

姓名のカナ漢字変換システム*

田 中 康 仁**

1. ま え が き

最近、漢字情報処理に関する研究が活発になっている。漢字情報処理の Hardware, Software については種々の問題点はあるが、技術の進歩と共に解決の方向にある。

しかし、漢字インプットについては新しい方式がなく、相変わらず人によって1文字1文字漢字入力装置(例えば漢字ペンコーダー、漢字ディスプレイ、ペンコーダー、漢テレ等)によって入力するのが普通である。このような入力方法では、種々の問題がある。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ない。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た。本稿では特に姓名のカナ文字変換システムを紹介したい。

2. インプット・システムに必要な条件

漢字インプット・システムに必要な条件は

- (1) 正確に
- (2) 早く
- (3) 安く

の3つの項目が上げられる。

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条件を考えてみる。漢字入力機器での入力はオペレーターが1文字1文字、3,000文字から4,000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り、また正確であるという保障を得るためには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1万5千文字である。このような入力では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年かかる。この1~2年間には1割~2割のデータが変更されてしまうし、また新しいデータも1割程度発生する。マスターの移行が終了すると入力に要した機器は必要なくなる。

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数々の問題が発生する。

安くという条件を考えてみると、原票の状態、コーディング化されているか否か、などにより異なるが、1文字の入力費用は2円~2円50銭程度といわれている。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が有ということが保障されていなければならない。マスターの移行のような場合を考えてみると、一時的に多量のデータが発生し、その後データ量が極端に少なくなる。このようなことを考えると漢字入力機器での入力は決して安い方法ではない。

3. システム開発過程

3.1 フィデアの段階

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たが、その頃学習研究社がカナ文字の漢字変換を考えていることを知った。その原理は簡単で、たとえば「タナカ」に対応して漢字の「田中」「田仲」「多中」を対応させ、人間の介入により「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来なかった。この方法は人間の介入があるから従来の漢字インプットと変らないではないかという意見もあるが、数千文字の文字盤の中から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る。さらに、オペレーターの誤読による似た文字への誤りもなくなる。そこでこの発想を理論的に発展させてみた。

3.2 カナ文字の姓名の分析

この原理はわかっても全国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

* Kana to Kanji Conversion Software for persons names by Yasuhiro TANAKA (Nippon UNIVAC Kaisha Co, Ltd.)

** 日本ユニパック (株) 金融システム推進部システム・マネジャー

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	1,111	4.33 %	27,828	3.88 %
イ	1,724	6.73	54,203	7.57
ウ	823	3.21	14,957	2.08
エ	287	1.12	6,436	0.89
オ	1,508	5.88	49,428	6.90
ア行計	5,453	21.29	152,852	21.35
カ	1,839	7.18	42,124	5.88
キ	717	2.80	17,250	2.40
ク	827	3.23	16,530	2.30
ケ	96	0.37	415	0.05
コ	1,082	4.22	30,151	4.21
カ行計	4,561	17.81	106,470	14.87
サ	859	3.35	40,556	5.66
シ	1,442	5.63	23,829	3.33
ス	438	1.71	22,862	3.19
セ	317	1.23	5,799	0.81
ソ	229	0.89	2,595	0.36
サ行計	3,285	12.83	95,704	13.36
タ	1,387	5.41	51,913	7.25
チ	269	1.05	2,368	0.33
ツ	605	2.36	9,500	1.32
テ	235	0.91	3,940	0.55
ト	923	3.60	10,732	1.49
タ行計	3,419	13.35	78,453	10.95
ナ	875	3.41	33,064	4.61
ニ	436	1.70	11,284	1.57
ヌ	89	0.34	874	0.12
ネ	78	0.30	1,273	0.17
ノ	317	1.23	7,012	0.97
ナ行計	1,795	7.01	53,507	7.47
ハ	934	3.64	25,695	3.58
ヒ	692	2.70	13,437	1.87
フ	771	3.01	22,083	3.08
ヘ	52	0.20	322	0.04
ホ	439	1.71	10,139	1.41
ハ行計	2,888	11.28	71,946	10.05
マ	752	2.93	26,463	3.69
ミ	914	3.57	21,232	2.96
ム	291	1.13	8,594	1.20
メ	43	0.16	258	0.03
モ	438	1.71	13,261	1.85
マ行計	2,438	9.52	69,808	9.75
ヤ	672	2.62	33,931	4.74
ユ	207	0.80	1,665	0.23
ヨ	446	1.74	16,506	2.30
ヤ行計	1,325	5.17	52,102	7.27
ラ	17	0.06	66	0.00
リ	78	0.30	500	0.06
ル	2	0.00	6	0.00
レ	10	0.03	15	0.00
ロ	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			21,216	2.96
合 計	25,605	100.	715,851	100.

