計算言語学 23－1
（1980. 6. 13）

# IDIOM形カナ漢字変換システムの試作

木村久正　白鳥嘉勇　小橋史彦　山階正樹　萩野輝雄
（日本電信電話公社　横須賀電気通信研究所）

## 〔1〕まえがき

情報処理技術の発展に伴ない，漢字情報を電子計算機に入力する必要性が高まっている。しかし，各種の入力方式が提案されているものの，まだ標準方式といえるものはない現状にある。こうした中で，カナ漢字変換方式は汎用のカナキーボードを使用して簡単な操作で入力が可能なことから注目され，各所で研究が進められている。[1)〜3)]

カナ漢字変換方式における最も重要な問題は同音語の同定法にあり，従来，大別して２つの観点から検討が進められていた。ひとつは，あくまで同音語の同定を電子計算機により自動的に処理しょうとするものである。すなわち，形態素解析法，構文解析法，意味解析法などの自然言語研究の成果を駆使して処理しょうとするものである。この場合，処理が複雑化して記憶容量，処理時間が膨大となるほか，意味解析法にいたってはアルゴリズムも確立しておらず，十分な効果をあげるにいたっていない。他のひとつは，人間の助けも借りて同音語を処理しょうとするものである。すなわち，同音語の同定に有効な情報を付加入力し，比較的処理アルゴリズムが確立している形態素解析法を組合わせて処理しょうとするものである。この場合，入力操作に多少の負担がかかるものの，付加の仕方が簡単であれば，当面では最も実際的な方法といえる。

我々は，後者の観点より，漢字の持つ図形的特徴に着目して簡単な符号を熟語の読みの前後に付加する入力方式（ＩＤＩＯＭ*方式）を既に提案している。[4)] この方式の特長は以下の通りである。

① 単純な変換アルゴリズムと小容量の辞書ファイルで高い カナ漢字変換率が得られる。
② 読み仮名通りに入力でき、素人にとって自己学習も可能であるため訓練期間が短かい。
③ ブラインドタッチによる高速入力が可能である。
④ 装置の小形化，経済化が図れる。

本報告では，このＩＤＩＯＭ方式の基本機能の確認を行うため、検討、試作した実験システムの概要と実験結果について述べる。

## 〔2〕IDIOM方式の原理

本方式は、日本語の性質を多角的に検討し、入力に利用できる性質を出来るだけ利用する事によって、従来問題となっていた同音語の同定について有効な解決策を与えるものである。その着眼点を以下に示す。

(1) 漢字の性質

漢字は（形・音・義）の性質をもつが、従来のカナ・漢字変換方式では「義（意味）」と，そのつながり方などの分析に重点をおき変換辞書が大きくなる方式が多く，そのわりに同音異義語の同定率はよくなかった。

本提案では、漢字が部首の集合であ

*IDIOM = Immediately DIstinguishing code Of Multitudinous kanji character
（人間が瞬間的に漢字の特徴を分類識別できる符号の意）

り形によって区別される事から，「音（よみ）」に「形」の情報を付加（入力時に）すれば漢字の同定に有効で，その結果高い変換率が得られるという観点で検討した。

「形」の情報としては，日常良く使われる字種の範囲（当用漢字レベル）では，部首の組合せ方が数種のパタンに分類される性質を応用した。[5)]

(2) 熟語の性質

漢字の読み方は複数あっても，漢字を組合せた熟語は読みがほぼ一義となる。（なお，熟語は日常良く使われる国語辞書の大半　約70%を占める）

さらに，前項（1）の付加情報の組合せは，同音異義語を分類できるケースを増大させる。

(3) 人間の図形認識力

人間が漢字を読む場合，まず大まかな図形的特徴を分類すると考えられ，上記程度のパタンの分類は瞬時（0.1秒以下）で行える点に着目する。[5)]

これ等の性質を利用することにより，高い変換率のカナ－漢字変換方式が可能となる。すなわち，従来の漢字指定方式（漢字で表記する単語の前後を漢字区切り符号で囲む）の漢字区切り符号のかわりに，本方式は数種の図形符号（ＩＤＩＯＭ符号）で囲み図形情報により同音語の同定を行う。次に例を示す。また，ＩＤＩＯＭ符号例を表１に示す。

