北海道大学工学部研究報告
第123号（昭和59年）

Bulletin of the Faculty of Engineering,
Hokkaido University. No. 123(1984)

日本語処理システムにおける前後の文字を利用する同音語の自動選択

伊藤太亮* 栃内香次 永田邦一

（昭和59年6月30日受理）

Automatic Homonym Selection Using Character Chain Matching In a Japanese Word-Processing System

Taisuke ITOH, Koji TOCHINAI and Kuniichi NAGATA

(Received June 30, 1984)

Abstract

In the Kana-Kanji translation system for the Japanese word processing, the manner in which to select homolyms is an important factor for the performance of the system.

We have developed a method for the automatic homonym selection by the use of the character chain matching.

Letters which appear just before and after a word are considered to be determined by the grammatical relation such as conjugation, the context of the sentence and the personal style of the writer, etc. Therefore, comparing a set of a word and letters just before and after it with those registered in the dictionary, homonyms can be selected as a suitable one.

In this paper we describe the construction of the homonym dictionary and selection algorithms used in the prototype system developed. Results of input experiments are also described, and it is concluded that about 50% of homonyms are automatically selected.

1. はじめに

かな漢字変換において，同音語の選択をどのように行なうかはシステムの性能を支配する重要な要素である。

我々は，同音語とその前後に出現する文字の組を登録し，これら三つ組のマッチングによって同音語を選択する方法を試みている。ある語の前後に現われる文字は，活用のような文法的関連，前後の語での意味的関連，および書き手の個人的なくせなどの総合によって決まると考えることができ，使用対象分野，個人への適応が可能で我々が研究中の研究者向き日本語入力システムに向いた方式と考えられる。

本論文では，この方式にもとづいて試作したシステムにおける具体的な同音語自動選択アルゴ

電子工学科 電子機器工学講座
*現在 王子製紙㈱

リズムおよび同音語辞書の構成について述べる。また，いくつかの文献を入力し，試作システムの性能を測定した結果について併せて報告する。

2. 前後の文字を利用する同音語の自動選択

2.1 同音語の自動選択手法

かな漢字変換方式では，漢字で表わされる語の読みをかな，あるいはローマ字で与え，これをキーとして，変換辞書を検索する。変換辞書には，読みとそれに対応する漢字の漢字符号との組が収録されており，入力した読みに一致するものがあれば，対応する漢字符号列が得られる。

一致するものがない場合は入力した読みに対応する語が辞書に登録されていないことを意味する。このような語を新出語，あるいは未登録語という。また，一致するものが2個以上あるときを同音異義語，あるいは簡単に同音語といい，その中から適切な1語を選択する操作が必要である。本論文ではこれらのうち同音語の選択処理を取扱う。

同音語の選択は，前後の文章との意味関連によって，候補語の中から一つを選び出す作業であるため，これをオペレーターが行なう場合には，一連の入力作業が中断され，実効的な入力速度が減少してしまう。

そこで，同音語の選択を自動的に行なう方法がいろいろ試みられているが，概ね以下の様な方式がある。

a）文法的な処理　自立語と付属語との接続の関係に規則性があることを利用する方式。動詞などの活用語尾が活用形で定っていることを用いる[1)]。

例：　コウシヨウナ　⟶　高尚な　○
　　交渉な，高唱な　×

b）意味論的な処理　動詞には，その前の主語，目的語に，特定の意味範囲の単語を要求する場合が多い。例えば，「行く」は，（動作主体）が（場所）へ行くとなり，助詞「が」の前には動作できる主体を表わす語がくる[1)]。

例：　センセイが行く　⟶　先生　○
　　宣誓　×

c）頻度情報による処理　入力文の対象分野によって，単語の使用分布状態に偏りがあることを利用する方法。

d）字づらによる処理　同音語とその前後に入力された，かなあるいはローマ字の組み合せによって，選択を行なうものである。字づらには，送りがな，活用語尾のような文法的な関連や，書き手の癖が表われるのを利用するものである。

以上四種のうち，a）とb）はその処理が複雑であり，品詞情報や活用法などかなり大量の基礎データが必要とされるため大型のシステムとなってしまう。またc）は文法的な根拠がないため正当性に問題がある。

