

前後接続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†

柄 内 香 次^{††} 伊 藤 太 亮^{††} 鈴 木 康 広^{††}

かな漢字変換システムにおいて、出現する多数の同音語がなるべく自動的に選択されることが望まれる。本論文は、語の接続関係を利用して同音語の自動選択を行う一手法を提案し、さらにそれをかな漢字変換システムに組み込んで実験を行った結果について述べたものである。近接の語との組合せによって同音語を選択する方式はいくつか報告されているが、本方式では同音語の前後に接続する各1文字の組を利用する。すなわち、あらかじめいくつかの文書から同音語とその前後各1文字からなる三つ組を抽出してその出現頻度とともに辞書に登録しておき、新しい文書を入力する際、入力文中に出現した同音語について同様な三つ組を抽出し、辞書中の三つ組と比較する。辞書中に同一の三つ組が存在し、かつその出現頻度があらかじめ設定した基準値をこえていればその三つ組を含む同音語を選択する。複数の同音語に同一の三つ組が存在する場合も同様に各々の出現頻度を比較し、その比が設定値をこえていれば多い方を選択する。基準を満たさない場合は手動で選択し、その結果を辞書に登録して使用者への適応を行う。この機能を組み込んだ実験システムを試作し、情報処理その他の論文等10種の資料を用いて入力実験を行った。その結果、辞書に約400語の同音語が登録された状態で入力同音語の70%強を自動選択することができた。また、選択率をさらに増大させる方策を見いだすことができた。

1. はじめに

かな漢字変換システムの性能を表す指標として、

- (1) 変換辞書に未登録の語の出現率、
 - (2) 同意語の出現率、および
 - (3) 誤変換の発生率、
- の三つがある¹⁾。

このうち(1)は、辞書の容量を大きくすれば原理的にはいくらでも低下させることができる。また、先に報告したように、処理対象がある専門分野の学术论文などに限定される場合は、辞書に対象および個々の使用者への適応機能をもたせることにより、2~3,000語程度の小容量の辞書でも実用上満足できる値におさえることができる^{2), 3)}。また(3)も、誤変換は変換されるべき正しい語が辞書中に存在しないために発生するものであるから、(1)と同様である。

これに対し、(2)は同音の漢字が多数存在するという日本語表記の性質そのものに由来しており、減少させることができない。しかも、辞書の容量を大きくすればその中に含まれる同音語も増加するから、上記(1)、(3)を減少させる手段として辞書の容量を大きくすることは(2)に対しては逆効果となる。

このように、かな漢字変換システムでは同音語の出

現は避けられない。それゆえ、出現した同音語の中から特定の1語を選択する操作は可能な限り変換システム内で自動的に行うべきである。この問題に対し、次に示す二つの手法が考えられる。

(1) 語の出現頻度による自動選択

一般に、一つの文書の中である1組の同音語の各々が等頻度で出現することは少なく、いずれか1語のみが高頻度で現れることが多い。そして、このことは処理対象分野および使用者が限定されればさらに著しい。そこで、一つの文書中である同音語が何回か出現するとき、最初だけ使用者が手動で選択し、以後はその選択結果に従って自動選択することが考えられる。先に報告したシステムではこれを同音語選択の固定化とよんでおり、この機能を用いることによって自動選択回数を出現した同音語総数の約1/2に減少させることができた³⁾。ただし、この方法では手動選択回数は最良でも同音語の語種数より少なくはならない。

(2) 語の間の関係を利用する自動選択

ある語とその近くに現れる語との間の関係を利用して同音語を選択することが考えられる。たとえば、各々の語に意味属性を付加して辞書に登録しておき、語の間の意味関連を利用して選択する方法が報告されている⁴⁾。この方式では各語の意味属性をあらかじめ辞書に登録しておく必要があり、意味属性の分類基準の確立や辞書作成にかかる手間などの問題点がある。これに対し、意味による分類を行わず、語と語の接続関係によって同音語選択を行う方法が考えられ、同音

† Kana-Kanji Translation System with Automatic Homonym Selection Using Character Chain Matching by KOJI TOCHINAI, TAISUKE ITOH and YASUHIRO SUZUKI (Department of Electronic Engineering, Faculty of Engineering, Hokkaido University).