〈例文〉「漢字の特徴を抽出する。」

〈入力操作〉

「□かんじ□の□とくちょう□を□ちゅうしゅつ□する」

〈特徴分類方法と同音異義語〉

漢字　特徴　抽出

同音異義語：幹事　特長

監事

表１　ＩＤＯＭ符号例

| 1種 | 2種 | 3種 | 呼名 | 漢字例 |
|---|---|---|---|---|
| 漢字区切り符号 | 田 | 田 | 縦割り | 漢，幹，特 |
| | 口 | 日 | 横割り | 字，監，吉 |
| | | 口 | その他 | 事，長，出 |

本提案の方式は，字形の特徴を符号化した少数のキーをカナキーボードに追加し，漢字部分を示す区切り符号と兼ねて付加する事により具体化できる。

## 〔3〕入力規則

① 漢字単語部分をＩＤＩＯＭ符号で囲み，他はべた入力とする。

② 漢字部分が複合語となる場合は単語単位で入力する。

③ 漢字１字からなる語（主に用言と接辞）を入力する場合は，後のＩＤＩＯＭ符号としてダミー符号（☒）を入力する。

④ カタカナ，漢数字，英小文字を入力する場合は，ダミー符号で前後を囲む。

⑤ 人名は人名識別符号（㊫）を単語の最初に入力する。

表２　入力例

| 項目 | | 入力語 | IDIOM入力列 |
|---|---|---|---|
| 漢字2字以上の語熟語 | 名詞 | 漢字 | 田カンジ日 |
| | | 計算機 | 田ケイサンキ田 |
| | サ変名詞 | 検討する | 田ケントウ田スル |
| | 形容動詞 | 簡単な | 日カンタン日ナ |
| | 複合動詞 | 取去る | 田トリサ口ル |
| 漢字1字の語 | 体言 | 花が | 日ハナ☒ガ |
| | 用言 | 開く | 口ヒラ☒ク |
| 複合語 | | 情報処理 | 田ジョウホウ田口ショリ田 |
| | | 青少年 | 日セイショウネン口 |
| 接辞 | 接頭語 | 全日本 | 口ゼン☒口ニッポン口 |
| | 接尾語 | 人間的 | 口ニンゲン口田テキ☒ |
| | 数詞1 | 第5 | 日ダイ☒5 |
| | 数詞2 | ／2回 | ／2口カイ☒ |

## [4] 変換アルゴリズム

本システムの変換機能は，次の6種類の処理から成っている。図1に処理の流れを示す。

(1) 英数かな処理：英大文字，数字，ひらがな，記号の変換。

(2) 非漢字処理：英小文字，カタカナ，漢数字等の変換。

(3) 名詞処理：複合語の構成要素の内，漢字2字以上からなる単語（例；図形的，情報処理）の変換。一般語辞書の検索，IDIOM符号・名詞の適合性判定のみを行い文法処理は行わない。

(4) 接辞処理：複合語（数字を含む）の構成要素の内，漢字1字からなる語（例；全日本，人間的，第1回）の変換。接辞辞書の検索，IDIOM符号・接辞種別（接頭語，接尾語，数詞前，数詞後，の4種）の適合性判定による。

(5) 一般語処理：アルゴリズムの概要を図2に，同音語処理例を表3に示す。

① 一般語辞書の検索

② IDIOM符号の適合性判定

③ カナ語幹部（活用語尾ではないが送りがなである部分，例；下がる，美しい）の適合性の判定

④ 活用語尾表（品詞・活用形ー活用語尾）による活用語尾の適合性判定

⑤ 付属語接続表（自立語＜品詞・活用形＞ー付属語の接続）による付属語接続の適合性判定。付属語間の接続関係は見ないので，処理時間の短縮，接続表の少容量化が可能となった。

