姓名のカナ漢字変換システム

田中康仁
（日本ユニバック(株) 金融SE部システム推進G）

1. まえがき

最近漢字を処理する研究が活発になってきました。漢字のHardware, Softwareの問題は種々ありますが技術の進歩とともに解決の方向を見いだしています。

しかし漢字インプットについては新しい方式がなく相変らず人手によって1文字1文字漢字入力装置（例えば漢字ベンコーダー，漢字ディスプレイ，ベンコーダー，漢テレ等）によって入力するのが普通です。このような入力方法ですと種々の問題があります。入力機器が多量に必要であるとか，いままでカナ文字で出来た種々のマスターの移行に長期間を要する，漢字処理システムの普及に大きな障害をおよぼす等数えあげればきりがありません。

そこでこの問題を解決する方法としてカナ文字から漢字への移行を考えてみました。特に姓名のカナ漢字変換システムをお話し申し上げ，御批判を頂きたいと思います。

2. インプット・システムに必要な条件

漢字インプットに必要な条件は

1) 正確に
2) 早く
3) 安く

の3つの項目があげられます。

従来使われている漢字入力機器での入力についてこの3つの条件を考えてみよう。漢字入力機器での入力はオペレーターが1文字1文字，3000文字から4000文字の文字盤から選ばなければならない。この操作を正確に行うには熟練した技術が必要である。又正確であるという保障を得るために校正を行わなければならない。

早くという条件について考えてみると漢字入力機器の入力では平均50～60文字/分，1日の作業量は約1万5千文字であります。このくらいの入力では200万件とか300万件のマスターを移行するには20台とか30台の入力機器で約1年～2年かかります。この1～2年間には約1割～2割のデーターが変更されてしまいます。又新しいデーターも約1割ぐらい発生します。マスターの移行が終りますと入力に必要な機器は必要なくなります。数社～数十社が同時に漢字システムの導入を行うとするとただ単純な算術での計算だけではすまされない問題が発生します。

安くという条件を考えてみます。漢字の入力費用は校正の度合い，原票の状態コーディング化されているか否か，などにより異りますが1字2円～2円50銭程度といわれています。この価格を維持するためには数年間定期的に一定量の入力データーが有るということが保障されていなければなりません。マスター移行のような場合を考えてみますと一時的に多量のデーターが発生し，その後データー量が極端に少なくなります。

このようなことを考えますと漢字入力機器での入力は決して安い方法ではありません。

3. このシステムの開発過程.

3.1 アイデアの段階.

昭和47年3月頃 新しい入力について何か変った方法はないか考えていました。その当時 東邦生命の高西部長から"学習研究社でカナ文→漢字変換が可能である"と話していた"ことを聞きました。これが可能であれば大変なことなのでこっそく学習研究社の後藤氏に会いその内容を聞いてみました。その原理は簡単で「タナカ」に対応して漢字の「田中」「田仲」「多中」を対応させ，人間の介入によって「田中」を選択する方法であった。我々は全て自動的に変換することを考えていたため、このような簡単な発想が出てこなかった。この方法は人間の介入があるから従来の漢字のインプットと変らないではないかという意見もあるが、数千の文字盤の中から2~3文字を選ぶのと数個の中から1つ選択するのでは大変な差があります。

そこでこの発想を理論的に発展させてみました。

3.2. カナ文字の姓名を分析する.

この大変興味深い方法について研究を行ってみた。原理はわかっても全日に氏名がいくつあるか、それがどのような頻度分布をしているか、などの基本的事項は全くわかっていない。柳田国男先生が約8~10万の姓があると書いているのはあるが、これも統計的な研究を行って書いたものではなく、一つの推測にしかすぎない。全国を網羅し重複して登録していないファイルがないものかと探してみたが適当なものがなかった。そこでそれに近いものとして第百生命，東邦生命のファイルを借り統計的処理を行ってみた。生命保険のファイルはすべて全国を網羅しており，しかも氏名が独特であるから保険に入らないという性質のもので