†† 北海道大学工学部電子工学科

語とその直後の語との接続関係を利用して同音語のべ語数の68.8%を選択できたことが報告されている⁵⁾。

本論文では、語の接続情報を利用した同音語選択の別な方式として、同音語とその直前、直後の各1文字との接続関係を用いる方法を提案する。これは、同音語とその前後の文字との組を辞書に登録しておき、文中に出現した同音語について同様な組を抽出してこの辞書と照合することにより選択を行うものである。ここで、同音語の辞書はあらかじめ作成されている必要はなく、かな漢字変換システムを稼働させて文書作成を行う際に、出現した同音語について手動による選択を行いながら自動的に上記の情報を収集して辞書を作り上げてゆき、十分な選択情報が蓄積された同音語について自動選択する方法を採用することができる。

このような方式による実験システムを作成し、情報処理分野その他の学術文献を用いて実験を行った結果、入力同音語のべ語数4,000語以上でその約60%を、8,000語以上では70%強を自動選択することができた。本論文では、以下、この方式の原理、実験システムの構成ならびに実験結果について報告する。

2. 文字接続を利用する同音語選択

かな（またはローマ字）で書かれた文の中で、読みが w である同音漢字語 W_1, W_2 が $\dots awb \dots$ という形で出現しているとする。ここで a, b は記号、数字、空白等を含む任意の文字である。また、文はべた書きではなく、漢字語にはそれを識別する記号がつけられているものとする。さらに、 W_1, W_2 のこれまでの出現頻度が、その前後に接続する文字との三つ組、 $a_i W_1 b_i, a_j W_2 b_j$ ごとに記録された表1のような辞書（同音語辞書）があるものとする。

いま、表1で $m_1 \geq M n_1, n_2 \geq M m_2$ であるとする。これは、 W_1, W_2 のこれまでの出現に際して、三つ組 $a_1 W_1 b_1$ は $a_1 W_2 b_1$ の M 倍以上、逆に三つ組 $a_2 W_2 b_2$ は $a_2 W_1 b_2$ の M 倍以上の頻度で出現したことを意味する。そこで、新たに出現した三つ組 awb につ

表1 同音語一前後接続文字三つ組
Table 1 Triplets of homonym and connected characters.

W_1		W_2	
三つ組	頻度	三つ組	頻度
$a_1 W_1 b_1$	m_1	$a_1 W_2 b_1$	n_1
$a_2 W_1 b_2$	m_2	$a_2 W_2 b_2$	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$

いて、

A : それが $a_1 w b_1$ ならば $w = W_1$, また

B : それが $a_2 w b_2$ ならば $w = W_2$,

と推定することができる。以下に二、三の例を示す。

例1 読み…「ようい」

～は ようい である → ～は容易である

～を ようい する → ～を用意する

例2 読み…「いこう」

～の いこう により → ～の移行により

それ いこう に～ → それ以降に

例3 読み…「かん」

～に かん して → ～に関して

～K かん の～ → ～K間の～

(Kは漢字を示す記号とする)

例3の「関」は、活用語尾をつけた「関し」あるいは「関して」を辞書に登録しておけば、「間」とは同音語にならない。しかし、実験システムでは漢字で表記される部分のみを辞書に登録する原則なので、「関」と「間」は同音語となる。なお、漢字部分の読みがかな1字の活用語（知る、見る、書く、など）については活用語尾1字をつけて登録することになっている³⁾。

上述の推定の確実性は、三つ組 $a_i W_1 b_i$ と $a_i W_2 b_i$ の出現頻度の比 m_i/n_i 、および W_1, W_2 の総出現頻度に影響されると考えられる。しかし、これらの出現の様子は著者、あるいは対象分野によって変動する。それゆえ、文書入力を行う過程で上記の三つ組を収集、蓄積できるようにし、出現頻度が小さいうちは手動で選択してその結果を蓄積し、ある閾値をこえた段階で自動選択する方式を採用すればよいことになる。

3. 実験システム

3.1 システム構成

先に報告した適応変換辞書を用いたかな漢字変換システム¹⁾に、前章で述べた同音語選択機能と同音語辞書を付加して実験システムを構築した。図1にシステム構成の概要を示す。なお、このシステムはPL/Iで書かれ、使用計算機は北大大型計算機センターのHITAC M-280Hである。

