企業名のカナ漢字変換システム

田　中　康　仁　（日本ユニバック）
大久保　静　子　（日本ユニバック）

はじめに

漢字システムに於いて重要な問題点の1つは入力作業である。入力作業はカナ文字ファイルから漢字ファイルへの移行作業と毎日毎日発生するトランザクション・データの入力作業である。これらの作業をさらに分析すると漢字の入力作業は次の3つの条件を満していなければならない。

この条件は

1. 安く　　2. 早く　　3. 正確に

である。

この三つの条件はどんなシステムにも必須であるが，これ以外に次の条件を特につけ加えておきたい。

4. 取り扱い易い

これは1，2，3という条件を満した入力システムでも前近代的な労働条件で人を働かせていたのでは決して良いものではない。又教育とか練習が特に必要なシステムを考えるとこのために要する費用も膨大なものになる。教育とか練習が必要なシステムでは人間の採用が限られるとか特別な手当を支払わなければならない等の問題も発生する。そこでここでは誰れでも取り扱えるシステムについて考えてみることにする。特に企業名についてカナ文字での入力を考えてみる。

この研究は3つの大きな部分から構成されている。

第1.　企業名のカナ漢字変換の可能性について
　　企業名単語の収集
第2.　企業名（カナ）の分かち書きについて
　　企業名単語が十分であるか
第3.　実用化への試行
　　評価（測定）改良

この研究は第3.の評価の一部分を残すだけになっている。

I　企業名のカナ漢字変換の可能性について

I.1.　企業名について

全国には一体どのくらいの企業があるのだろうか又それはどのような資料によって判るのだろうかと思い調査を行った。総理府統計局では5年毎に事業所統計を行いその内容を業種別，企業名（アイウエオ順）に並べ会社企業名鑑として出版している。最近の統計は昭和49年に実施しその内容が会社企業名鑑として出版されている。昭和50年3月に発行されたこの資料によると企業数は270,697件事業所数は323,637件である。これら企業も常にその姿を変えている。新しく起る企業，企業と企業の合併による名称の変更，倒産する企業，……時代の流れに従って企業名も少しずつ変化している。

このため個々の企業毎にカナ文字名と企業名の漢字を対応させることは変動す

る企業名を取り扱うには不便な事が判る。さらにカナ文字の表記方法には曖昧さがある。例えば「工業」は「コウギョウ」とカナ付けをするが「コウギョオ」「コオギョウ」……等と誤っている場合もある。又企業名は姓，名等に比べ桁数が長いという特徴がある。桁数が長くなればなるほどカナ文字の表現種類が増加する。このように考えると企業名のカナ文字に漢字の企業名を1対1に対応させることは不可能に近くなる。

I.2　企業名を単語に分解する

企業名を調べると殆んどの企業が「株式会社」という単語を持っている。さらに幾つかの企業名を頭に思い浮かべると面白いことがわかる。

例えば　○○銀行，田中商店，山田工業大阪支店を単語に分解すると

○○，銀行，田中，商店，山田，工業，大阪，支店になる。

これら単語は他の企業名にもしばしば使われている。さらにこれらの単語は次のようにも分類される。

1. 人名，地名　　2. 名称（東洋，日本）　　3. 業種名（銀行，保険　……）
4. 形態名（支店，工業，株式会社　……）　　5. その他の単語

このように分類された単語は次のように合成することも出来る。

$$\text{企業名} = \left\{\begin{matrix}\text{人名}\\\text{地名}\end{matrix}, \text{名称}\right\} \oplus \{\text{業種名}\} \oplus \{\text{形態名}\}$$

更に一般的に書けば次のようになる。

$$\text{企業名} = \{\text{単語}\} \oplus \{\text{単語}\} \oplus \{\text{単語}\} \oplus \cdots\cdots \oplus \{\text{単語}\}$$

但し，この単語は企業名に使われているものである。

企業名を単語に分割し個々の単語の使用頻度を調べ使用頻度の高い順に単語を並べその累積使用率を％で表わしてみるとそのグラフは次のようになる。

I.3　企業名単語ファイルの作成

総理府統計局で出版している企業名鑑をデータ収集の基礎資料にした。この資料は全国の全業種の企業名と事務所が記載しており又資本金も一定レベル（300万円以上）に制限しているためデータ収集の対象資料として最適である。