またd）は入力かな文字列（あるいはローマ字列）の意味的冗長性を利用した方法であるため，入力文字の区切り方や，入力文献の種類による影響が大ではあるが，データの蓄積をシステムの使用と同時に行なえることと，選択アルゴリズムの容易性から我々は，字づらによる処理を採用した。

2.2 字づらによる同音語自動選択

我々が試作した，日本語処理システムは，研究者が自分の専門分野に関する論文等を作成する

のに適した日本語入力方式として，個々の使用者に適応した比較的小容量の変換辞書を用いてローマ字入力にて，かな漢字変換を行なう方式である[2)~4)]。

この方式では

○漢字，かな，カタカナ，などの文字種の区別は使用者の指定になる。

○変換辞書は動的に構成され，使用につれて，その収録語が使用者に適応してゆくため，小容量（2,500語）である。

同音語の発生は，漢字語のみに限られ，しかも変換辞書容量が小さいため同音語の発生率は20％と低くなっている。

そこで，このように小容量の変換辞書に登録されている語の同音語の選択に，同音語と前後の文字との組のマッチングによる方法（字づらによる方法）を用いることとした。この方式の概要は次のようになる[5)]。

1）ある文の中で読み w をもつ同音語 w_i，w_j が各々

……x_i，w_i　y_i……

……x_j　w_j　y_j……

という形で使われているものとする。ここで，x_i，y_i，x_j，y_j は空白を含む任意の1文字。

例：　……が多い……，　……は大き……

2）もし $w_i \neq w_j$ のとき，$(x_i, y_i) \neq (x_j, y_j)$ がつねに成立するなら，それにより，w_i，w_j を識別できる。現実には，漢字語の前後につく文字は種々あり，上記2）が常に成立するとはいえないが，後述するような方法によって，同音語の相当数について，いずれかであるかを推定することができる。

図1に試作システムの処理の流れを示すが，このうち，破線の中が同音語処理を行なう部分である。

ここで同音語辞書とは，各同音語について，前後にとりうる文字と，その使用頻度を記録したものをいい，詳細は，後述する。

同音語が発生した場合

1）まず，その前後の1文字 x，y を抽出する。

2）語の読み w と，x，y の3つの組によって同音語辞書を検索する。

3）特定の選択条件を満足するかどうか調べる。

4）条件を満足した場合には，その語を自動選択する。

5）条件を満足しなかった場合は，正しい語を使用者が選択して，w，x，y とともに，その語の漢字符号等を同音語辞書に登録する。

以上のような操作をくりかえし行なう。

2.3　同音語辞書の構造

各同音語について，前後にとりうる文字の組と，それぞれの頻度が登録されている。1つの同音語に対して，前後にとりうる文字の種類は語によって大幅に異なるが，試作システムは最大23種まで登録できるようにした。その他，辞書検索の効率を高めるためのポインタや辞書管理用のデータが登録されている。同音語辞書の構成とその内容例を図2に示す。