3.2 同音語辞書

図2に同音語辞書の構造を示す。図に示すように、同音語辞書は1語あたり288バイトからなる。はじめの72バイトはその同音語が漢字語辞書に登録されている形のまゝの写しである。残り216バイトのうち、192バイトをその同音語に関する前後の文字接続情報

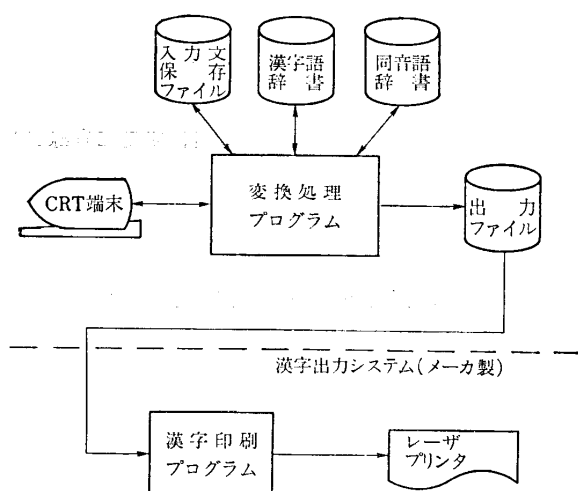


図 1 実験システムの構成

Fig. 1 Schematic of the experimental system.

にあてている。すなわち、その同音語がこれまでに出現した際にその前後に接続していた文字 a 、 b の組をその頻度とともに記録している。前後接続文字の組は 24 組まで登録することができ、これをこえた場合は出現頻度の小さい組を削除する。その際、削除された組の出現頻度合計を辞書の N_0 欄に記録する。

同音語の前後に接続する文字には漢字、ひらがな、カタカナ、数字、記号などあらゆる種類がある。同音語辞書の a 、 b 欄は各々 2 バイトの長さを持ち、これらの情報を次のような形式で記録している。

(1) ひらがなはそのローマ字つづりをそのまま記

録する。母音など、ローマ字 1 字の場合は左づめとし、また拗音の場合は、たとえば kya (きゃ) ならば ky のように最初の 2 字を記録する。

(2) ひらがな以外の文字に対しては、以下に示す字種記号を左づめで記録する。

- イ) 漢字…K
- ロ) 数字…N
- ハ) 句読点、空白…P
- ニ) カタカナ、英字、記号…S

(3) 入力文字列の先頭、末尾など前後いずれかの接続文字がない場合は、左づめで「X」を記録する。

同音語は出現順にこの辞書に書き込まれている。同音語選択処理を行う際は同一の読みを持つ数個の同音語の情報がすべて必要になるので、ポインタ N_p により、これらの間を連結している。

工学諸分野の学術論文において、論文 1 篇あたりの漢字語総数は 1,000~2,000 語で、その 20% 前後が同音語をもち、同音語の語種はこの約 1/2、すなわち漢字語総数の 10% 程度である³⁾。それゆえ、同音語辞書への新出同音語の累積速度はあまり大きくないと判断される。そこで、本実験システムでは同音語辞書の容量を 500 語とし、漢字語辞書で行っているような使用頻度、履歴による入換え⁶⁾は行っていない。

3.3 同音語選択アルゴリズム

実験システムはローマ字入力を採用しており、大文字、小文字によって漢字、かなの字種指定を行ってい

72バイト																	同音語数
4				20				20				20					
語番号		語 の 読 み						補 助 情 報				漢 字 符 号				n	h
k	l	N_p	$a_1 b_1 $	n_1	$a_2 b_2 $	n_2	$ $	3	4								8
9			10														17
18			19					$ $	22	$ $	23	$a_{24} b_{24} $	n_{24}	c_p	N_0	N_t	
8			8			---											

k : 何番目の同音語であることを示す番号

l : (a, b) 組の登録個数 (1~23)

N_p : 次の番号の同音語を指すポインタ

a_i, b_i : この同音語の前後に各々接続している文字

n_i : 当該 (a_i, b_i) 組の出現回数

ただし、 (a_{24}, b_{24}) は前が空白、後が不明という特別な組にあてられる。

c_p : (a_i, b_i) 組を登録するための管理用情報

N_0 : オーバフローして登録されていない (a, b) 組の個数

N_t : 当該同音語の出現頻度合計

なお、最初の 72 バイトは当該の語が漢字語辞書に登録されている形のままの写し。

図 2 同音語辞書の構造

Fig. 2 Structure of the homonym dictionary.