はないので統計を取ってす十分意味があると判断したからです。

また力ナ文字の姓と名を区別するために1桁のスペースが入っているから姓と名を区別することが容易にできるという利点もある。

その結果を表にしてまとめてみました。

表1 姓の頻度調査 (第百生命)

	種類	総種類に対する割合	件数	総件数に対する割合
アイウエオ	1,111	4.33%	27,823	3.88%
	1,724	6.73	54,203	7.57
	823	3.21	14,957	2.08
	287	1.12	6,436	0.89
	1,508	5.88	49,428	6.90
ア 行 計	5,453	21.29	152,852	21.35
カキクケコ	1,839	7.18	42,124	5.88
	717	2.80	17,250	2.40
	827	3.23	16,530	2.30
	96	0.37	415	0.05
	1,082	4.22	30,151	4.21
カ 行 計	4,561	17.81	106,470	14.87
サシスセソ	859	3.35	40,556	5.66
	1,442	5.63	23,829	3.33
	438	1.71	22,862	3.19
	317	1.23	5,799	0.81
	229	0.89	2,595	0.36
サ 行 計	3,285	12.83	95,704	13.36
タチツテト	1,387	5.41	51,913	7.25
	269	1.05	2,368	0.33
	605	2.36	9,500	1.32
	235	0.91	3,949	0.55
	923	3.60	10,732	1.49
タ 行 計	3,419	13.35	78,453	10.95
ナニヌネノ	875	3.41	33,064	4.61
	436	1.70	11,284	1.57
	89	0.34	874	0.12
	78	0.30	1,273	0.17
	317	1.23	7,012	0.97
ナ 行 計	1,795	7.01	53,507	7.47
ハヒフヘホ	934	3.64	25,685	3.58
	692	2.70	13,437	1.87
	771	3.01	22,083	3.08
	52	0.20	322	0.04
	439	1.71	10,139	1.41
ハ 行 計	2,888	11.28	71,946	10.05
マミムメモ	752	2.93	26,463	3.69
	914	3.57	21,232	2.95
	291	1.13	8,594	1.20
	43	0.16	253	0.03
	438	1.71	13,261	1.85
マ 行 計	2,438	9.52	69,803	9.75
ヤユヨ	672	2.62	33,931	4.74
	207	0.80	1,665	0.23
	445	1.74	16,506	2.30
ヤ 行 計	1,325	5.17	52,102	7.27
ラリルレロ	17	0.06	66	0.00
	78	0.30	500	0.06
	2	0.00	6	0.00
	10	0.03	15	0.00
	26	0.10	80	0.01
ラ 行 計	133	0.52	667	0.09
ワ	291	1.13	13,405	1.87
ワ 行 計	291	1.13	13,405	1.87
調査より除外したデーター			21,216	2.96
合 計	25,605	100.	715,851	100.

表2 姓の頻度分布

順位	累計パーセント
～ 50	27.81
～ 100	37.21
～ 200	48.40
～ 300	55.35
～ 500	63.80
～ 1000	74.59
～ 2000	83.61
～ 3000	87.89
～ 5000	92.30
～ 10000	96.66
～ 15000	98.38
～ 20000	99.21
～ 25000	99.91

図1 姓の頻度分布 (第百生命)

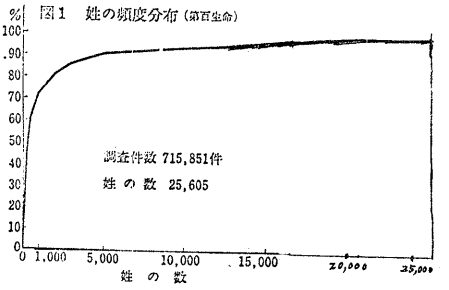


表3 姓の頻度分布 (東邦生命)