⑥ 同音語優先順位の決定（かな部分の整合文字数の多い語を優先し，同数の場合は頻度の高い語を優先する。）

⑦ 連濁処理（辞書に該当語のない場合），連濁（例；ほレ＋そら→ほレぞら）があったと仮定し，濁音を除去して再検索する。

(6) 人名処理：姓および名の変換。

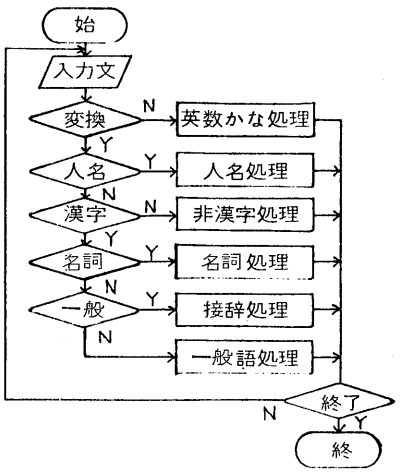



図1 変換処理の流れ図

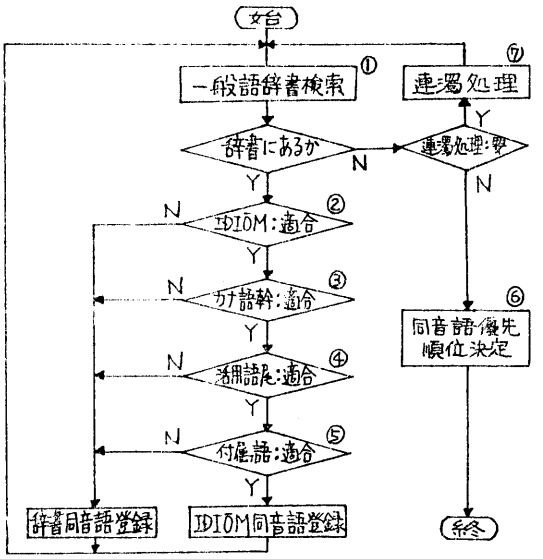



図2 一般語処理アルゴリズムの概要

表3 同音語処理例

<例> □サ ☒カバ ラナイ → 下がらない

| IDIOM符号 適合語［品詞］ | 差［名詞］ | 差す［サ五］ | 去る［ラ五］ | 退く(け)［下一］ | 下(げ)［下一］ | 下(が)［ラ五］ |
|---|---|---|---|---|---|---|
| カナ語幹 | — | — | — | × | × | ○ |
| 活用語尾 | — | × | × | | | 未然 |
| 付属語接続 | が | | | | | ない |
| 整合文字数 | 2 | 0 | 0 | 0 | 0 | ⑤ |

(注) IDIOM符号の適合性判定により「咲く」「指す」等は除外済み。

## 〔5〕同音語分布と推定変換率

一般語辞書の構成を図3に示す。以下一般語辞書における同音語分布と推定変換率を示す。

(1) 一般語辞書の同音語分布

頻度を一様とした場合の同音語分布を図4に、頻度を考慮した場合の同音語分布を図5に示す。頻度を考慮した時ではキー数2～3の場合、1キーの場合と比べると一意に定まる単語の出現度数は1.4～1.5倍に増加し、累積同音語の出現度数は70～80%に減少する。

図4と図5とを比較すると、頻度一様な場合と比べて頻度を考慮した場合では、同音語分布曲線の収束が遅くなっている。これは、同音語構成語数(以下同音語数と略す)が多い単語ほど頻度が高く、同音語数1～8程度で頻度は同音語数に伴なって増加することによる。

(2) 同音語分布からみた推定変換率

図5を基に単語変換率*，一般文変換率**の推定を行ないその結果を図6に示す。また、図6において実線は頻度を考慮した場合，破線は頻度を一様とした場合の変化率を示している。3キーの場合の単語変換率は頻度を考慮した場合92%であり、一般文変換率は97%程度が見込める。

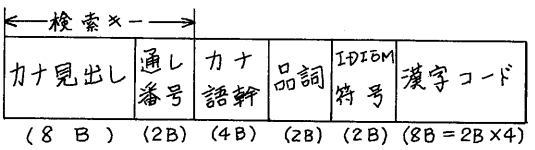



図3 一般語辞書の構成(24,000語)