る²⁾。すなわち、先頭が大文字で、次に大文字、数字、記号、制御用文字 (C, X, L) または空白が現れるまでの文字列が漢字で書かれた語 (漢字語) である。出現した漢字語について漢字語辞書を検索することによって同音語の有無が判明し、ない場合は直ちにその語に対する漢字符号列が得られて変換が終了する。同音語がある場合は、以下に述べるアルゴリズムにより自動選択処理が行われる。

以下、出現した漢字語 w について 2 種の同音語, W_1, W_2 が存在し、これまで出現した際の前後接続文字の組とその出現頻度等とともに前述の形式で同音語辞書に登録されているものとする。また, W_1, W_2 各々の全出現頻度を N_{i1}, N_{i2} とし、同様に登録されているものとする。 w に対し、以下の手順により同音語の選択処理が行われる。

(1) 入力文中から w とその前後に接続している文字 a, b との三つ組 awb を抽出する。 a, b がひらがなの場合はそのまま以後の処理を行うが、漢字語の一部やその他の数字、記号等である場合は 3.1 節で述べた字種記号に変換する。以下、二、三の例を示す。

例 1 ...niKan_su... →

ni	Kan	su
----	-----	----

例 2 ...haSenkeiKairo... →

ha	Senkei	K
----	--------	---

例 3 ...3Syou_deha... →

N	Syou	de
---	------	----

(2) 同音語辞書中の W_1, W_2 を検索し、入力と同一の三つ組があるかどうかを調べる。

(2 a) もし、いずれか 1 語、たとえば W_1 のみに aW_1b なる三つ組が存在し、かつ $N_{i1} \geq N$ ならば、 $w = W_1$ とする。 $N_{i1} < N$ のときは自動選択を行わず、手動選択に移行する。ここで N は自動選択を行うか否かを判定する閾値で、外部から設定可能である。

(2 b) 三つ組 aW_1b, aW_2b がともに存在する場合、各々の出現頻度を n_1, n_2 とし、

① $n_1/n_2 \geq M$, かつ $N_{i1} \geq N$ ならば $w = W_1$,

② $n_2/n_1 \geq M$, かつ $N_{i2} \geq N$ ならば $w = W_2$,

とする。ここで M は前記 N と同じく、自動選択を行うか否かを判定する閾値である。

③ これらの条件が満足されない場合は自動選択を行わず、手動選択に移行する。

(2 c) 入力と同一の三つ組 aW_1b, aW_2b が辞書中に存在しない場合はそのまま手動選択に移行する。

(3) 手動選択が行われたとき、以下により同音語辞書を書き換える。以後、 W_1 が選択されたものとす

る。

① W_1 の全出現頻度 N_{i1} について

$N_{i1} + 1 \rightarrow N_{i1}$ とする。

② 三つ組 aW_1b が同音語辞書にすでに登録されている場合は、この組の出現頻度 n_i について

$n_i + 1 \rightarrow n_i$ とする。

③ 三つ組 aW_1b が未登録の場合は、

(i) 三つ組登録欄 (1~24) にあきがあればそこに登録する。

(ii) あきがない場合は既登録の三つ組の中で出現頻度最小のものをいだし削除し、そこに登録する。また、削除された三つ組の出現頻度を N_0 に加える。

以上のアルゴリズムにより、

(1) 語の出現頻度総数が N 回以上であり、

(2) 三つ組 aW_1b が aW_2b の M 倍以上出現している、

という二つの条件が満足される同音語 W_1 が自動選択される。条件が満足されない場合は手動で選択されるが、同一の三つ組がくり返し出現するとやがて条件が満足されて自動選択に移行する。

閾値 N, M を小さく設定すれば、三つ組の出現頻度や 2 種の三つ組間の頻度の比が小さくても自動選択が行われ、自動選択率が増加する。反面、三つ組の出現傾向が安定しないうちに自動選択が行われ、誤選択が増加する可能性がある。文書中の同音語の現れ方は使用者、分野によって異なるので、 N, M は使用者ごとに設定できるようにすべきである。