	姓の種類		件数	
	数	比率	数	比率
アイウエオ	1,613 2,427 1,483 458 2,138	3.98 5.93 3.66 1.13 5.27	125,379 201,335 56,163 25,381 191,556	4.73 7.60 2.12 0.95 7.23
ア行計	8,119	19.98	599,814	22.66
カキクケコ	2,850 1,107 1,207 172 1,657	7.03 2.73 2.98 0.42 4.09	157,192 65,642 68,296 1,943 114,420	5.93 2.48 2.58 0.05 4.32
カ行計	6,993	17.26	406,893	15.37
サシスセソ	1,257 1,778 709 475 347	3.10 4.39 1.75 1.17 0.85	173,697 90,820 88,053 21,787 9,296	6.56 3.43 3.32 0.82 0.35
サ行計	4.566	11.27	383,635	14.49
タチツテト	2,007 430 921 373 1,381	4.95 1.06 2.27 0.92 3.41	191,411 12,095 36,432 13,861 43,049	7.23 0.45 1.37 0.52 1.62
タ行計	5,112	12.62	296,848	11.21
ナニヌネノ	1,311 689 160 129 500	3.23 1.70 0.39 0.31 1.23	123,452 43,225 3,954 4,772 23,439	4.66 1.63 0.15 0.18 0.88
ナ行計	2,789	6.88	198,942	7.51
ハヒフヘホ	1,392 1,052 1,135 112 685	3.43 2.59 2.80 0.27 1.69	94,808 49,091 83,335 1,512 40,030	3.58 1.85 3.14 0.05 1.51
ハ行計	4,376	10.80	268,776	10.15
マミムメモ	1,105 3,394 505 105 602	2.72 8.38 1.24 0.25 1.48	94,750 79,657 30,638 1,121 46,418	3.58 3.00 1.15 0.04 1.75
マ行計	5,711	14.10	252,584	9.54
ヤユヨ	1,052 315 635	2.59 0.77 1.56	121,772 6,839 61,407	4.60 0.25 2.32
ヤ行計	2,002	4.94	190,018	7.17
ラリルレロ	54 165 37 21 75	0.13 0.40 0.09 0.05 0.18	228 1,791 143 38 390	0.00 0.06 0.00 0.00 0.01
ラ行計	352	0.86	2,590	0.09
ワ	479	1.18	46,911	1.77
合計	40,499	100.	2,647,011	100.

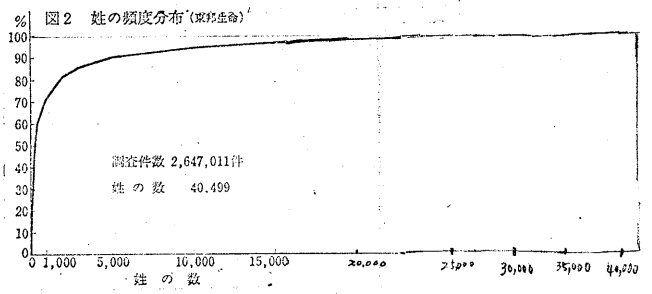
表4 頻度上位60までの姓 (東邦生命)

順位	苗字	順位	苗字
1	佐藤	31	太田(大田)
2	鈴木	32	三浦
3	高橋	33	岡田
4	伊藤	34	村上
5	渡辺	35	藤田
6	斎藤	36	長谷川
7	田中	37	中島
8	小林	38	坂井(酒井)
9	佐々木	39	工藤
10	山本	40	坂本
11	中村	41	前田
12	加藤	42	小野
13	吉田	43	青木
14	阿部(安倍)	44	山下
15	山田	45	金子
16	木村	46	近藤
17	松本	47	竹田(武田)
18	井上	48	新井(荒井)
19	山口	49	中野
20	林	50	松田
21	菊池(菊地)	51	中川
22	橋本	52	竹内(武内)
23	森	53	田村
24	清水	54	柴田(芝田)
25	遠藤	55	千葉
26	池田	56	上田(植田)
27	山崎	57	藤井
28	佐藤	58	西村
29	石川	59	福田
30	小川	60	岡本

表4 頻度上位60までの姓 (東邦生命)