1語あたり，72×4＝288バイトをとり，その内容は，以下に示す通りである。

先頭の72バイトは，当該同音語に関する変換辞書の内容のコピーであり，「語の読み」と「漢字符号」以外の同音語の自動選択には用いない。

K：同音語番号……一つの読みに対して登録順に1，2，3…とつける。

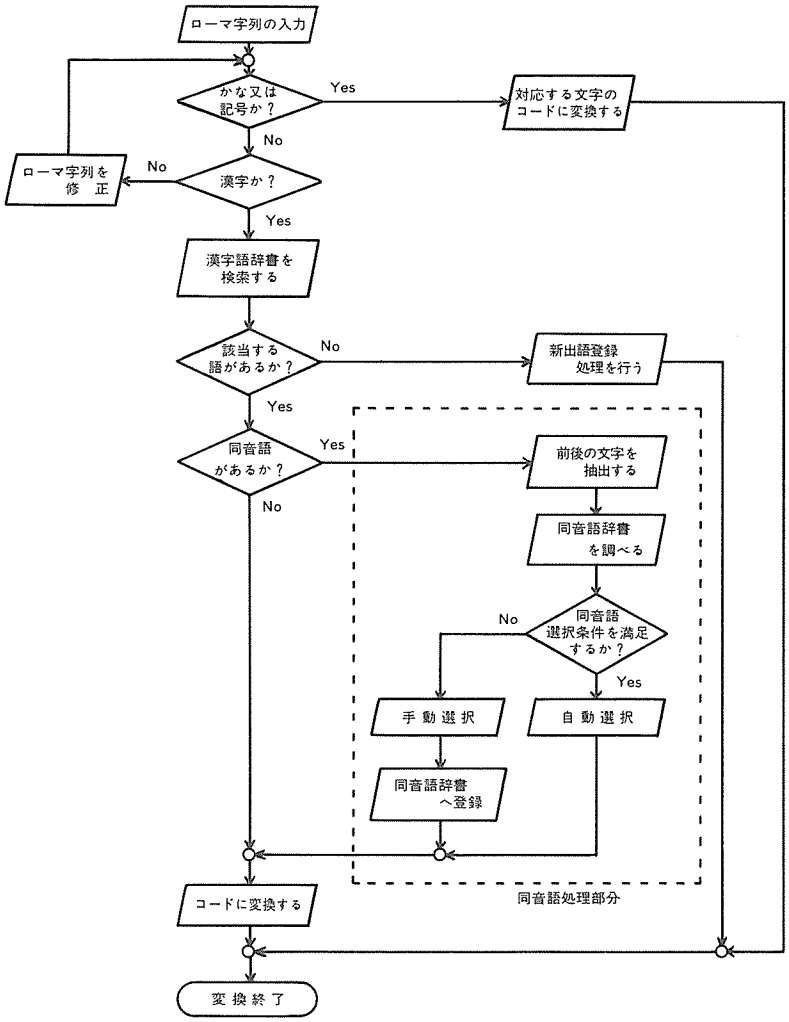


図1　処理の流れ

l：登録されている（x, y）組の個数。（1～23）

N_p：次の同音語番号のものの所在を示すポインタ。

x_i：語の前の文字

y_i：語の後の文字　　} （$i=1$～24）

n_i：当該（x_i, y_i）組の出現頻度

C_p：（x, y）組の登録管理用ポインタ

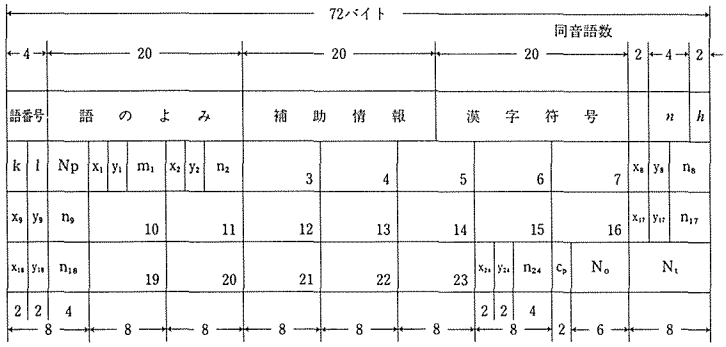
N_o：登録されている（x_i, y_i）$i=1$～23組以外のものの出現頻度の和

N_t：当該同音語の出現頻度合計

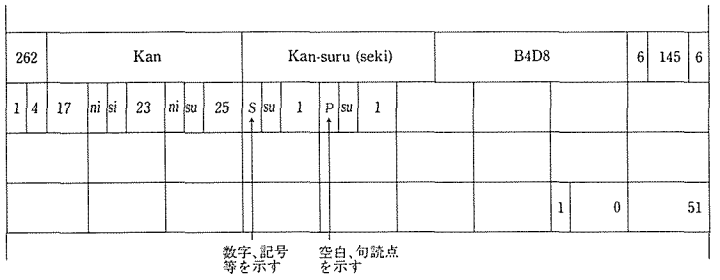
これらのうち，前後につく文字（x_i, y_i）は下記の規則に従って入れる。

1）ひらがな　　ローマ字つづりをそのまま入れる。ただし，よう音を含む場所や外来音などで3文字をこえるときは頭2文字を入れる。また，1文字のみの場合は左づめとし，2文字目は空白とする。