的事項は全くわかっていない。柳田国男先生が8~10万の姓があると書いておられるが、これも統計的な研究を行なわれたものではなく、一つの推測にすぎ

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

表-2 姓の頻度分布

(第百生命)	
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

のファイルを借り統計的处理を行った。生命保険のファイルは全国を網羅しており、しかも氏名が独特であるから、保険に入れないという性質のものではないので、統計を取っても十分意味があると判断したからである。

またカ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出来るという利点もある。

その結果は表-1～5、図-1～2 のようになる。

70 万から 264 万件にデーターを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。このことから姓の数は約4万位またこれより増加してもそれは変換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はなく個人にとって一代限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(表-6 p236 参照)。

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 (p237 参照) 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知恵をしぼる

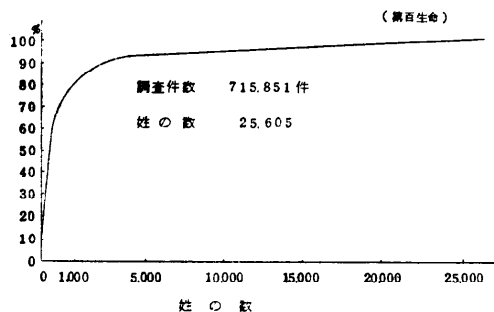


図-1 姓の頻度分布

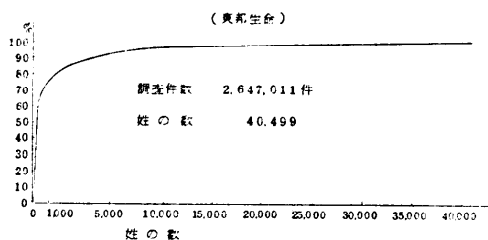


図-2 姓の頻度分布

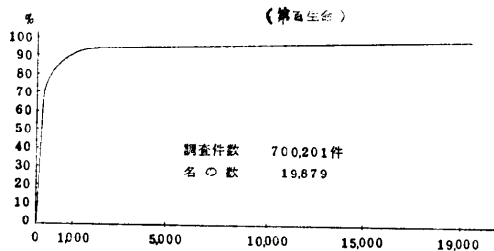


図-3 名の頻度分布

にもかかわらず、名にも姓以上の片寄りがあり、代表的名前でかなりの部分を占めていることがわかる。

さらに図-3 を図-1、図-2 と比べてみると、名前の方の方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。