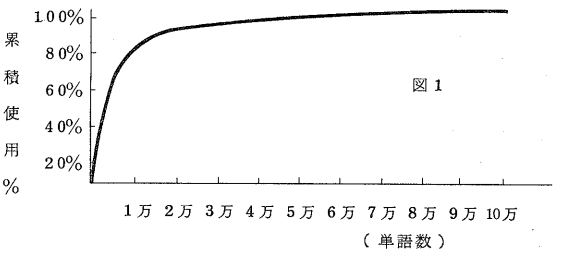


図1

第1回企業名単語収集状況

1. データ収集期間　：　S 49年12月～S 50年8月
2. 総データ件数　：　205,677件
3. 重複データ件数　：　98,761件
4. 桁数オーバ　：　8,685件（漢字5桁以上）
5. 正しいデータ　：　98,321件
6. 異った発音　：　62,115件
7. 調査資料　：　会社企業名鑑
　　会社四季報

第1回の企業名単語収集で約10万件弱の単語を集めた。但し第1回の企業名単語収集作業では事業所名を省略したので第2回目の作業でこれらの入力をした。

第2回企業名単語収集作業状況をまとめてみると次のようになる。

第 2 回 企 業 名 単 語 収 集 状 況

1. データ収集期間 ： S50年9月～12月
2. 入 力 件 数 ： 21,306件
3. 第1回データ ： 98,231件
4. 重 複 デ ー タ ： 16,098件
   桁 数 オ ー バ （カナ10桁以上，漢字6桁以上削除）
5. 正しいデータ ： 103,439件
6. 異った発音 ： 66,341件
7. 調 査 資 料 ： 会社企業名鑑（事業所名）

I.5 企業名単語ファイルの構造

収集した企業名単語ファイルの性質を調べるために次の統計資料を作成した。

1. アイウエオ順単語数調査（表1）
2. 漢字の桁数分布 （表2）
3. カナ文字桁数分布 （表3）
4. 同音異字単語の状況（表4）

これら統計資料によりシステム・デザインに必要な情報が得られる。特に表4は同音の漢字の書き方が異っているものはどの程度あるか調べたものである。頻度6程度で98.13%である。頻度7以上のものは使用頻度の高いものを選べばほぼ満足ゆくシステムができる。これまでの段階で研究の第1段階が終了した。

アイウエオ順単語数統計（表1）

カナ	件数	%	カナ	件数	%	カナ	件数	%	カナ	件数	%	カナ	件数	%	カナ	件数	%
ア	3,870		サ	3,726		ナ	2,346		マ	3,567		ヤ	1,835		ル	190	
イ	4,140		シ	7,381		ニ	1,915		ミ	2,869		ユ	1,128		レ	483	
ウ	1,909		ス	3,250		ヌ	141		ム	531		ヨ	1,649		ロ	590	
エ	2,260		セ	1,980		ネ	255		メ	710		小計	4,612	4.45	小計	3,124	3.02
オ	3,631		ソ	1,142		ノ	604		モ	1,139		ラ	526		ワ	914	
小計	15,810	15.28	小計	17,479	13.73	小計	5,261	5.08	小計	8,816	8.52	リ	1,335		合計	103,439	100.0
カ	6,366		タ	5,137		ハ	4,128										
キ	4,715		チ	1,717		ヒ	2,872										
ク	1,913		ツ	1,345		フ	3,409										
ケ	1,569		テ	1,919		ヘ	653										
コ	5,548		ト	4,093		ホ	2,004										
小計	20,111	19.44	小計	14,211	5.08	小計	13,066	12.63									

同音異字単語の状況（表4）

頻度	件数	%	頻度	件数	%
1	51,456	77.56	17	22	0.03
2	7,920	11.94	18	22	0.03
3	2,922	4.40	19	9	0.01
4	1,482	2.23	20	19	0.03
5	819	1.23	21	9	0.01
6	516	0.77	22	7	0.01
7	327	0.49	23	10	0.01
8	221	0.33	24	6	0.01
9	153	0.23	25	9	0.01
10	104	0.15	26	2	0.00
11	69	0.10	27	4	0.00
12	42	0.06	28	1	0.00
13	50	0.07	29	1	0.00
14	38	0.05	30	3	0.00
15	28	0.04	30以上	38	0.05
16	32	0.05	合計	66,341	100.0