4. 自動選択実験

4.1 実験資料および方法

複数の分野の文献を用い、自動選択率、誤選択率を測定する実験を行った。実験に用いた資料は表 2 に示す 7 種で、資料 6 (機械工学) 以外は情報処理関連の文献である。全資料を通じてのべ 9,767 語の同音語が出現しており、その語種は 402 語である。

これを実験システムに順に入力し、自動選択率および誤選択率の推移を求めた。表 3 に結果を示す。ここで、実験は同音語辞書が空の状態を開始し、入力の累積につれて出現した同音語を順次登録した。

4.2 実験結果

4.2.1 自動選択率の推移

図 3 に、入力同音語の累積に対する自動選択率の推移を示す。ここで横軸は出現した同音語ののべ語数で

表 2 実験に使用した資料
Table 2 Collected experimental data.

No.	内 容	漢字語総数	同音語数
1	穂鷹良介: データベース要論 (共立出版), pp. 1-108	7,829	1,273
2	中原啓一: 情報検索 (電子通信学会), pp. 1-91	8,096	1,150
3	斉藤 康: 北大工学部修士論文 (1982), [かな漢字変換]	4,169	831
4	岡沢好高: " (1983), ["]	2,362	510
5	橋津正晶: " (1976), [性能評価]	4,202	872
6	機械工学に関する論文 4 篇 ^{*1)}	5,011	843
7	情報処理に関する論文 3 篇 ^{*2)}	4,670	1,023
8	" 3 篇 ^{*3)}	4,279	1,062
9	" 4 篇 ^{*4)}	5,307	1,133
10	" 4 篇 ^{*5)}	5,673	1,070

- ^{*1)} 1) 有江, 木谷, 他: 北大工学部研究報告, No. 106, p. 1 (1981)
 2) 園田, 谷口, 他: " No. 106, p. 9 (1981)
 3) " : " No. 106, p. 21 (1981)
 4) 飯田, 古川 : " No. 107, p. 33 (1982)
- ^{*2)} 1) 前島, 桂, 他: 情報処理学会論文誌, Vol. 23, p. 16 (1982-1)
 2) 山本, 中崎, 他: " Vol. 23, p. 58 (1982-1)
 3) 松山, 三浦, 他: " Vol. 23, p. 142 (1982-2)
- ^{*3)} 1) 木村 : " Vol. 23, p. 162 (1982-2)
 2) 有田 : " Vol. 23, p. 260 (1982-3)
 3) 田村, 坂根, 他: " Vol. 23, p. 321 (1982-3)
- ^{*4)} 1) 高藤, 小林, 他: " Vol. 23, p. 333 (1982-4)
 2) 長岡, 中村, 他: " Vol. 23, p. 358 (1982-4)
 3) 酒井, 落水 : " Vol. 23, p. 487 (1982-5)
 4) 中所 : " Vol. 23, p. 545 (1982-5)
- ^{*5)} 1) 吉住, 堀越 : " Vol. 23, p. 591 (1982-6)
 2) 寺田, 関, 他: " Vol. 23, p. 707 (1982-6)
 3) 有澤 : " Vol. 23, p. 267 (1982-3)
 4) 吉村, 日高, 他: " Vol. 24, p. 40 (1983-1)

表 3 自動選択率および誤選択率
Table 3 Automatic selection rate and erroneous selection rate.

資料 No.	漢字語 総 数	同音語数	自動選択 同音語数	自 動 選択率 (%)	誤 選 択 同音語数	誤選択率 (%)
1	7,829	1,273	395	31.0	1	0.08
2	8,096	1,150	647	56.3	0	0
3	4,169	831	402	48.4	4	0.48
4	2,362	510	312	61.2	9	1.76
5	4,202	872	536	61.5	20	2.29
6	5,011	843	508	60.3	26	3.08
7	4,670	1,023	662	64.7	8	0.78
8	4,279	1,062	628	59.1	45	4.24
9	5,307	1,133	832	73.4	12	1.06
10	5,673	1,070	771	72.1	17	1.59
計	51,598	9,767	5,693		142	