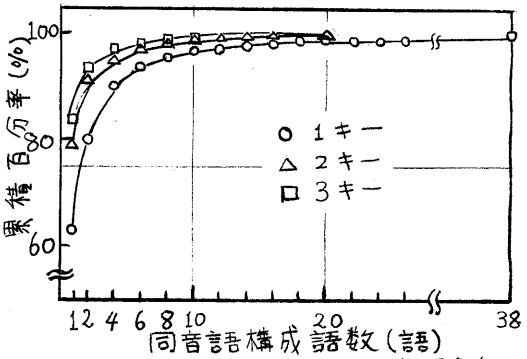



図4 頻度を一様とした同音語分布

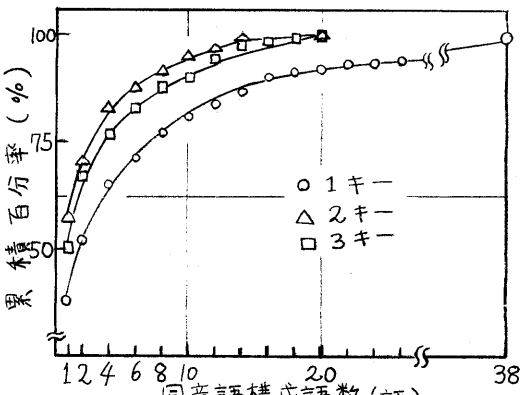



図5 頻度を考慮した同音語分布

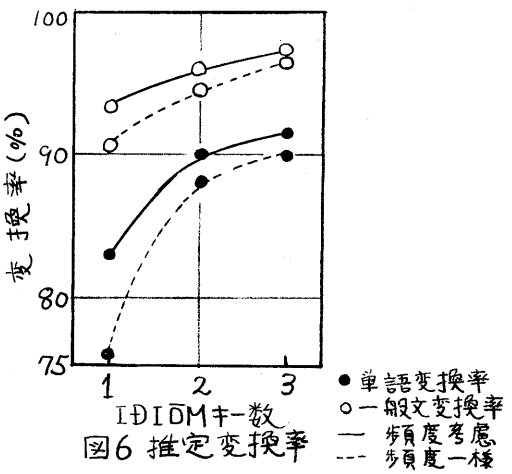



図6 推定変換率

---

*単語変換率：一意に定まる単語の出現度数に同音語中のオ1頻度の単語の出現度数を加えた度数の総出現数に対する割合

**一般文変換率　日本語文は漢字と仮名の出現比率が2：3であると仮定した変換率

## 〔6〕修正機能

修正機能として、置換、挿入、削除の各機能を用意している。置換および挿入では、同音語列挙あるいは同音字列挙による選択修正機能を有する。

(1)同音語列挙法

修正位置にカーソルを移動し、列挙キーを押下することにより、同音語を列挙表示する。該当の単語番号を指定することにより、単語単位で修正する。なお、IDIOM同音語（同音語のうちIDIOM符号も同じ語）は辞書同音語（辞書に登録されている同音語）に比較して、最大同音語構成語数が、38語から20語に減少する。（図4）

(2)同音字列挙法

同音語を列挙しても、該当同音語がない場合、漢字の読み＋IDIOM符号を入力することにより列挙された同音字（IDIOM同音字）より該当文字を選択するために用いる。

文字種3278字について、辞書同音字（辞書に登録されている同音字）と前記IDIOM符号3キーによる同音字の分布を図2に示す。IDIOM同音字は辞書同音字に比べその累積同音字数は約78%に減少し、最大同音字構成字数も132より83に減少する。