漢字の桁数と件数（表2）

	件数	%
1	1,498	1.448
2	53,234	51.464
3	27,458	26.545
4	15,935	15.405
5	5,314	5.137
合計	103,439	100.0

カナ文字桁数と件数（表3）

桁数	件数	%
1	105	0.101
2	4,330	4.186
3	25,131	24.295
4	44,778	43.289
5	15,599	15.080
6	6,237	6.029
7	3,582	3.463
8	2,291	2.214
9	1,386	1.340
合計	103,439	100.0

Ⅱ. 企業名（カナ）の分かち書きについて，企業名単語が十分であるか。

Ⅱ.1　企業名の研究方法

これまでの段階で企業名の特性，企業名単語の収集について述べた。しかしこの方法を進めても研究に新しい面を加えることが出来ないため研究方法を変えることにした。企業名は姓や名前の様に単語に分割されずカナ文字の文字列として連なっている。この連らなった文字列を人間の手作業によって分割することは大変な作業である。この作業を半自動的に行う方法とそのための単語について研究する。さらに分かち書きに使用する単語と企業名単語の照合により不足する単語の調査を行う，又カナ文字の種々の統計を取ることにより性質を分析することにした。これらのことは3つの部分的目的に表わすことができる。

カナ文字企業名の研究目的

1. 企業名単語（カナ文字）の統計を取り性質を調べる。
2. 企業名単語が10万語で十分であるか。
3. 分かち書きは最長一致法やパターン・マッチィングでどの程度正確に行えるか。

上記3つの部分的目的を満すカナ文字企業名を探したが見つからないため独自に作成することにした。

カナ文字ファイル作成用資料は次のものを使用した。

1. 資 料 名 ： 企業名鑑（50年度版）
2. データ数 ： 約27万件
3. 調査日数 ： S51年1月～5月
4. 調査用紙 ： 次のものを使用した。

企業名カナ漢字変換テスト用紙　　PAGE

企業 No.	重複件数	企業名（カナ文字、分かち書き）
1 2 3 4 5 6 7 8		5 タナカ ショウテン
2 3 4 5 6 7 8 9		1 ニホンバシ ショウジ

この調査用紙は独自にデザインし簡易印刷した。調査用紙の記入要領は企業名をカナ文字で分かち書きし，同一企業名は重複件数を記入し入力を省力した。

株式会社，有限会社等のカナ文字は省略した。

この調査用紙の記入において企業名を分かち書きにした理由は種々の統計を取りたいため，分かち書きのテストをコンピュータで行った場合に正確さの判定を自動的に行わせ人手による分析を少なくするためである。単純な用紙であるが分かち書きを行わせたためこれ以後の作業を楽にした。

Ⅱ.2　企業名のカナ文字統計

企業名のカナ文字統計は次の4つのものを調べた。

1. 企業名の桁数統計（全体）
2. 企業名単語の桁数統計
3. 企業名単語に分かち書きされる項目件数統計
4. 同一企業名の発生頻度統計

調査結果をつぎに述べる。

企業名の桁数統計によると平均9桁で最長桁数は43桁である。3桁～29桁のデータでほぼ99.84%を占めている。桁数の長いものはヒラガナ表示のもの英字をヒラガナで表現したものが多い。このデータはファイルのデータの長さを決めるために使われる。（但しこのデータ表の中には株式会社，有限会社等の文字は含まない）

企業名単語の桁数統計は分かち書きした個々の単語の桁数である。最多桁数は4桁である。8桁以内の桁数で95.85%を占めている。9桁以上のデータは分かち書きを正しく行わなかったものがほとんどである。