あり, 縦軸はそれをほぼ 200 語ごとに区切った各区間ごとの自動選択率である*。

自動選択を行うか否かの閾値は, 以下の考察にもと

* 入力文献 1 篇ごとに区切って行ったので, 各資料末尾の区間は一般には 200 語にならない。

づき, $N=10$, $M=2$ に設定した。

1) 資料中の漢字語総数は 50,000 語で, 工学系学会雑誌論文 40 篇程度相当である。したがって $N=10$ の語で平均 4 篇に 1 回の出現率であり, これより少頻度の同音語は頻度比較による自動選択の対象とするには出現傾向の安定性が十分でないと考えられる。

2) 同音語の出現傾向には個々の文献ごとに偏りがあり, 最初に出現した際の選択に固定する単純な方法でも誤選択はそれほど多くはない³⁾。そこで, 三つ組 aW_{1b} と aW_{2b} の頻度比較によって選択を行う本方式では, 頻度比 M が 2 以上という自動選択基準で十分な精度が得られると考えられる。

図 3 からわかるように, 自動選択率は全資料を通じて滑らかに推移し, 資料の境界でも大幅な変動は見られない。むしろ資料 2 におけるように同一資料中での区間ごとの変動の方が大きい場合もある。したがってこの実験では資料のちがいによる同音語出現傾向の明らかな差異は認められないといえる。

前述のように, 実験は同音語辞書が空の状態を開始

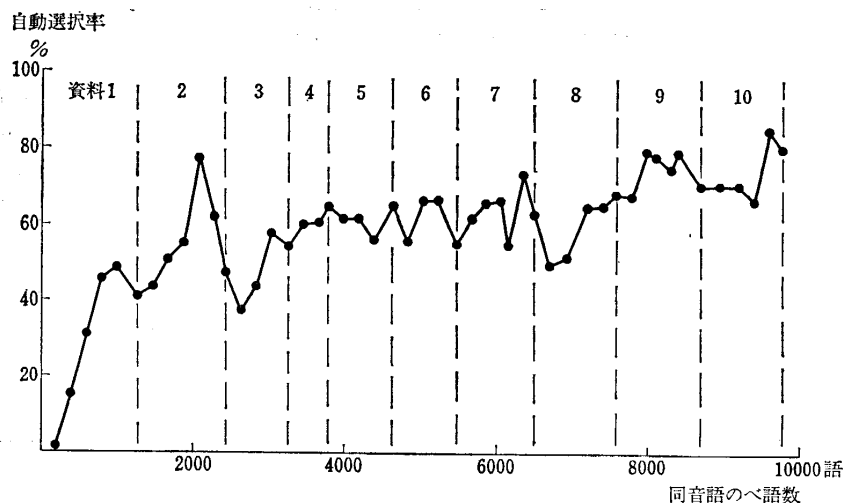


図 3 同音語自動選択率の推移

Fig. 3 Change in the rate of automatically selected homonyms.

したので、自動選択率は最初の4区間（のべ800語）ではほぼ0から50%前後まで急速に増大する。その後はゆるやかに増加し、同音語のべ入力語数が8,000語をこえた後、70%以上に達している。

4.2.2 誤 選 択

同音語とその前後に接続する文字との三つ組 awb について、入力中のある時点で選択条件が満たされて W_1 が選択されると、以後同一の三つ組 awb が出現した場合つねに W_1 が選択される。そこで、本来 W_2 であるものが W_1 に誤選択される場合が生じる。

表3に示されるように、本実験において、平均誤選択率は同音語総数の約2%で、漢字語総数に対する比率は極めて小さい。本実験システムには、誤選択になった同音語に対し、同一の三つが再度出現したときに再度誤選択されることを防止する機能は設けていない。したがって入力の累積とともに同一の三つ組による誤選択がくり返し発生することになる。表4に、本実験において発生した誤選択のうち、くり返し多数回発生したものの例を示す。この中でも特に「語」を「後」と、また「時、次」を「字」と誤ったものが非常に多く、誤選択総数の45%強を占める。それゆえ、同一の誤選択のくり返しを防止する機能の付加により、誤選択をさらに減少させることが可能である。

4.2.3 閾値 N , M の設定

前述のように、ある同音語 W_1 が自動選択されるのは次の二つの条件が満足されたときである。

- 1) W_1 のこれまでの出現頻度が N 以上。
- 2) 三つ組 aW_1b の出現頻度が W_1 の同音語 W_2 の同様な三つ組 aW_2b の出現頻度の M 倍以上。