例：　「kya」……ky　　「a」……a＿　　（＿は空白



例：



図－2　同音語辞書の構造

図 2　同音語辞書の構造

2）漢字　　K＿とする。

3）句読点，空白，ピリオド，コンマ，疑問符，感嘆符，改行，改段，改頁，タブ　P＿とする。

4）カタカナ，英字，数字，記号　　S＿とする。

5）入力エラーにより判定不能となった場合や入力が未完の場合　　X＿とする。

また，n_{24}は特に，y_{24}＝X＿であるものの出現頻度合計であり，

x_{24}＝＿＿　　y_{24}＝X＿とした。

同音語辞書の内容は，登録順に配列される。

2．4　同音語選択手順

変換辞書を検索して，同音語が出現した時，以下の手順により選択処理を行なう。

1）前後にくる文字（x, y）を抽出する。

x, yは２．３で述べた同音語辞書中の文字データ（x_i, y_i）と同様な規則で符号化する。

2）当該同音語について同音語辞書を $k=1$ から順に検索し，（x, y）組の同一のものがあるか否かを検索する。N_pを使用して検索効率を高めている。

3）いずれか１個の同音語についてのみ

$(x_i, y_i)=(x, y)$　となる

（x_i, y_i）組が存在する場合，以下の処理を行なう。

① $N_t \geqq N$　のとき，これを該当する同音語として選択する。

② $N_t < N$　のとき，手動選択処理を行なう。　6)参照

4） 2個以上の同音語について

$(x_i, y_i) = (x_j, y_j) = \cdots\cdots = (x, y)$ となる

(x_i, y_i)組，(x_j, y_j)組……が存在する場合，以下の処理を行なう。

① $N_i/N_j \geqq M$であり，$N_{ti} > N$のとき，N_iの方を該当する同音語として選択する。

②①の条件が満たされないとき，手動選択処理を行なう。 6)参照

5）いずれの同音語にも，$(x_i, y_i) = (x, y)$ となる (x_i, y_i) 組が存在しない場合は，手動選択処理を行なう。

6）手動選択処理

同音語の候補語を画面に表示し，使用者に選択を促す。

手動選択によって選ばれた同音語の情報と前後の文字の組を同音語辞書に登録する。（次頁参照）

注：N，Mはパラメータ（現状 N＝10，M＝2）

以上のことから分かるように，今回の方法は，発生した同音語の全数を自動選択するものではなく，使用につれて同音語とその前後の文字の組を学習していくことにより，以前と同じような使われ方をした語を自動的に選択して，手動選択回数の低減を計るものである。

2.5 同音語辞書への登録手順

同音語とその前後にとりうる文字の情報が記録されている同音語辞書への新しい情報の登録は同音語の手動選択処理が行なわれた時に実行される。

つまり，この登録処理は，同音語の使用状態を前後の文字との関係および頻度によって学習する機態とみなすこともできる。

具体的には，以下の手順による。ただし，(x, y)は，同音語の前後の1文字とし，同音語辞書制御用の変数はそれぞれ，

KMAX：同音辞書の全容量（現状は500語），および

KCNT：同音辞書の実登録語数，である。

1）選択した同音語が，すでに同音語辞書に登録されている場合，辞書中のその語の項目について以下の操作を行なう。

①y＝X＿であれば

$n_{24} + 1 \longrightarrow n_{24}$

$N_t + 1 \longrightarrow N_t$ とする。

② $l < 23$ ならば，$l + 1 \longrightarrow l$ とし，l番目の欄に (x, y) 組を書込む。

また $l \longrightarrow n_l$，$N_t + 1 \longrightarrow N_t$とする。

③ $l = 23$ ならば $C_P + 1 \longrightarrow C_P$とし，$C_P$番目の欄から順に$N_l$を調べ，$N_l$が最少のものを見出す。