企業名単語に分かち書きされる項目件数は最多項目数は2である。分かち書きは多くの場合4項目以内でほとんどが占められ，まれには8項目にまで分割されることがある。

同一企業名の発生頻度統計は同じ発音の企業名が全国ではどの程度ありその重複度合を調べたものである。19万8千件の企業名に重複がないことがわかる。
企業名カナ名に対応して企業名漢字を対応させることは大変労力のかかることがわかる。（表6参照）
この中で発生頻度の最も多いものはマツモトショウテンであった。

同一企業名の発生頻度（表6）

重複度	カナ見出(件数)	累積%
781～100	13	1.0
～ 33	129	3.0
～ 22	336	5.0
～ 10	1,276	10.0
～ 6	3,064	15.0
～ 4	5,955	20.0
～ 2	1,500	25.0
～ 1	198,824	100.0
合　計	272,667	－

Ⅱ.3　企業名単語10万語で十分であるか？

分かち書きされたカナ文字企業名を項目ごとに分解しアイウエオ順に分類し企業名単語マスターと照合した。その結果次のことが判った。

1. 調査日時：S51年9月
2. 分かち書き項目総数：558,978件
3. 企業名単語マスターとの照合結果

	件数	種類
マッチ：	477,165件	25,076件
アンマッチ：	81,813件＊	41,296件＊
合　計	558,978件＊	66,372件＊

＊の中には不注意による分かち書きのなされていないデータが23,000件程度入っている。これは種類，件数とも同数入っている。それ故補正して考える必要がある。

企業名のカナ文字単語は約4万件程度で十分であり頻度順分析結果は表7のとおりである。

10万語の収集した企業名単語によるマッチング率は90%（補正後）であった。アンマッチ・データのうち頻度の高いものの入力を行えば98%程度のマッチ率に容易に向上することも明らかになった。

企業名単語分析表（表7）

種類件数	単語累積%
2	5%
4	10%
20	20%
64	30%
167	40%
379	50%
814	60%
1,818	70%

種類件数	単語累積%
4,688	80 %
16,932	90 %
38,424	95 %
60,783	99 %
63,577	99.5%
65,815	99.9%
66,372	100.0%

（補正前）

企業名単語の頻度の高いもの（表8）

1	コウギョウ	7	デンキ
2	ショウテン	8	セイサクショ
3	ケンセツ	9	ジドウシヤ
4	ショウジ	10	ニホン
5	サンギョウ	11	ハンバイ
6	ショウカイ	12	セイサクジョ

入力もれの発生したものは次のようなものであった。

1. カタカナ表示のもの　　例　ホテル，ナショナル等
2. 頭文字マのつく文字が少なかった　　例　マルイ，マルタカ等

これは入力の際データが一部抜けたか又は調査ミスが原因であった。
この調査により不足データが明確になり，この不足したデータはその後追加入力しデータの不足を補った。（約1万件程度の入力データであった。）

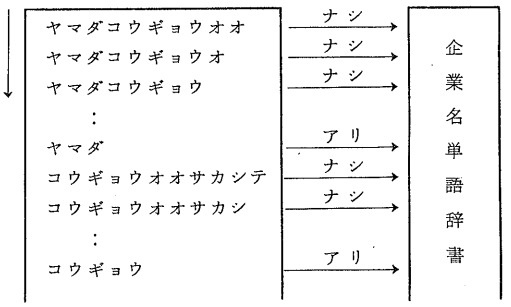
Ⅱ.4　分かち書きは最長一致法で正確に行えるか。

最長一致法は分かち書きを行う場合普通用いられるものである。これについて簡単に説明する。
ヤマダコウギョウオオサカシテンを分解するには次の図2のようにする。
この方法によると結果は次のようになる。
ヤマダ△コウギョウ△オオサカ△シテン

分かち書きを最長一致法でテストする際カナ文字の途中にある空白は左につめ一連の文字列としこのテストを行った。
テスト結果が分かち書きした入力データと一致するか比較した。これは人手による分析の省力化のためであった。

最　長　一　致　法　　（図2）

ヤマダコウギョウオオサカシテンの分解
濁点も含め12桁のカナ文字を比較の対象とする。



12桁の文字列の途中に空白がある場合はその空白までの文字列を分解する

見出し語として使用頻度の高いカナ文字企業単語1万語を採用し行った。しかしこれはあまり良い結果は得られなかった。そこでカナ文字企業名単語4万語を採用し行った。この結果は大変良かった。