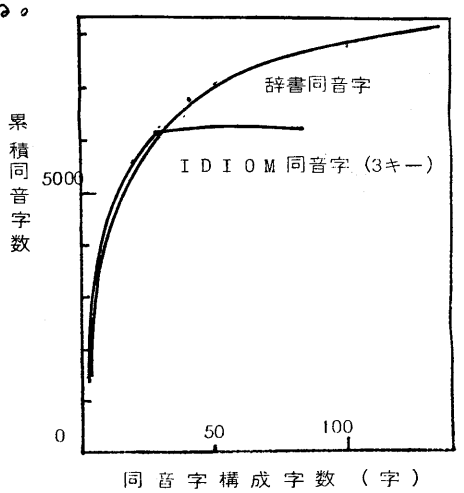



図7 同音字分布図

## 〔7〕システム構成

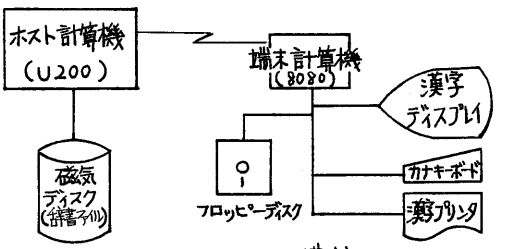
本方式の変換性能、入力操作性等を確認するため、実験システムを作成した。図8にシステム構成を示す。

(1)ハードウェア

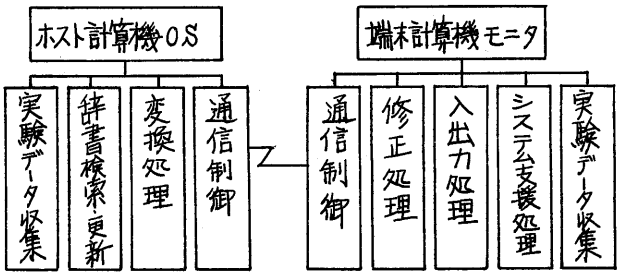
変換処理はホスト側で行ない、非同期回線でキーボード／ディスプレイと結合している。操作部はディスプレイ端末の汎用カナ鍵盤を用い、図形的特徴の入力はテンキーのキートップを変更して流用している。

(2)ソフトウェア

ホスト側においては、変換処理およびそのサポートとして、辞書ファイルの検索、更新処理、キーボード／ディスプレイ端末との通信制御の処理を行なう。この他文字の出現頻度などの実験データ収集を行なう。キーボード／ディスプレイ側ではディスプレイ、プリンタ、フロッピィディスク等の入出力処理、ホスト側との通信制御、修正処理および操作性の実験データの収集を行なう。




(a) ハードウェア構成




(b) ソフトウェア構成

図8. システム構成

## 〔8〕実験と評価

### 8－1 変換性能

(1)実験方法

入力データとして、一般文11編（天声人語8編、高校国語教科書3編）を取扱った。ただし、1編は350～400字（漢字かな混り文）で、固有名詞は対象から除いた。入力データの総文字数は4028字，IDIOM変換単位数*は906、漢字変換語数は694語、同音語を有する語数は400語である。

(2)実験結果

IDIOM符号の効果、文法処理の効果を確かめるため、表1に示す4種類の変換処理について実験を行った。

表4．変換処理種類

| 種類 | IDIOM符号数** 符号内容 | 文法処理 |
|---|---|---|
| A | 1個（漢字区切り） | 有 |
| B | 2個（縦割り、その他） | 有 |
| C | 3個（縦割り、横割り、その他） | 有 |
| D | 3個（縦割り、横割り、その他） | 無 |

以下にその結果を示す。

《変換率》 一意変換率（処理により一意に変換された割合）と正解率（第一頻度で正しく変換された語を含む割合）とを表2に示す。

表5．一意変換率と正解率

| 一意変換率(%) 処理種類 | A | B | C | D |
|---|---|---|---|---|
| IDIOM変換単位 | 67.9 | 79.5 | 82.5 | 71.3 |
| 漢字変換語 | 58.1 | 73.2 | 77.1 | 62.5 |
| 同音語を有する語 | 27.3 | 53.5 | 60.3 | 35.0 |

| 正解率(%) 処理種類 | A | B | C | D |
|---|---|---|---|---|
| IDIOM変換単位 | 87.4 | 91.9 | 92.6 | 85.1 |
| 漢字変換語 | 85.6 | 89.5 | 90.3 | 80.5 |
| 同音語を有する語 | 71.5 | 81.8 | 83.3 | 66.3 |

