。そこで、漢字変換500回を基準にして何回分かの使用結果をまとめて集計している。なお、この場合にも各利用者の漢字用ユーザ辞書はほとんど空の状態から始めている。この複数の利用者による全体の利用状況は1ヶ月間で約300時間、約15万字である。

これらの実験より、漢字変換に関して次の①～⑦の割合を求めた。

- | | |
|---------------------------------|-----------|
| ① ユーザ辞書で同音語がなく変換できた割合 | k4/K |
| ② システム辞書で同音語がなく変換できた割合 | k8/K |
| ③ ユーザ辞書で同音語があったが、選択作業なしに変換できた割合 | (k5+k6)/K |
| ④ ユーザ辞書で同音語があり、利用者が選択して変換できた割合 | k7/K |
| ⑤ システム辞書で同音語があり、利用者が選択して変換できた割合 | k9/K |
| ⑥ 変換できなかった、あるいは、変換しなかった割合 | k2/K |
| ⑦ その中で辞書にないので変換できなかった割合 | k3/K |

ただし、 $k = k0 + k1 = k2 + k4 + k5 + k6 + k7 + k8 + k9$ は全変換回数を示す。

また、かな変換については変換できなかった割合だけを集計している。

なお、ここでは使用結果を自動的に作成しているため、漢字に変換できたが考えていたものではないという場合のデータは正確につかめていない。これには、タイプミスによるもの、ユーザ辞書で変換されたことによるもの、およびシステム辞書で変換されたが違うものであった場合が考えられる。この中で処理系自体に問題があるのは、2つ目のユーザ辞書で変換されたが違うという場合である。これがおこると利用者は、直接システム辞書を検索して変換しようとするはずである。これは図2のデータから $k1/(k0+k1)$ により求めることができるが、全体で約2%以下である。

5. 実験結果

Keditの適応機能を評価する指標としては、利用者が選択を行わねばならない割合および漢字変換の失敗率の減少が挙げられる。

まず、第一の実験の結果をグラフに表したものが図3である。図中1~9はKeditの使用回数を示しており、1~4回目と5~9回目では文献が異なっている。図3-1の棒グラフは、前章の①, ②, ③, ④, ⑤, ⑥の割合を下から順に■, □, ▨, ▩, ▪, ▫で示している。図3-2の棒グラフは左側が漢字変換の回数、右側がかな変換の回数を示している。また、実線は⑥の割合、破線は⑦の割合、一点鎖線は全変換（漢字変換とかな変換を合わせた）回数に対する変換失敗の割合、点線はかな変換の回数に対する変換失敗の割合を示している。図3-3は、漢字用ユーザ辞書の累積登録状況を示し、下から順に利用者が登録した語数、同音語の選択により自動登録された語数、複合語として登録された語数を表している。

図3-1をみると、全体としてしだいにシステム辞書やユーザ辞書で選択を必要とする割合（▪+▩）と変換できなかった割合（▫）が減少し、ユーザ辞書による変換の割合（■+▨）が増えている。また、図3-3をみると、ユーザ辞書の登録語数がしだいに増えてきている。これらから、適応機能の効果が予想した通りであることがわかる。なお、5回目にユーザ辞書による変換の割合（■）が減少しているが、これはここで入力する文献が別のものに移ったためと考えられる。また、4回目と9回目にもやや減少しているのは、これらは「あとがき」の部分にあたることから本文と比較して多少語調が変わるのではないかと推測される。辞書の登録状況に関しても、4回目から5回目にかけての