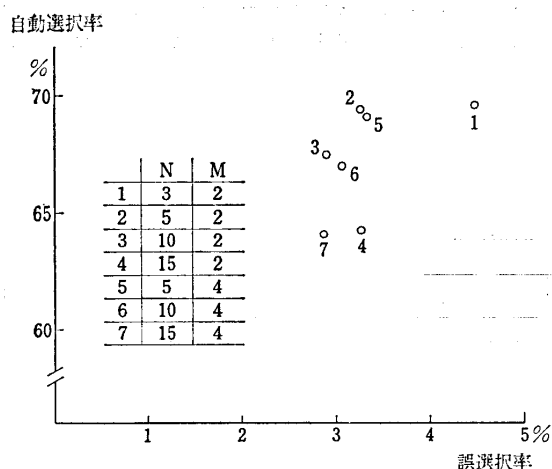
表 4 誤選択例

Table 4 Examples of erroneous selection.

誤選択された同音語			三つ組による分類		具 体 例
正	誤	回数	三 三 組	回数	
語	後	33	S 語 の	5	…カタカナ後の…
			S 語 P	5	…カタカナ後.
			S 語 K	4	…カタカナ後数…
			K 語 S	4	…日本後 (…
			そ の 他	15	
後	語	4	K 後 K	2	…送信語処理…
			K 後 P	2	…入力語,
時	字	28	K 時 の	23	…検索字の…
			S 時 の	4	…コンパイル字の…
			の 時 K	1	…の字分割…
次	字	4	N 次 K	3	…2字記憶…
			N 次 S	1	…1字エネルギー…
過程	家庭	12	K 過程の	9	…生産家庭の…
			K 過程に	3	…作成家庭に…
分	文	9	K 分 の	4	…開発文の…
			N 分 の	2	…2文の…
			そ の 他	3	
文	分	1	N 文 は	1	…960 分は…
項	孔	7	N 項 K	7	…3孔系列…
高	孔	1	S 高 K	1	…) 孔性能…

注) K…漢字, S…カタカナおよび記号, N…数字, P…句読点

したがって、 N , M の設定値は自動選択率、誤選択率に影響を与えるが、その程度は現実の文書における同音語の出現状況によって変動する。

図4 パラメータ N, M の効果Fig. 4 Effect of the parameters N and M .

今回行った実験では前述のように $N=10, M=2$ としたが、この適否を確認するために前記の資料のうち7~10を用い、 N, M を変えて実験を行った。ここで、同音語辞書の初期状態は資料6まで入力した時点の状態とした。この結果を図4に示す。

これからわかるように、 N, M を大きくすれば自動選択率が若干低下する代りに誤選率も減少する。しかしそのちがいはわずかであり、 $N=5\sim 10, M=2\sim 4$ であればほぼ一定とみなしてよい。

5. 考 察

以上の実験により、資料に出現する同音語の60~70%を自動選択できることが確かめられた。この値は文献5)に報告されている値(68.8%)とほぼ同程度であるが、文献5)では17,000語弱の同音異形語を収容する辞書を用いているのに対し、本論文の方式は400語程度の辞書を用い、さらに同音語に接続する語の代りに前後各1文字のみの情報によって同程度の結果を得ている。もちろん、対象としている文書が全く異なっているので単純な比較はできないが、このような小容量かつ簡単な構成の辞書によっても同程度の結果が得られることが示されたといえる。

しかしながら、これらの結果は最終的なものではなく、今後さらに検討すべき課題がいくつかある。以下、その主なものについて考察する。

(1) 自動選択を行うか否かを定める閾値 N, M について、 $N=5\sim 10, M=2\sim 4$ が妥当であるという実験結果が得られたが、より広範な分野について一般的に成立するか否かは明らかでない。

(2) 実験システムにおいて、「語」は漢字の部分

表5 自動選択率の改善
Table 5 Improvement of selection rate.