分析結果

1. 分かち書きがコンピュータと人手の結果が一致したもの　　85%
2. コンピュータの分かち書きの方がより詳しく行われたもの　　5%

アイチポンプコウギョウ　アイチ△ポンプ△コウギョウ（コンピュータ）　アイチ△ポンプコウギョウ（人手）

3. どうしょうもないもの

アールエーステーション　　アル△エース△テーション（コンピュータ）　　アル△エー△ステーション（人手）

オオタケオリモノ　　オオタケ△オリモノ（コンピュータ）　　オオタ△ケオリモノ（人手）

4. 単語の追加，削除により改良の余地のあるもの

この結果より95%程度のデータは正しく分かち書きが行なえることがわかる。
処理速度は次のとおりである。

1. 1件の処理時間	： 92.2 ms	4. I／O	： 49.5 ms
2. CPU	： 32.5 ms	5. 1秒間の処理件数	： 11件
3. CC／ER	： 10.1 ms	6. 使用言語	： COBOL

最長一致法は文字列の途中で企業名単語ファイルに該当する文字が無い場合，それ以後の分かち書きへの分解が行われない又行われても誤まりを起し易いという欠点がある。　この欠点を補い分かち書きの速度を向上させるためには企業名単語の中で使用頻度の高い単語についてパターン・マッチイング法を使い最長一致法の欠点を補うことが良い方法である。

パターン・マッチイング用の単語としては前記表8のように使用頻度の高い単語で4～5桁以上の

長い単語を採用すべきであろう。（有意語の選定）

さらに最長一致法に“一部後退機能”をもたせる。この機能は最長一致法が正しく行われない場合次に長い単語を採用するものである。

分かち書きについてもほぼ正確に行えることが確認できた。ここまでの段階で第2段階の研究が終了した。

Ⅲ. このシステムの実用化

Ⅲ.1 企業名カナ漢字変換の流れとファイル構成

カナ文字の企業名より漢字の企業名への変換は既に述べた方法により次の3段階をへて行う。

1. オリジナル・データの整備（カナ文字のエラー・チェック）

   タナカショウテン　，　ヤマダコウギョウオオサカシテン

2. これらのカナ文字を企業名カナ単語ファイルによって次のように分かち書きに変える。
   （最長一致法とパターン・マッチ法を採用）

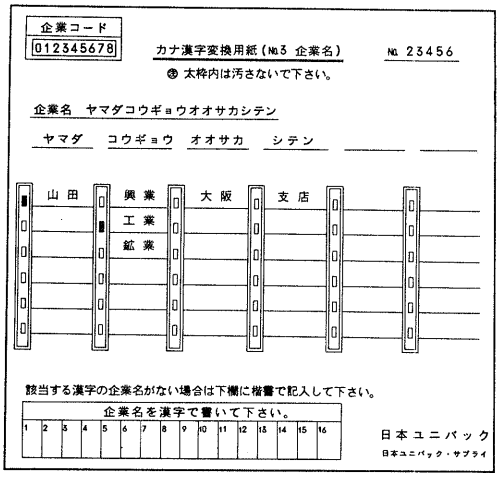
   タナカ△ショウテン　，　ヤマダ△コウギョン△オオサカ△シテン

3. 分かち書きに漢子をあてはめ漢字企業名を合成する。

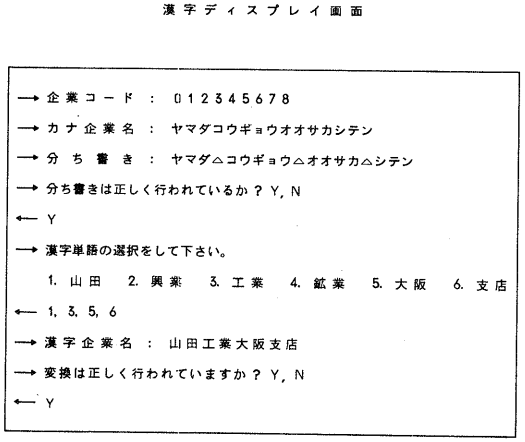
   田中商店　　山田工業大阪支店

同音異字語の選択はOCR用紙に漢字プリンターでプリント・アウトする方法と漢字ディスプレイを使う2種類がある。OCR用紙による変換方法は次の用紙にプリント・アウトし人間の視覚によって同音異字語の選択を行いマークを付け処理する。