623 件の異なった姓を調査し、同音異字の発生頻度を表にすると図-4 (p235 参照) のようになる。

これによると、1つの呼び方について平均1.7個の異な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。

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	3.98	125,379	4.73
	2,427	5.93	201,335	7.60
	1,483	3.66	56,163	2.12
	458	1.13	25,381	0.95
	2,138	5.27	191,556	7.23
ア行計	8,119	19.98	599,814	22.66
カ	2,850	7.03	157,192	5.93
キ	1,107	2.73	65,642	2.48
ク	1,207	2.98	68,296	2.58
ケ	172	0.42	1,343	0.05
コ	1,657	4.09	114,420	4.32
カ行計	6,993	17.26	406,893	15.37
サ	1,257	3.10	173,697	6.56
シ	1,778	4.39	90,820	3.43
ス	709	1.75	88,053	3.32
セ	475	1.17	21,787	0.82
ソ	347	0.85	9,296	0.35
サ行計	4,566	11.27	383,635	14.49
タ	2,007	4.95	191,411	7.23
チ	430	1.06	12,095	0.45
ツ	921	2.27	36,432	1.37
テ	373	0.92	13,861	0.52
ト	1,381	3.41	43,049	1.62
タ行計	5,112	12.62	296,848	11.21
ナ	1,311	3.23	123,452	4.66
ニ	689	1.70	43,225	1.63
ヌ	160	0.39	3,964	0.15
ネ	129	0.31	4,772	0.18
ノ	500	1.23	23,439	0.88
ナ行計	2,789	6.88	198,942	7.51
ハ	1,392	3.43	94,808	3.58
ヒ	1,052	2.59	49,091	1.85
フ	1,135	2.80	83,335	3.14
ヘ	112	0.27	1,512	0.05
ホ	685	1.69	40,030	1.51
ハ行計	4,376	10.80	268,776	10.15
マ	1,105	2.72	94,750	3.58
ミ	3,394	8.38	79,657	3.00
ム	505	1.24	30,638	1.15
メ	105	0.25	1,121	0.04
モ	602	1.48	46,418	1.75
マ行計	5,711	14.10	252,584	9.54
ヤ	1,052	2.59	121,772	4.60
ユ	315	0.77	6,839	0.25
ヨ	635	1.56	61,407	2.32
ヤ行計	2,002	4.94	190,018	7.17
ラ	54	0.13	228	0.00
リ	165	0.40	1,791	0.06
ル	37	0.09	143	0.00
レ	21	0.05	38	0.00
ロ	75	0.18	390	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.

これから全国には約6万以上の姓があり、またこれくらいのデータは集めなければならないし、名についても同様の調査を行うと平均約3.8個の異なった名があることがわかる。

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13万3千ということになり、これから全国には約13万以上の名があり、またこれくらいのデータが必要であることがわかる。

3.4 姓名のマスタ・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

表-5 頻度上位 60 までの姓

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(沼井)
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	後藤	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	7,967
2	ス ズ キ	42,163	32	オ ミ ウ	7,815
3	タ カ ハ	36,045	33	オ カ ダ	7,738
4	イ ト ウ	28,470	34	ム ラ カ	7,731
5	ワ タ ナ	28,396	35	フ ジ タ	7,596
6	サ イ ト	27,603	36	ハ ガ ガ	7,564
7	タ ナ カ	25,825	37	ナ カ ジ	7,352
8	コ バ ヤ	22,393	38	サ カ イ	7,231
9	サ サ キ	21,197	39	ク ド ウ	7,146
10	ヤ マ モ	19,709	40	サ カ モ	7,056
11	ナ カ ム	18,873	41	マ エ ダ	7,044
12	カ ト ウ	18,376	42	オ ノ	6,859
13	ヨ シ ダ	18,346	43	ア オ キ	6,840
14	ア	16,330	44	ヤ メ シ	6,794
15	ヤ マ ダ	15,176	45	カ ネ コ	6,744
16	キ ム ラ	11,195	46	コ ン ド	6,709
17	マ ツ モ	11,951	47	タ ケ ダ	6,561
18	イ ノ ウ	11,762	48	ア ラ イ	6,447
19	ヤ マ グ	11,585	49	ナ カ ノ	6,289
20	ハ ヤ シ	10,911	50	マ ツ ダ	6,257
21	キ ク チ	10,429	51	ナ カ ガ	6,149
22	ハ シ モ	10,232	52	タ ケ ウ	6,097
23	モ	10,036	53	タ ム ラ	6,006
24	シ ミ ズ	9,900	54	シ バ タ	5,968
25	エ ン ド	9,519	55	チ バ	5,943
26	イ ケ ダ	9,484	56	ウ エ ダ	5,911
27	ヤ マ ザ	8,605	57	フ ジ イ	5,838
28	ゴ ト ウ	8,595	58	ニ シ ム	5,701
29	イ シ カ	8,315	59	フ ク ダ	5,637
30	オ ガ ワ	8,169	60	オ カ モ	5,480