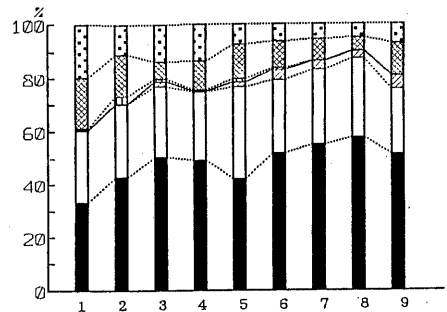


図3-1 文献入力による漢字変換状況

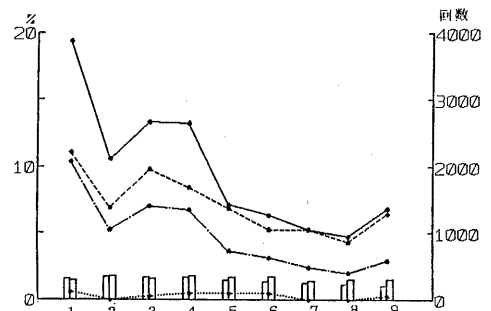


図3-2 文献入力による変換回数と失敗状況

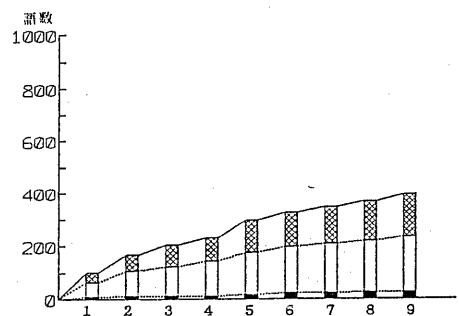


図3-3 文献入力によるユーザ辞書登録状況

伸びが大きいのは同じ理由と考えられる。

図4, 図5は第二の実験で二人の利用者について5ヶ月間の使用結果を第一の実験と同じ形式でまとめたものである。いずれも、第一の実験結果と同じ傾向を示しているが、使用量が大きくなるとどの割合についてもかなり増減の波が出てくる。これは、第一の実験でも文献が変わると選択や失敗の割合が増えたように、同じ利用者が書く文章であっても、少しずつ語いが異なってくるためと考えられる。次に、図4-1, 5-1では特にシステム辞書で選択を必要とする割合(○)が使い始めの頃と比べて後半では著しく減少している。また、図4-3, 図5-3では、システム辞書からユーザ辞書へ自動登録された語数(□の部分)がしだいに飽和状態に達してきている。これはシステム辞書で同音語をもつ言葉でも、利用者が頻繁に使うものは最初のうちにほとんどユーザ辞書に登録されてしまい、その後はユーザ辞書で変換され選択の必要がなくなるためと考えられる。

さらに、適応機能の有効性を詳しく調べるため、図4, 5の利用者を含めた5名の使用結果から、漢字変換の回数2000回を一つの期間とし各期ははじめの500回について自動選択(③)、利用者による選択(④+⑤)、漢字の変換失敗(⑥)、かなと漢字を合わせた失敗の各割合(いずれも

回数比)を表2~5にまとめた。表中の'A', 'B', ...は利用者を示し、'I'~'VII'は第何期目かを表している。また、5月以前は適応機能を持たないエディタを利用していたが、そのときの使用結果を比較のために第'0'期として載せている。表2によると、しだいに自動的に選択される割合が増えてきており、同音語に関する適応機能が有効的に働いていることがわかる。表3については、第I期では利用者が選択しなければならない割合が非常に大きい。しかし、時間がたつにつれてしだいに小さくなり、第0期に比較して十分小さくなっている。これは、ユーザ辞書への自動登録とともに、複合語がつけられたことによっても選択の割合が減少していると考えられる。表4, 5でも第I期では変換に失敗す