資 料 No.	自動 選 択 率 (%)	
	方 法 1	方 法 2
1	31.0	35.0
2	56.3	68.0
3	48.4	55.0
4	61.2	71.6
5	61.5	64.9
6	60.3	66.3
計	51.1	57.9

方法1 前後各1文字の接続を利用

方法2 方法1で選択できなかった語に対し後続2文字の接続による選択を加える

のみを切り出すのが基本である。それゆえ「～に関し」のような場合は「関」が1語となり、「間、感」などと同音語になるが、逆に「監視」とは同音語にならない。したがってこのような語が同音語総数や自動選択率に与える影響は複雑で、出現した個々の同音語についてより詳細な分析を行う必要がある。

(3) 同音語の前後1文字より多数の文字接続を利用すれば、さらに選択精度を上げることができると考えられる。これを確かめるために、前後各1文字の三つ組と、同音語と後続2文字との三つ組の双方を同音語辞書に登録し、前者で選択できなかった同音語についてさらに後者による選択を試みる方式の実験を行った。その結果、表5に示すように選択率が若干向上することが認められた。しかし、これにより同音語辞書、選択アルゴリズムともに複雑化することになり、それに見合うだけの選択率改善が常に得られるかどうかは未検討である。

(4) 本方式では、いったん選択基準を満足した同音語は、以後同一の三つ組が出現するごとに同じ語が選択され、前述のように同一の誤選択がくり返し発生する可能性がある。そこで、誤選択になった同音語については蓄積された頻度情報をリセットする等の処理を行って、くり返し発生を避けるようにする必要があるが、現在は個々の使用者任せになっている。

6. お わ り に

同音語とその前後に接続する文字との三つ組を辞書に登録し、入力文に現れる同様な三つ組との比較を行って同音語を選択する方式の実験を、情報処理および機械工学に関する10種の資料を用いて行った。その結果、辞書への登録語数が400語程度になれば、出現

した同音語の約 70% を自動選択できることがわかり、この方式の有効性が確かめられた。これを先に報告した同音語選択の固定化機能と組み合わせることにより、さらに良好な自動選択率を得ることができると考えられる。また、この実験では出現した同音語総数の 2% 前後の誤選択が生じたが、誤選択になった同音語について同音語辞書を更新する機能の付加によりこれを減少させることが可能である。

本論文はこのような方式による同音語自動選択の可能性を示したもので、今後さらに検討すべき点も多い。また、より大量かつ広い分野の資料を用いて入力実験を行い、多くの分野における性能の検証を行うことも今後に残された課題である。これらについて今後さらに検討を進める予定である。

謝辞 本研究に際し、種々ご討論いただき、適切な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感謝します。

参 考 文 献

- 1) 森 健一, 河田 勉: かな漢字変換, 情報処理, Vol. 20, No. 10, pp. 911-916 (1979).
- 2) 枅内香次, 齊藤 康: 適応型変換辞書を用いるかな漢字変換, 情報処理学会論文誌, Vol. 24, No. 2, pp. 209-213 (1983).
- 3) 枅内香次, 岡沢好高: 適応変換辞書方式かな漢字変換システムの性能測定, 情報処理学会論文誌, Vol. 26, No. 4, pp. 733-739 (1985).
- 4) 牧野 寛, 木澤 誠: べた書き文の仮名漢字変換システムとその同音語処理, 情報処理学会論文誌, Vol. 22, No. 1, pp. 59-67 (1981).
- 5) 中野 洋: 同音語の判別, 情報処理学会自然言語処理研究会資料, 33-4 (1982).
- 6) 枅内香次, 伊藤太亮, 荒木健治, 鈴木康広, 永田

邦一: 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, No. 119, pp. 119-126 (1984).

(昭和 60 年 1 月 7 日受付)

(昭和 60 年 10 月 17 日採録)



枅内 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。電子通信学会, 日本音響学会各会員。



伊藤 太亮 (正会員)

昭和 29 年生。昭和 52 年東京農工大学工学部電気工学科卒業。同年王子製紙(株)に入社。昭和 57 年 4 月より 2 年間、北海道大学工学部電子工学科研究生として、日本語文書処理の研究に従事。現在、生産設備の計画、制御設計に従事。運転制御装置のマン・マシンインタフェースの改善に興味をもっている。



鈴木 康広

昭和 35 年生。昭和 57 年北海道工業大学電気工学科卒業。昭和 60 年北海道大学大学院工学研究科修士課程情報工学専攻修了。現在同大学大学院博士後期課程在学中。語の接続関係を利用した日本語情報処理の研究に従事。電子通信学会会員。