順位	姓	件数	順位	姓	件数
1	サトウ	54,689	31	オオタ	7,937
2	スズキ	42,163	32	ミウラ	7,815
3	タカハシ	36,045	33	オカダ	7,738
4	イトウ	28,470	34	ムラカミ	7,731
5	ワタナベ	28,396	35	フジタ	7,596
6	サイトウ	27,603	36	ハセガワ	7,554
7	タナカ	25,825	37	ナカジマ	7,352
8	コバヤシ	22,393	38	サカイ	7,231
9	ササキ	21,197	39	クドウ	7,146
10	ヤマモト	19,709	40	サカモト	7,056
11	ナカムラ	18,873	41	マエダ	7,044
12	カトウ	18,376	42	小ノ	6,859
13	ヨシダ	18,346	43	アオキ	6,840
14	アベ	16,330	44	ヤマシタ	6,794
15	ヤマダ	15,176	45	カネコ	6,744
16	キムラ	11,195	46	コンドウ	6,709
17	マツモト	11,981	47	タケダ	6,561
18	イノウエ	11,762	48	アライ	6,447
19	ヤマグチ	11,585	49	ナカノ	6,239
20	ハヤシ	10,911	50	マツダ	6,257
21	キタチ	10,429	51	ナカガワ	6,149
22	ハシモト	10,232	52	タケウチ	6,097
23	モリ	10,036	53	タムラ	6,006
24	シミズ	9,900	54	シバタ	5,968
25	ニンドウ	9,519	55	チバ	5,943
26	イケダ	9,484	56	ウエダ	5,911
27	ヤマザキ	8,605	57	フジイ	5,838
28	ゴトウ	8,535	58	ニシムラ	5,701
29	イシカワ	8,315	59	フクダ	5,637
30	オガワ	8,169	60	オカモト	5,480

図2 姓の頻度分布(東邦生命)



70万から264万件にデーターを増加させても増加した割合で姓の数は増えず，累積曲線はほぼ一致していた。このことよりカナの姓の数は約4万位又はこれより増加してもそれは管理システムを考える上で大きな影響はあたえないと思われる。

名についての統計資料

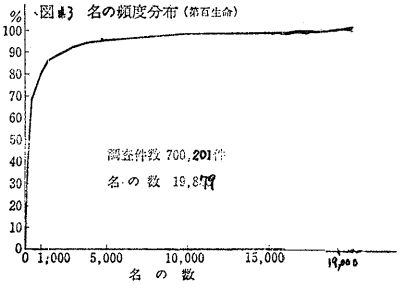
表6 名の分布 (第百生命)