ついで $N_o + n_l \longrightarrow N_o$とし，

(x, y)組を，その場所に書込み，

$1 \longrightarrow n_l$，$N_t + 1 \longrightarrow N_t$とする。

2）選択した同音語が，同音語辞書にない新たな語である場合，以下の手順により項目を追加する。

① KCNT＜KMAX ならば，KCNT＋1 ⟶ KCNT

②新出同音語を KCNT の示す場所に書込む。

③ (x, y) 組を (x_1, y_1) に書込み，$n_1 = 1$とする。ただし，y＝X＿のときは，(x_1, y_1) へ

の書込みは行なわず，$n_{24}=1$とする。

④初期設定を行なう。

$l=C_p=N_t=1$，　$N_p=N_o=0$，　$x_{24}=$＿＿，　$y_{24}=X$＿

⑤辞書をKCNTから逆向きに検索し，最初に出現するこの語の同音語を見出す。その同音語番号をK'，N_pポインタをN_p'とする。

⑥$K'+1$を新しい同音語のK欄に書込む。

⑦KCNTをN_p'に書込む。（それまで，$N_p'=0$である）

⑧先頭までさかのぼっても見出せないときは，これが最初の同音語である。そこで，

$K=1$　とする。

⑨KCNT＝KMAXのときは，同音語辞書の拡張を検討する。

3. 試作システムの性能

すでに作成され稼動している日本語入力システムに上記同音語辞書と関連機能を組込み，試験使用を行なった。

3.1 入力資料

入力文として，表1に示す情報工学に関する4つの文献を用いた。漢字語総数は21,937語で，そのうち同音語は3,764語であった。

内容的には，資料1と2がデータベースと情報検索ということで近い関係にあり，資料3と4は日本語処理システムに関するもので近い関係にある。

表1　入力資料一覧

No.	著者	書名	出版社など
1	穂鷹良介	データベース要論	共立
2	中原啓一	情報検索	電子通信学会
3	斎藤康	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	北大工学部修論
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	北大工学部修論

3.2 同音語自動選択特性

資料1，2では，1章ごと，資料3，4では，2章ごとに分割して入力し，同音語数，自動選択数を集計した。

実験結果を表2に示す。出現同音語のべ数3,764語のうち47%にあたる1,756語が自動選択され，誤選択数はのべ14回であった。ただし，ここでの誤選択数とは，自動選択した結果が誤っているものだけを指し，人為的な選択ミスは含んでいない。

また，図3，図4に自動選択率と同音語辞書の収容語数の推移を示す。ここで横軸は，出現同音語のべ数である。また，自動選択率は同音語100語ごとの自動選択数として算出した。なお，図4から分かるように，同音語辞書は空の状態から実験を開始した。

表2 実 験 結 果

文献	章	文字総数	漢字語数	同音語数	自動選択数	誤選択数
1	1	3,162	430	76	0	0
	2	4,404	531	82	2	0
	3	7,509	1,006	135	5	0
	4	25,499	2,887	517	187	1
	5	22,751	2,975	463	201	0
	小 計	63,325	7,829	1,273	395	1
2	1	8,898	1,501	299	135	0
	2	9,059	1,635	222	122	0
	3	34,229	4,960	629	390	0
	小 計	52,186	8,096	1,150	647	0
3	1，2	5,522	1,155	222	86	2
	3，4	10,805	2,323	463	238	2
	5，6	3,305	691	146	78	0
	小 計	19,632	4,169	831	402	4
4	1，2	5,328	1,066	212	126	4
	3，4	5,209	634	267	164	4
	5	656	143	31	22	1
	小 計	11,193	1,843	510	312	9
合 計		146,336	21,937	3,764	1,756	14

これらの結果から，以下のことが考察される。

1）今回の実験は，同音語辞書が空の状態から始めたことを考えると，使用が進み同音語辞書にある程度の語が蓄積された後においても，約50%の自動選択率は十分に期待できると思われる。

2）図3，図4において入力資料が変わる点において，自動選択率が低下し，同音語辞書収容語数が急増する傾向が見られる。これは，資料の内容の変化および著者の使用語種の変化によるものと考えることができる。

資料2と3の間の変動が大きいのは，資料の内容のジャンルが変わったためと考えられる。

3）自動選択された語を品詞別に調べてみると，表3のようになり，名詞が最も多い。

また，品詞別の特徴としては，動詞，形容詞は，送りがなや語尾変化による判定がなされたものが多く，名詞は，入力資料に関係の深い語，つまり出現頻度の大きい語が多く選択されている。