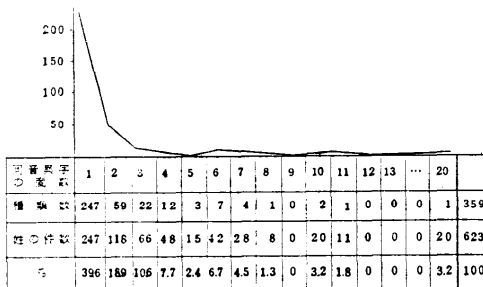


図-4 同音異字の姓の発生頻度

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。この内容については省略する。

第一次マスター作成 S48.6.

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

第二次マスター作成 S49.7.

総件数 異なった発音
姓 71,048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)

第三次マスター作成 S50.2. 予定

総件数 異なった発音
姓 約10万件 約6万件
名 約15万件 約4万件

姓名マスターは第三次マスターまで作成する予定であり、この時点で姓約10万、名約15万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

(変換されたデータ数(カナ文字段階) / 総変換対象データ数) × 100(%) すなわ

ち、与えられたカナの音に対応する漢字を選び出す) かまたは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

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	2.05	125	0.62	15,915	2.27	7,206	1.02
ア行計	1593	8.01	517	2.60	40,048	5.71	16,105	2.30
カキクケコ	854	4.29	234	1.17	34,478	4.96	9,461	1.33
カ行計	2809	14.13	747	3.75	91,349	13.04	29,199	4.17
サシスセソ	534	2.68	160	0.80	12,791	1.82	7,324	1.04
サ行計	2,719	13.67	722	3.63	77,142	11.01	22,860	3.26
タチツト	992	4.99	224	1.12	45,308	6.47	6,413	0.91
タ行計	2,913	14.65	819	4.11	100,352	14.33	29,667	4.23
ナニノ	362	1.82	111	0.55	3,872	0.55	2,379	0.33
ナ行計	714	3.59	203	1.02	18,281	2.61	6,548	0.93
ハヒフヘホ	330	1.66	122	0.61	8,206	1.17	5,639	0.80
ハ行計	1,528	7.68	456	2.29	59,237	8.45	21,742	3.10
マミムメモ	550	2.76	163	0.81	41,325	5.90	9,050	1.29
マ行計	1,547	7.78	516	2.59	67,975	9.70	28,081	4.01
ヤユヨ	289	1.45	93	0.46	11,531	1.64	4,047	0.57
ヤ行計	1,048	5.27	311	1.56	59,035	8.43	20,081	2.86
ラリルレロ	20	0.10	12	0.06	36	0.01	59	0.01
ラ行計	471	2.63	133	0.66	7,435	1.06	4,006	0.57
ワヰヱヲ	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	19,879		1000		700,201件		100.0	

の測定を行った。

姓名マスターは第一次マスターを使用。

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	4,803件

2. 名 96% の変換率

インプット・データ	マッチした種類	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

228 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
2,246,663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	24,787件

表-7 名の頻度分布

順	位	果 計 (%)
～	5 0	27.8 2
～	1 0 0	40.5 6
～	2 0 0	54.3 7
～	3 0 0	61.9 0
～	5 0 0	70.7 8
～	1,000	81.3 4
～	1,500	86.4 3
～	2,000	89.4 8
～	3,000	92.9 8
～	4,000	94.8 9
～	5,000	96.0 6
～	10,000	98.4 7
～	15,000	99.3 2
～	19,879	100.0 0

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

これらテスト結果は第一次マスター・データによるものであるが、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片寄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

$$\left(\frac{\text{変換されたデータ数(漢字での段階)}}{\text{総変換対象データ数}} \right) \times 100(\%)$$
 すなわ