	男		女		男		女	
	名前の数	%	名前の数	%	件数	%	件数	%
アイウエオ	408 554 194 253 184	2.05 2.78 0.97 1.27 0.92	125 151 52 88 101	0.62 0.76 0.26 0.44 0.50	15,515 13,724 930 6,693 2,782	2.27 1.96 0.13 0.95 0.39	7,205 2,648 812 5,248 191	1.02 0.37 0.11 0.74 0.02
ア 行 計	1,593	8.01	517	2.60	40,048	5.71	16,105	2.30
カキクケコ	854 807 313 364 471	4.29 4.05 1.57 1.83 2.36	234 236 71 60 146	1.17 1.63 0.35 0.30 0.73	34,473 19,166 6,229 16,871 14,605	4.96 2.73 0.88 2.40 2.08	9,461 12,824 2,113 3,825 976	1.35 1.83 0.30 0.54 0.13
カ 行 計	2,809	14.13	747	3.75	91,349	13.04	29,199	4.17
サシスセソ	534 1,293 307 393 187	2.68 6.52 1.54 1.97 0.94	160 320 129 54 49	0.80 1.60 0.64 0.32 0.24	12,791 46,324 5,979 10,319 1,759	1.82 6.61 0.74 1.47 0.25	7,324 8,032 4,085 3,061 358	1.04 1.14 0.58 0.43 0.05
サ 行 計	2,719	13.67	722	3.63	77,142	11.01	22,860	3.26
タチツテト	992 411 357 300 853	4.99 2.06 1.79 1.50 4.29	224 186 132 67 210	1.12 0.93 0.66 0.33 1.05	45,308 4,936 8,586 9,731 32,691	4.67 0.57 1.22 1.38 4.65	6,413 6,639 2,470 2,728 11,417	0.91 0.94 0.35 0.38 1.63
タ 行 計	2,913	14.65	819	4.11	100,352	14.33	29,667	4.23
ナニヌネノ	362 97 12 15 228	1.82 0.48 0.06 0.07 1.44	111 22 5 1 64	0.55 0.11 0.02 0.01 0.32	3,872 304 18 48 14,638	0.55 0.04 0.01 0.01 2.00	2,379 24 56 1 4,088	0.33 0.01 0.01 0.— 0.58
ナ 行 計	714	3.59	203	1.02	18,281	2.61	6,548	0.93
ハヒフヘホ	330 581 426 85 106	1.66 2.92 2.14 0.42 0.53	122 153 142 6 33	0.61 0.76 0.71 0.03 0.16	8,206 42,240 7,590 806 235	1.17 6.04 1.08 0.08 0.03	5,639 9,033 6,939 11 46	0.80 1.29 0.99 0.01 0.01
ハ 行 計	1,528	7.68	456	2.29	59,237	8.45	21,742	3.10
マミムメモ	550 540 151 37 264	2.76 2.71 0.75 0.18 1.32	163 245 33 22 53	0.81 1.23 0.16 0.11 0.26	41,325 21,243 1,481 287 3,639	5.90 3.03 0.21 0.04 0.51	9,050 17,531 514 235 751	1.29 2.50 0.07 0.03 0.10
マ 行 計	1,547	7.78	516	2.59	67,975	9.70	28,081	4.01
ヤユヨ	289 256 503	1.45 1.28 2.53	93 84 134	0.46 0.42 0.67	11,531 14,132 33,372	1.64 2.01 4.76	4,047 6,084 9,945	0.57 0.85 1.42
ヤ 行 計	1,048	5.27	311	1.56	59,035	8.43	20,081	2.86
ラリルレロ	20 345 12 56 38	0.10 1.73 0.06 0.28 0.19	12 84 13 21 4	0.06 0.41 0.06 0.10 0.02	36 6,448 37 526 388	0.01 0.92 0.01 0.07 0.05	59 1,895 337 1,695 30	0.01 0.27 0.04 0.24 0.01
ラ 行 計	471	2.63	133	0.66	7,435	1.06	4,006	0.57
ワ	79	0.39	34	0.17	779	0.11	279	0.03
ワ 行 計	79	0.39	34	0.17	779	0.11	279	0.03
男女別計	15,421	77.53	4,458	22.38	521,633	74.45	178,568	25.46
合 計	名前の数	19,879		100.0	件 数	700,201件		100.0

表7 名の頻度分布

順 位	累 計(%)
～ 50	27.82
～ 100	40.56
～ 200	54.37
～ 300	61.90
～ 500	70.78
～ 1,000	81.34
～ 1,500	86.43
～ 2,000	89.48
～ 3,000	92.98
～ 4,000	94.89
～ 5,000	96.06
～10,000	98.47
～15,000	99.32
～19,879	100.00

図3 名の頻度分布 (第百生命)



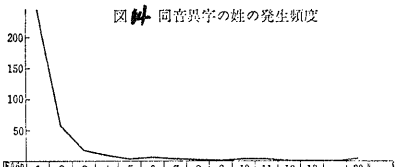
名前について姓の統計と同様なことを行ってみた。名前は世襲のものでなく個人にとって一代かぎりのものであるから、膨大な種類があるだろうと想像していた。しかし統計を取って調べてみるとこの予想はみごとにはずれてしまった。