OCR用紙による変換方式　図3



漢字ディスプレイによる変換方式　図4



漢字ディスプレイ方式は日本ユニバック総合研究所で開発中である。

OCR又はシートによる変換も日本ユニバック総合研究所で開発し，ほぼ終了した。

漢字ディスプレイ方式のプロセスとファイル構成は次のようになる。

企業名カナ・インデックス（4万件），企業名単語ファイル（10万件）をディスク・ファイルにロードし，カナ・データを読み，分かち書きを行い，漢字単語を引き，漢字ディスプレイで同音語の処理をし，漢字に変換を行う。

企業名カナ・インデックスと企業名単語は同一化すべきであるが今回の開発では別ファイルとした。

Ⅳ. このシステムの評価

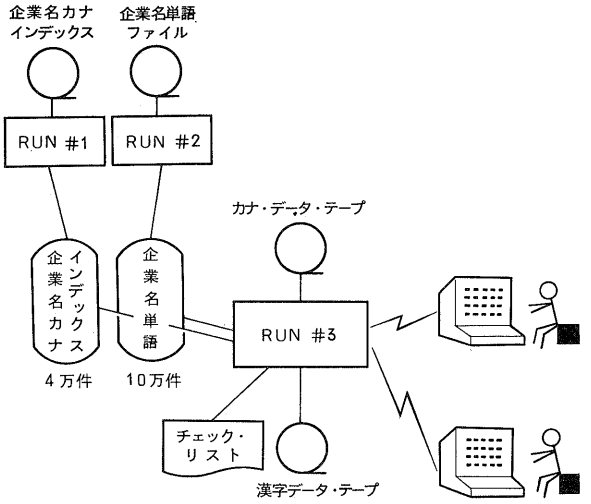
この入力方式はComputer による機械処理を多くし期間と労力を減らす方法である。今後我々の社会では単純労働の安い賃金は得られにくくなっていることを考えるとこの方法は大変良いシステムともいえる。

このような変換方式で中規模の銀行を考え直接入力方式と変換方式の比較を机上で行ってみることにした。

前提条件

1. 業　　種 ： 銀　行
2. デ ー タ 量 ： 200万件(全データ)
3. 支　店　数 ： 100支店
4. データの性質 ： 姓名データ8割　**
   企業名データ2割

プロセスとファイルの構成



5. 直接入力方式には地方支店への撮影機材運搬費用，出張旅費は含まず。
6. 変換システムでは支店でのマーク付作業は含まず。
7. 未変換率2割と想定している。
   （実際には1割以下の予定）

直接入力と変換システムの比較

方式	作業工程	作業期間	コスト	データの正確性	チェックシステム	その他
直接入力方式	原票撮影 原票複写 入力 校正 訂正入力 支店への問合せ アンマッチチェック 訂正確認 ブルーフリスト検収	データ件数 200,000件 100店舗 17ケ月	原票撮影　円 原票複写　円 入力　円 アンマッチ チェック　円 ブルーフリスト 作成　円 TOTAL　円 直接入力1件費用をaとする。	入力ミス 校正ミス 不明原票の扱い等の点で精度には限度が有る。 カナ文字との関連がとりにくい。	原票との照合 エラーリストの訂正 支店への問合せ アンマッチ・リストの訂正	地方支店の場合原票撮影の機材運搬出張旅費がさらに必要。
変換システム	支店説明会 コンピュータ処理 漢字プリント 支店チェック キーパンチ 補正入力 ブルーフリスト検収	データ件数 2,000,000件 100店舗 6～7ケ月	コンピュータ処理円 漢字プリント　円 キーパンチ　円 補正入力　円 シート　円 TOTAL　0.3a円 変換方式コスト ＝0.3a円	基本的にユーザのファイルに忠実である。 一連作業の中に支店チェックが組み込まれている。 精度はよい カナ文字との関連がとりやすい。	支店チェックによるデータは信頼性が高い 支店別ブルーフリストによりデータの抜けがチェックできる。	地方支店のマスター漢字化も容容易に行える。