ち、与えられたカナに対応する正しい漢字を選び出す

はこれより数 % 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到達する予定である。

対象変換数 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・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「カゾオ」があらわれていたりする。このため「カズオ」は処理されても「カゾオ」はアンマッチになる等の例がみられ、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. 清音と濁音の相違、ザワ、サワ等

3. 長音、促音などの表わし方の違い

これらのデータのためには、カズオはカゾ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. 変換用紙の設計

姓名の変換用紙は、第5章にサンプルを示しているが、この用紙を作成するにあたり二、三の統計を取ってみた。まず姓名の桁数は直感的に2桁が平均であるということはわかるが、何桁が最高でありどのような分布になっているかと問いかねられると行きづまってしまう。第一次マスターで集めた姓名を分析すると表-8 のようになる。

カナ文字の桁数を分析すると表-9 のようになる。

表-8

桁 数	姓	名
1	1,531	1,891
2	50,973	29,949
3	9,748	6,602
4	209	266
5	6	15
6	0	2
7	0	1
8	0	0
計	62,477件	38,726件

表-9

桁 数	姓	名
1	59	6
2	4,060	928
3	22,278	13,558
4	30,838	19,140
5	4,574	3,642
6	549	1,240
7	88	204
8	7	8
9	0	0
10	2	0
11	0	0
計	62,455件	38,726件

変換用紙上に1つの姓に対してまた名に対して何個の漢字姓名を表わしておけば十分であるか、また変換の作業を行う人は何個ぐらいが最適であるか、漢字プリンターでプリントする場合にプリントが遅くならないか等の課題を解決するため、1つの発音に対する漢字姓名の分布を調べてみた。使用したデータは第一次マスター・データである。(表-10 次頁参照)

表-10 からわかるように、姓は1つの発音に対して15～20個ぐらいでほとんどを占めているが、名の方は1つの発音に対して30以上のものがある。

表-10

頻度 姓名	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	30以上	合計
姓	24,779	6,706	2,428	1,288	623	362	227	160	88	55	46	34	26	15	11	9	7	6	1	2	3	0	1	1	1	1	0	0	1	0	0	36,885
名	5,727	1,684	889	511	367	216	183	153	128	91	71	80	59	44	33	40	25	22	23	24	21	24	16	12	12	17	8	7	15	9	145	10,655

それ故姓よりも名の方に多くの表示場所をもうけなければならない。

同一の発音の姓・名の中でもよく調べてみると使われている頻度がかなり異なることがわかる。例えば「タナカ」という発音でほとんどの人が「田中」で次に「田仲」その次が「多中」である。名を調べてみると次のようになっている。これは 25 万人の名の頻度を調べたものを紹介しよう。

例えば「アキオ」を取りあげてみると

1	アキオ	昭夫	193	17	アキオ	明生	8
2		秋雄	125	18		章郎	6
3		昭男	105	19		晃雄	5
4		秋男	95	20		彰雄	5
5		明夫	83	21		明郎	4
6		昭雄	88	22		晁雄	3
7		秋夫	76	23		旭男	3
8		明男	70	24		顕雄	3
9		明雄	40	25		晃男	3
10		章夫	22	26		秋郎	3
11		章雄	16	27		昌夫	3
12		章男	14	28		晃生	2
13		昭生	13	29		暉夫	2
14		秋生	12	30		晶夫	2
15		昭郎	11	31		章生	2
16		彰男	8	32		明郎	2

頻度 1 件のものを並べてみると次のようになる。

安起郎, 安喜男, 安喜夫, 暉雄, 晁生, 晁男, 晁夫, 顕男, 顕夫, 顕郎, 晃夫, 晃郎, 皓男, 皓郎, 秋尾, 穉雄, 昭大, 昭尾, 昭朗, 照夫, 昌雄, 晶男, 晶雄, 彰百, 彰郎, 彰朗, 暉夫, 椿男, 旦夫, 哲男, 哲郎, 斌夫, 斌郎, 明広, 萌夫, 耀雄, 亮男, 亮夫, 朗男, 朗夫, 朗雄 (1 件のもの 41 件)