《誤処理》 各処理の割合および誤処理の割合（IDIOM符号2、文法処理：有の場合）を表3、表4にそれぞれ示す。ただし、表4の全体欄は、漢字変換語全体に対する誤処理の割合である。

表6．処理割合（%）

| 処理内容 | 一般語 | 名詞 | 接辞 | その他 |
|---|---|---|---|---|
| 処理割合 | 53.1 | 11.9 | 11.6 | 23.4 |

表7．誤処理の割合（%）

| 種類 処理 | 一般語 | 名詞 | 接辞 | 全体 |
|---|---|---|---|---|
| 頻度処理誤り | 4.4 | 6.5 | 17.1 | 6.6 |
| 辞書未登録誤り | 1.3 | 0.9 | 5.7 | 1.9 |
| 文法処理誤り | 1.0 | —— | 1.9 | 1.0 |
| ひらがな出力 | 0.8 | 1.9 | 1.0 | 1.0 |
| 合計 | 7.5 | 9.3 | 25.7 | 10.5 |

(3)考察および評価

① IDIOM符号1から2で変換率、特に一意変換率、が大きく向上しており IDIOM符号の効果が現われている。IDIOM符号2から3で変換率があまり変化していないのは、横割りとその他の出現頻度割合が1対3と大きく異なるためと考えられる。

② 文法処理の効果は大きい。特に漢字1字からなる語に対して有効に働いていることが、データの解析により確かめられている。

③ 文法処理誤りは少なく、新しく導入した整合文字数による同音語優先順位決定法および名詞処理は処理の簡易化の面から有効な方法といえる。

〈誤例〉 初めて（接続可）→ 始めて（接続可）
[名詞]+「て」 [下一段連用]+「て」

④ 接辞処理が最も誤り率が高い。これは、(i)接辞辞書が小さい（292語）ことと、(ii)〈例〉肺病病み→屋みのように接辞ではない語（病）を接辞（接尾語：屋）と見なし、付属部（み）を無視して処理するために誤変換される場合が若干起こること、によると考えられる。

---

* 本アルゴリズムで変換する単位で、漢字部＋（かな部）で構成されている。[6]

** 1文字用のダミー符号（□）を別に含む。

## 8－2 操作性

IDIOM形入力方式の操作性を明らかにするため、次のような操作実験を行った。

(1) 実験装置

装置は、キー入力操作を表示部とIDIOMキーをJISかな配列の外側（テンキー部）に配置したキーボードとからなり、各キーの使用回数および入力時間間隔を測定する。

(2) 実験方法

かなキーボードの操作に無経験な女性3名について、IDIOMルール（2および3キー）にもとづき社説（朝日）を入力する。この際、原稿は毎回異なるものを用い、1回を30分として行う。入力操作後、原稿とキー操作結果とを対照して、入力速度、および誤入力率等を測定する。

また、入力操作の一部をVTRで撮影し、入力動作を分析する。

(3) 実験結果

① 社説の漢字かなの比率

入力文235編（計25万字）に対し、漢字は42％（熟語分34％、一文字分8％）、かな51％、その他（記号等）7％であった。

② 一文字あたりのタッチ数

3名の総入力タッチ数は、519122タッチであり、総入力文字数250693で割ると、2.07タッチ／字となる。なお、IDIOMキーのタッチ数は総タッチ数の25％を占めている。

③ 入力速度

各オペレータは、80回（40時間）後には、ほぼ50字／分の入力速度に達した（図9）。なお、IDIOMキーをブラインドタッチできるような位置（例えば、スペースキー位置）に配置すれば、さらに入力速度の向上が期待できる。

④ 漢字認識時間

入力時間間隔およびVTRによるオペレータの頭部動作時間測定等から、IDIOMキーの認識時間は、0.1秒程度であり、かな入力操作の障害とはならない。

⑤ 誤入力率（誤タッチ数／入力タッチ数）

誤タッチ回数は、入力回数の増加に伴なって減少し、0.5％程度となる。このうちIDIOMキー誤りは$\frac{1}{3}$以下である。（図1）

⑥ IDIOMキー数の影響

入力速度および誤入力率とも、2から3キーの範囲では、ほぼ同じ傾向を示しており差はない。

以上の結果、IDIOM形入力方式は、素人の場合でも、短期間に簡単に習熟できることが明らかになった。また、かな入力速度の向上に伴なって、さらに高速入力が可能である。