前に述べた比較表からも判るように多量のデータについては変換方式が有利であることがわかる。特に次のような点を特徴としてあげておきたい。

1. 漢字ファイル作成期間の短縮
2. 漢字ファイル作成コストの低減
3. 漢字ファイル作成の正確性
4. 現有ファイルの有効利用（カナ文字との連動）

このシステムを作成して判ったことであるが直接入力方式ではカナ文字との連動が全くつかないことがある。

例えば次の入力を考える。

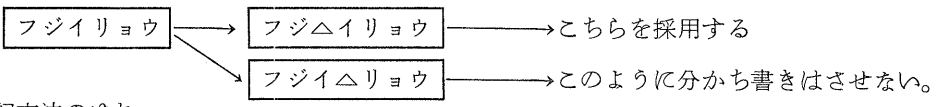
コンピュータのカナ文字 ： オカヤマアサヒコウトウガッコウ又はアサヒコウトウガッコウ
原　　票 ： 岡山県立岡山朝日高等学校
直接入力によると ： 岡山県立岡山朝日高等学校
変換方式によると ： 岡山朝日高等学校又は朝日高等学校

変換方式の方がコンピュータのカナ文字と連動が行われることがわかるであろう。

V. 今後の研究

今後の研究テーマとしては次のようなことを考えている。

1. 企業名単語の分析
   企業名単語は姓，名，地名，名称，企業名独特のもの，その他から構成されているのでそれらの重複度合いの調査。
2. 単語の接続
   単語には前に接続を許すもの，許さないもの等がある。これにより変換のあいまいさを減らす。
3. あいまいさの除去
   ある場合には単語の分割をさけ長い単語を採用する。
   ex　接続させる　化学工業
4. 分かち書きのあいまいさの除去
   カナ見出しの中に特別のカナ文字列は強制的に変換するように固定する。



5. 表記方法のゆれ
   表記方法の種類が多数あるもの
   アルミニュウム　　エンジニアリング
   アルミニューム　　エンジニヤリング
6. 企業名を複合語として研究を行う。
   これらについて今後調査，研究を続けてゆきたい。

VI. このシステムの研究と開発過程

この研究は2年半の期間をかけて行われた。このように長期にわたった理由はこの作業にたずさわる専門の人がアサインされず筆者の一部分の時間で行われたためである。又このような経験が他に無かったためでもある。

1.	'74年 8月 ～ 11月	企業名に使われる単語の分析
2.	'74年12月 ～ '75年4月	単語の整理（一次）
3.	'75年 5月 ～ 7月	単語辞書の磁気テープ化（一次）
4.	'75年 8月 ～ 10月	企業名のカナ漢字変換システムのドキュメント化
5.	'75年 8月 ～ 10月	単語の整理（二次）
6.	'75年11月 ～ '76年4月	カナ文字統計の準備
7.	'76年 1月	SHARE-11に発表（日本ユニバック総研発行）
8.	'76年 5月 ～ '76年9月	カナ文字統計資料作成，シミュレーション
9.	'76年10月 ～ '76年12月	ディスプレイを用いたシステム，OCR方式のシステム開発
10.	'77年 1月 ～	今後の研究

Ⅷ. おわりに

企業名のカナ漢字変換システムについての変換率の測定，操作性，運用等については今後の研究にまかせたい。なおこのシステム作成にあたって助言して下さった社会保険庁の上田陸奥夫氏，又このシステム作成のために協力して下さった人々に深く感謝致します。

参考文献

1. 文書情報処理に関する研究48-S004（情報処理開発センター）
2. Computer Report 1974/12 P35～P45 "JAPATICにおける漢字検索システム"
3. 昭和50年度第16回大会講演論文集（情報処理学会）
   昭和51年度第17回大会講演論文集（情報処理学会）

資　料

1. 会社企業名鑑（S46年度版，S50年度版）（総理府統計局編集）
2. 会社四季報（東洋経済社）
3. 電話帳　職業別