名前の種類は姓の種類よりも少ないのである。頻度順に名前を並べ名前の累計%を作成すると表7のようになる。約3,000の名前で93%。1万で98%になる。親達は、子供に立派になってほしいと和裏をはぼるにもかかわらず、名にも姓以上の先等りがあり、代表的名前でかなりの部分を占めていることがわかる。この図3を図1,図2と比べてみると名前のグラフの立ちが急であることがわかる。東邦生命のファイルについて同様の分析を行ってみたがほぼ同じ結果を得た。

3.3 同一発音の姓名の発生頻度

カナ文字ファイルだけの分析では不十分であるため一つの発音の姓名についてそれがどの程度の種類を持っているか分析を行った。
6そう件の姓を調査し同音異字の発生頻度を表にすると次のようになる。

図14 同音異字の姓の発生頻度



	1	2	3	4	5	6	7	8	9	10	11	12	13	…	20	
頻度数	247	59	22	12	3	7	4	1	0	2	1	0	0	0	1	359
姓の件数	247	118	66	48	15	42	28	8	0	20	11	0	0	0	20	623
%	39.6	18.9	10.6	7.7	2.4	6.7	4.5	1.3	0	3.2	1.8	0	0	0	3.2	100

これによると1つの呼び方について平均約1.7個の異った姓があることがわかる。
東邦生命の呼び方の件数(姓)4万件と掛け合わせると約6万8千件ということになる。
これから約全国には6万以上の姓がありそれくらいのデータは集めなければならない。
名についても同様の調査を行うと平均約3.8個の異った姓があることがわかった。
東邦生命の呼び方の件数(名)3万5千件と掛け合わせると約13万千ということになる。これから全口には約13万以上の名がありそれぐらいのデータが必要であることがわかった。

3.4. 姓名マスタ・データの収集

このシステムを作成するにあたって一番困難であったのはどのようにしてデータを集めるかであった。これはコンピューターの分析力ではできない仕事でこつこつとした調査と多くの協力者によってデータを集めることができた。この方法について具体的方法,内容は省略する。

オー次マスター作成 S.48.6.

	総件数	異った発音
姓	62,468件	37,212件
名	38,726件	10,655件

オ二次マスター作成 S.49.7.

	総件数	異った発音
姓	71,044件	約4万5千件 *
名	93,099件	約2万5千件 *

(*未調査のため予想数)

オ三次マスター作成 S49.12.予定

	総件数	異った発音
姓	約12万件	約6万件
名	約13万件	約3~4万件

姓名マスターはオ三次マスターまで作成する予定でありこの時点で姓約12万名約13万件になる予定である。

3.5. 姓名マスターの変換率の推定

姓名のカナ漢字変換用マスターを上記予定で作成しているがこれがどの程度の変換率か,又どのような問題点があるか分析してみる必要がある。そこで約54万件のカナ・ファイルと約228万件のカナ・ファイルについてカナ文字段階での変換率の測定を行った。
姓名のマスターはオー次マスター使用.
54万件のファイルで実験
1.姓 98.%の変換率

インプットデータ	マッチ・データ	アンマッチ・データ
547,210件	535,166件	12,044件

姓の種類	マッチした種類	アンマッチの種類
21,040件	16,237件	4803件

2. 名 96%の変換率

インプットデータ	マッチデータ	アンマッチ・データ
547,093件	525,752件	21,341件
名の種類	マッチした種類	アンマッチの種類
13,240件	6,781件	6,459件

22万件のファイルで実験

1. 姓 96.5%の変換率

インプット・データ	マッチデータ	アンマッチ・データ
2246,663件	2,166,954件	79,709件
姓の種類	マッチした種類	アンマッチの種類
53,390件	28,503件	24,887件

2. 名 95%の変換率

インプットデータ	マッチデータ	アンマッチ・データ
2,281,277件	2,165,988件	115,289件
名の種類	マッチした種類	アンマッチの種類
36,096件	9,184件	26,912件