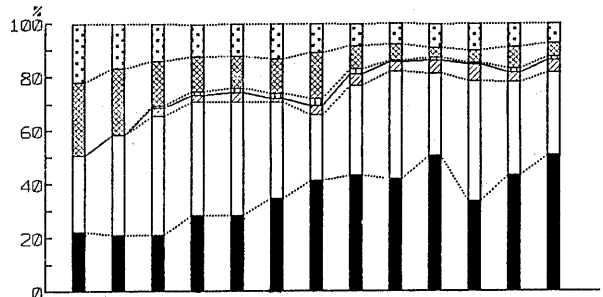


図4-1 利用者Aによる漢字変換状況

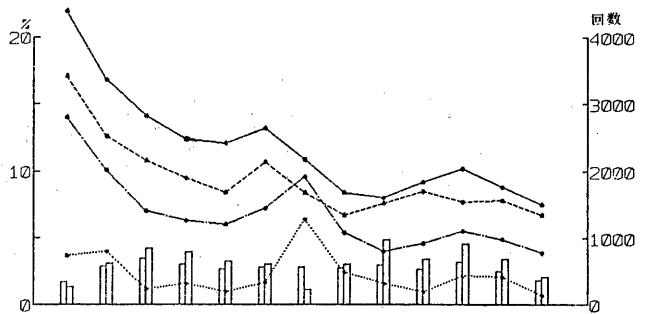


図4-2 利用者Aによる変換回数と失敗状況

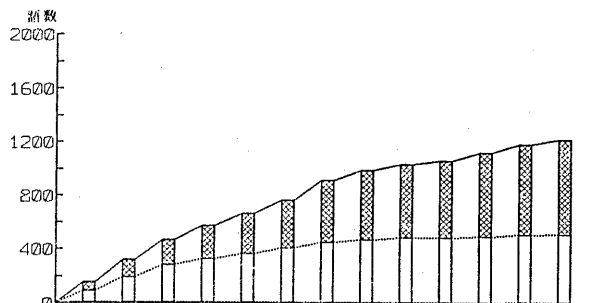


図4-3 利用者Aによるユーザ辞書登録状況

る割合が大きい、やはりしだいに小さくなってきている。また、表に載せていない他の利用者についても同様で、適応機能のない第0期と比較して良くなってきている。以上から、確かにK e d i tの適応機能は漢字変換の効率を向上させているといえることができる。

さて、図3-2、4-2、5-2から、かな変換と漢字変換の失敗の割合を比較すると、後者の方がかなり高い。漢字変換の失敗のうち、辞書に見出し語がない場合（破線）がほとんどである。この失敗した語について各利用者ごとに調べてみると、その語が辞書から