4）誤選択の全数を表4に示す。その原因として考えられものを以下に示す。

①今回用いた文字区分が，漢字はK＿，カタカナはS＿などと大まかであったために，日本語の冗長性を損うこととなり，判定不能となった。（表4の1，4，5など）

②今回の方法では，前後1文字づつしか取り出さないため，それだけでは判定できないものがあった。

例 …て表わ… …て現わ… 表4の2，3

③一度誤選択が発生すると，それを人為的な手段で訂正することを行なわないために，同じ誤りをくり返す。（表4の5，11，14など）

実際にこの日本語入力システムを使用する場合には，誤選択の訂正を随時行なうために，誤選

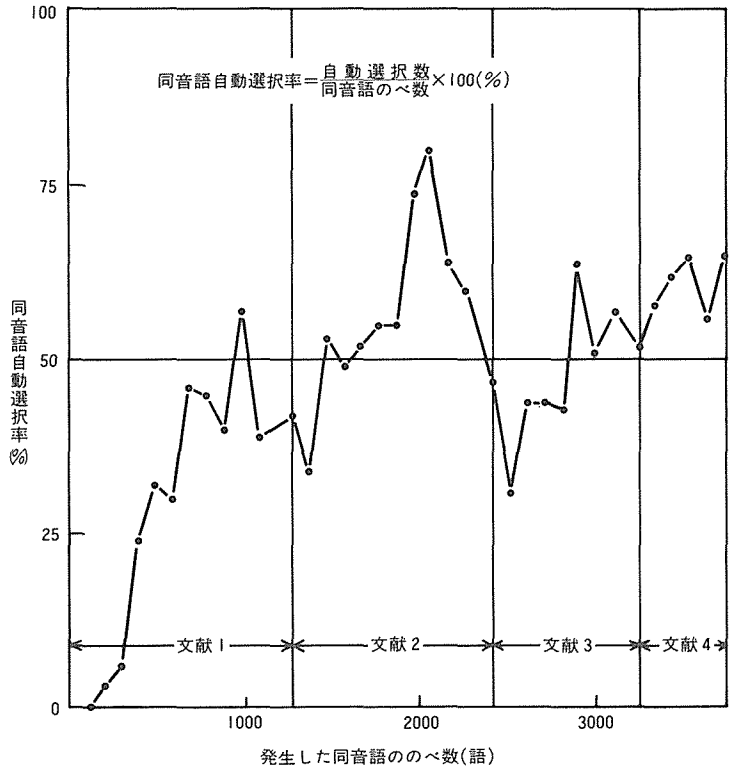


図3　同音語自動選択率の推移

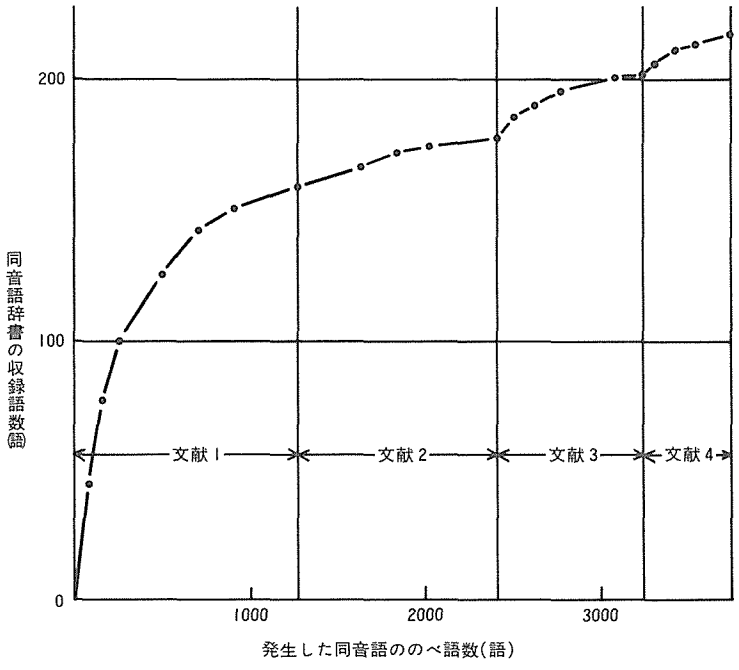


図4　同音語辞書収容語数の推移

表3 自動選択同音語の品詞分類

文献	章	自動選択数	名詞	動詞	形容詞	副詞
1	1	0	0	0	0	0
	2	2	2	0	0	0
	3	5	3	2	0	0
	4	187	77	104	6	0
	5	201	82	106	10	3
	小計	395	164	212	16	3
2	1	135	102	29	2	2
	2	122	92	26	3	1
	3	390	181	190	13	6
	小計	647	375	245	18	9
3	1, 2	86	28	35	16	7
	3, 4	238	142	68	16	12
	5, 6	78	54	14	6	4
	小計	402	224	117	38	23
4	1, 2	126	77	33	14	2
	3, 4	164	77	62	20	5
	5	22	10	7	5	0
	小計	312	164	102	39	7
合計		1,756	927	676	111	42

表4 誤選択例一覧

	誤選択部分	正解
1	，使用自身	仕様
2	1文字で表われた	現
3	初めて表われる	現
4	19文字文の	分
5	出力字の	時
6	多段シフト形入力形式	型
7	(タブレット)形入力形式	型
8	語の仕様について	使用
9	個人仕様の実験	使用
10	条件化における	下
11	検索字の見出し	時
12	求めた法が	方
13	話の仕様の調査	使用
14	登録字の負担	時

択の発生頻度は今回の実験よりも低下すると考えられる。