第一次データによるテストであるが姓名とも95%以上の%を示している。これにより姓名マスターを作成する作業にはあまり先走りがなく集められていることがわかる。

第一次データは姓6万2千件名3万8千件であるため少し不十分であることも同時にわかる。このテストはカナ文字でのテストで実質変換率はこれより2～3%下るはずである。

しかし第二次，第三次データではこの変換率も向上し最終的には実質変換率で次の目標になる予定である。

対象顧客数 500万件 実質変換率 姓 99%
名 98%

3.6 データ・ファイルの問題点

アンマッチ・データを分析してみるとジ，ヂやズ，ヅの使われ方がまちまちであるためアンマッチになった例が多くみうけられた。例えば「カズオ」と「カヅオ」があらわれていたりする。そのため「カズオ」は処理されても「カヅオ」はアンマッチになる等の例がみうけられた。

このようなものの例としては次のものがある。

1. ジとヂ，ズとヅ
2. 清音と濁音の相異　サワ，サワ等
3. 長音，促音などでの表わしかの異い。

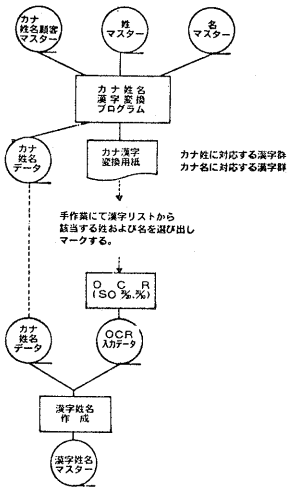
これらのデータのためには　カズオはカズオと等しいという処理をほどこすことにより変換率を向上させることにした。

4. 姓名マスターを使用しての変換プロセス

カナの姓名から漢字の姓名に変換する際，人間の介入を最小とし誰れでも行える方法としてOCR伝票を使用することを考えた。OCR伝票以外のものとしては漢字ディスプレイが考えられるがオペレーター1人に対する投資額が増えるのでこれについては

別の面から考えると，ここでは研究の対象からはずした。

4.1. 変換プロセス。



変換作業フロー

変換作業は三つに分割される。

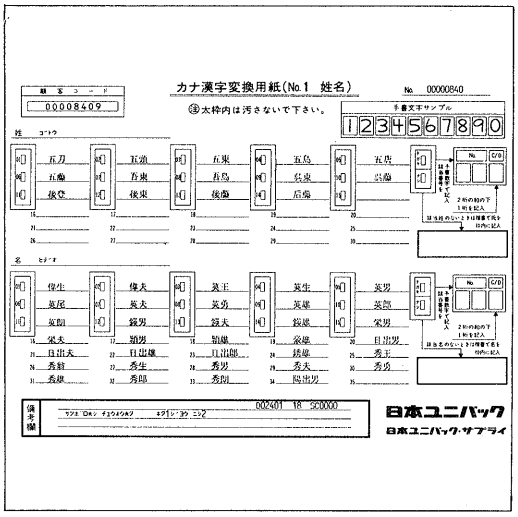
(1) 顧客のカナ姓名ファイルと漢字姓および名マスターから「カナ姓名漢字変換プログラム」を使用し，カナ姓に対応する漢字群およびカナ名に対応する漢字群を漢字プリンタを使用して印字する。

(2) リストされた姓および名の漢字群から手作業にて該当する姓および名を選び出し用紙上にマークをつける。

(3) 用紙上のマークをOCRで読むことによりカナ姓名に対応する漢字を決定する。

なおこのシステムで変換不可能な特殊な姓名に関しては漢字入力装置を使用して別途入力する。

4.2. 変換に使用するOCR用紙。



この用紙の記入方法はゴトウに対応するものに，ヒデオに対応するものにマークそれは手書きの数字を記入することによって入力する。印刷されたもの以外の姓名については用紙の姓名それぞれの欄に手書きの欄がありそこに記入し漢字入力装置で入力する。

5. おわりに

カナ文字の姓名を漢字の姓名に早く安く，正確に変換できることがわかった。今後はさらに単に姓名という分野でのカナ漢字変換ではなく，もっと別な分野のカナ漢字変換に考えてゆきたい。最後にこのシステムを作成するにあたって多大な協力をよせて下さった東邦生命，第一生命の方々に深く感謝します。