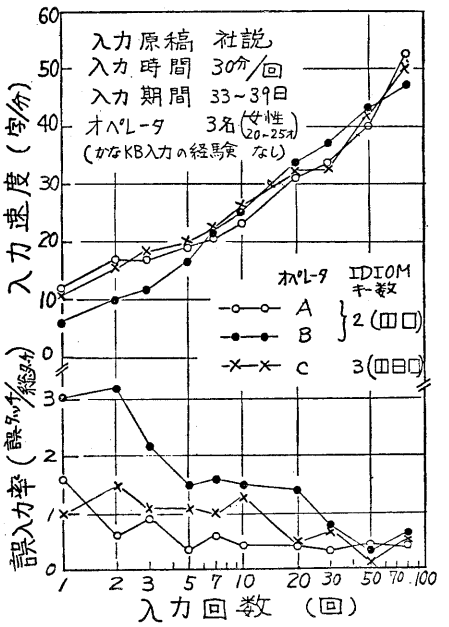



図9 IDIOM入力における習熟特性

## 〔9〕むすび

漢字の図形的情報を利用したIDIOM形かな漢字変換システムを試作し、その変換性能、入力操作性について実験評価した。その結果、次の点を明らかにした。

① 少ないIDIOM符号数（2～3）においても、同音語を一意に決定する割合は、従来方式の2倍以上となり、高い効果が得られる。

② IDIOM符号により、同音語が絞られるため、簡略形文法処理法の効果も高く，簡易な処理アルゴリズムで高い変換率が得られる。

③ カナキーボード未経験者でも、40時間の訓練で50字／分の入力速度が得られ、習熟が早い。また、IDIOM符号位置の最適化により、さらに高速入力が期待できる。

④ IDIOMキーの誤入力率は低く、かつ、漢字認識時間も短い。

今後、装置の小形化、処理アルゴリズムの最適化、編集処理機能の拡充を図り、文書処理分野等へ適用していく。

## 謝辞

本研究の遂行にあたり適切な御指導と御助言をいただいた入力装置研究室杉山室長、入出力方式研究室高野調査役に感謝いたします。また、実験評価面で協力いただいた現長岡技術科学大学の原助手、プログラム作成面で協力いただいた入出力方式研究室の壁谷主任、鈴木氏に感謝いたします。

## 文献

1) 栗原他：仮名文の漢字混り文への変換について、九州大学工学集報、NO39、（1967）
2) 牧野他：カナ漢字変換の一方法、情報処理、18－7、（1977）
3) 森他：計算機への日本語情報入力、信学会研資、EC78－23、（1978）
4) 杉山他：特徴分類（IDIOM）形日本語入力方式の検討、信学会研資、EC79－13、（1979）
5) 渡辺：漢字と図形、NHKブックス
6) 壁谷他：IDIOMにおける端末側機能、54年度信学会部門別全国大会、NO449

## 〈付録〉

□タンジュン□ナ□ヨヘンカン□□アルゴリズム□ト□シヨウ□□ヨウリヨウ□ノ□ジショ□□フアイル□デ□タカ□イ□カナ□□カンジ□□ヘンカン□□リツ□ガ□エ□ラレル IDIOM□ガタ□□ニユウリヨク□□ホウ□ニツイテ□サキ□ニ□テイアン□シタガ、ソノ□コウカ□ヲ□タシ□カメルタメ、□システム□□シサク□ヲ□オコ□ナツタ。

(a) 入力例

単純な変換アルゴリズムと小容量の辞書フアイルで高いカナ漢字変換率が得られるIDIOM形入力法について先に提案したが、その効果を確かめるため、システム試作を行なった。

(b) 変換例

付図　入力例と変換例