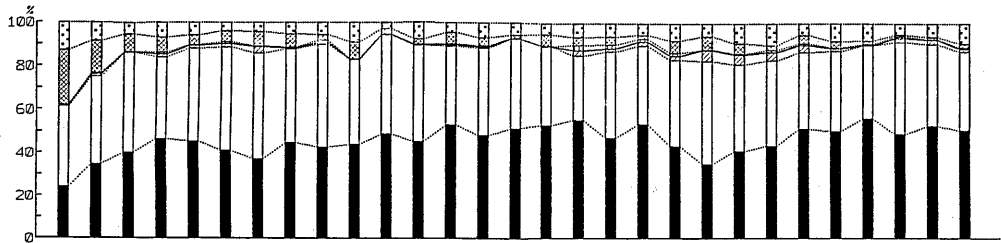


図5-1 利用者Bによる漢字変換状況

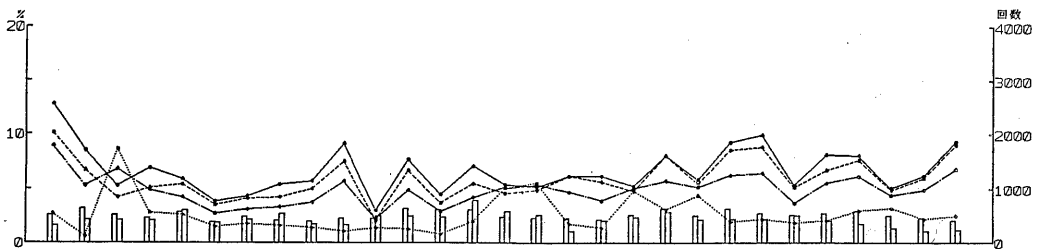


図5-2 利用者Bによる変換回数と失敗状況

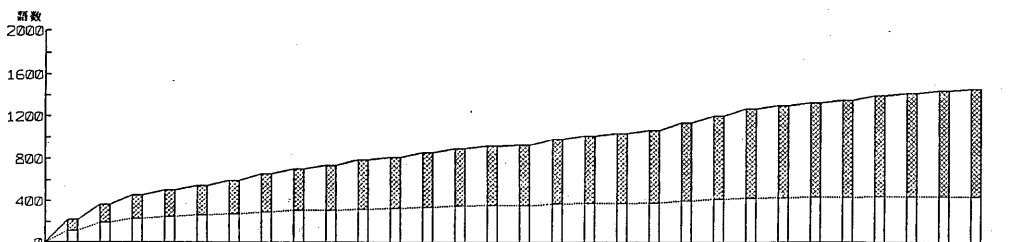


図5-3 利用者Bによるユーザ辞書登録状況

	A	B	C	D	E
O	-	-	-	-	-
I	0.0	0.3	0.0	0.0	0.5
II	3.3	1.5	0.0	2.2	0.2
III	3.3	1.0	0.9	4.2	1.4
IV	4.3	0.8	1.9	4.5	3.4
V	-	2.4	2.0	4.5	-
VI	-	5.1	-	-	-
VII	-	3.1	-	-	-

表2 自動選択された割合(%)

	A	B	C	D	E
O	35.5	19.5	8.9	21.7	13.7
I	27.4	25.1	20.1	20.9	21.7
II	13.5	4.6	7.8	13.8	6.1
III	6.7	3.5	3.3	12.8	4.6
IV	6.4	5.5	7.0	7.7	1.8
V	-	6.3	4.9	7.9	-
VI	-	6.3	-	-	-
VII	-	4.8	-	-	-

表3 利用者が選択した割合(%)

	A	B	C	D	E
O	10.9	5.2	3.3	4.7	8.0
I	22.0	12.8	7.6	7.7	9.5
II	12.1	5.9	5.2	7.6	5.9
III	8.0	5.7	4.4	14.0	4.4
IV	7.5	4.5	5.5	8.7	3.8
V	-	6.2	2.5	6.8	-
VI	-	5.9	-	-	-
VII	-	5.4	-	-	-

表4 漢字変換失敗の割合(%)

	A	B	C	D	E
O	6.9	3.8	3.4	3.4	4.2
I	14.0	8.9	4.1	5.6	6.3
II	6.0	4.2	2.7	4.8	3.0
III	4.0	3.7	2.5	10.9	2.5
IV	3.9	2.9	4.6	4.7	3.2
V	-	4.7	1.4	3.7	-
VI	-	5.2	-	-	-
VII	-	3.7	-	-	-

表5 かな+漢字変換失敗の割合(%)

もれているもの、送りがなの間違い、タイプミスに分かれる。このうち、タイプミスと考えられるものが50%~60%を占めており、漢字変換の失敗の多さの一因となっている。タイプミスの多さについては、日頃漢字で書いている言葉を改めてローマ字で表記することの難しさに原因があると考えられる。

さいごに、漢字用ユーザ辞書の内容については使わないような複合語が次々に登録されて、辞書を大きくしている。各利用者の登録語数が1000語程度で自動的に整理するコマンドを実行すると、約300個の複合語が削除される。また、同音語の登録状況は利用者によって異なるが、1000語について30~40語程度である。

6. おわりに

以上、利用者が用いる用語や語法をユーザ辞書に登録し漢字変換の効率向上をはかる適応機能について述べ、その効果を実際の使用結果から確認した。今後、漢字変換の失敗の大部分を占める入力のタイプミスに対処していけば、変換の効率はさらに良くなると考えられる。現在は変換に失敗した場合、その入力ローマ字列をそのまま画面に残し、誤りを訂正し再入力を行えるようにしている。そこで、この修正作業から各利用者ごとに入力のタイプミスの傾向をもとめ、変換に失敗した場合にはそれをもとに自動的に正しい表記に訂正して、再び辞書を検索する方法を検討している。

7. 参考文献

- 1) 森、河田：「かな漢字変換」，情報処理，vol.20，no.10，pp.911-916 (1979)
- 2) 駒谷、中西：「適応機能をもつユーザ辞書を備えた日本語スクリーンエディタ」，情報処理学会第29回全国大会，6J-2，pp.1483-1484 (1984)
- 3) 栃内、斉藤：「適応型変換辞書を用いるかな漢字変換」，情報処理論文誌，vol.24，no.2，pp.209-213 (1983)
- 4) 藤崎、大河内、諸橋：「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」，情報処理論文誌，vol.23，no.1，pp.1-8 (1982)