5）詳細のデータは示していないが，同音語辞書収容語1語当りの前後にとりうる文字の組の実登録個数は，名詞以外の場合，ほとんどが10種以下と少ない。また，名詞では，個数は多いが，1種当りの出現頻度が少ない。ゆえに，前後にとりうる文字の組の1語当りの登録個数を今回の23種より多少少なくしても，同様の自動選択率が得られると考えられる。

4. おわりに

日本語処理システムにおいて，前後の文字を利用する同音語の自動選択方式を開発し，この機能を組込んだシステムの特性を調査した。その結果を以下にまとめる。

1）同音語の選択の50%以上を自動的に行うことが可能となった。

2）今回のシステムは，対象分野や使用者のくせなどに適応する。それゆえ，使用分野が変化した際には，一時的に自動選択率が低下する。

3）今回の方式は，字づらによる選択であるが，実験結果から逆に判断すると，動詞，形容詞の語では，送りがなや語尾変化による判断をする文法的な選択方式とみなせるし，名詞の語の場合は，使用頻度による選択とみなされるものが多い。

4）誤選択は，自動選択された同音語数の1%以下であるが，前後の文字の分類が大まかであったことが主な原因である。

5）前後1文字だけでは誤選択されてしまう可能性のあるものもあるので，抽出文字の数や位置の検討や文字の分類の細分化を行なう必要がある。

6）同音語辞書の各語ごとの前後の文字の登録数（現状23組）をもう少し少なくできる可能性がある。（自動選択率への影響が小さい）

本システムは，現在，学会の予稿や論文，その他の資料等の作成に使用され，十分実用になっている。さらに，今回の方法を他の文字種（カタカナやひらがな，記号）にも応用し，べた書き文入力によるかな漢字変換システムについて検討する予定である。

謝　辞

本研究を行なうにあたり，種々有益な御意見，御討論をいただいた研究室の各位に感謝いたします。

参考文献

1）牧野　寛，木澤　誠：情報処理学会論文誌，Vol.22，No. 1，pp.59～67 (1981)
2）栃内香次，斎藤　康：情報処理学会論文誌，Vol.24，No. 2，pp.209～213（1983）
3）斎藤　康，岡沢好高，栃内香次，永田邦一：北大工学部研究報告，No.108，pp.43～52（1982）
4）栃内香次，永田邦一：電子通信学会技術研究報告，EC82-20 (1982)
5）伊藤太亮，栃内香次，永田邦一：情報処理学会第27回全国大会，2H-6（1983）