のようになる。

同一の発音を取りあげてみてもこのようにかたよりがあることは興味のあることである。同一の発音の中を頻度順にならべることににより変換率を向上させることが出来る。

このような頻度調査は変換用紙に姓・名の漢字のパターンを並べた用紙を作成した後マークをつけやすく

するためである。それと名のように 30 を越える同一発音があるとき、そのうちのどれとどれを用紙にプリントするかということは変換率を向上させる上で重要な要素である。

5. 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、誰れでも行なえる方法として 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 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。

図-5 にその変換のプロセスを示す。

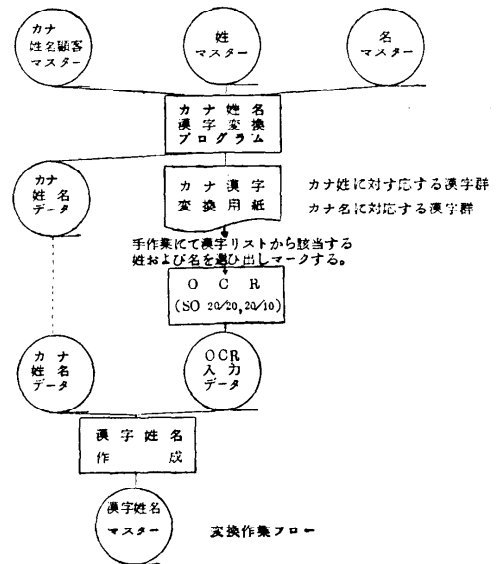


図-5 姓名マスターを使用しての変換のプロセス

変換作業は次の三つに分割される。

- (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。

5.2 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、また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機器で入力する。

6. このシステムの評価

このシステムのプログラムは、昭和 49 年 4 月作成し当社のダイレクト・メールを使い稼働させてみた。この時はプログラムの作成に主眼がおかれ、データのコンバージョンはさほど重点をおかなかった。ダイレクト・メールのようにカナ文字のシステムで十分使われているものを漢字のシステムに変換するためには、その原票になるものが確実に保管されていない場合は変換を行うのが少し困難である。

第二回のテストは昭和 49 年 10 月実施した。

データ件数 22,938件

姓アンマッチ件数 344件 1.5% (カナ文字レベル)

名アンマッチ件数 427件 1.9%

エラー・データ件数 20件

(エラー・データは姓と名のあいだにブランクが入っていなかったものである。)

機械的に変換出来たものは約 85% ぐらいであった。実質変換率があまりよくない理由は、

マスター・ファイルの変換対象データが女性がおもであったので

(1) 女性の名前が少なかったこと。

かな文字の名前が少なかった

(2) 名前が十分頻度順に並んでいなかった等のためであった。

しかし、これ等の問題点も第三次マスターでは解決し姓・名の頻度順整理も終る予定である。これらの作業は昭和 50 年 2 月までには完了する予定である。

第 2 回の変換作業は原表のそろっている業務で行ったため原表がないという問題は発生しなかった。

この変換システムは人間が介入して原表と照合するという部分を除いてしまえば、機械のプロセスが多いため、コンバージョンが早く行えるという利点があった。正確性という面では、変換用紙上に印書されたものを選択するという作業があるため、誤記、誤読などによるミスが発生する隙がないため正確に変換出来ていた。このため校正作業を行わないですぐに正確な入力作業が完了した。

費用という面から考えると、開発費をどのように分配するか、機械プロセスをどのように費用として考えるか等むづかしい問題がある。しかしこの面もデータ件数が数十万件から数百万件になると十分安い費用で処理出来ると思われる。

7. 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、別な分野、例えば、法人名、地名、一般文のカナ漢字変換を考えてゆきたい。最後にこのシステムを作成するにあたって学習研究社の後藤氏および多大な協力をよせて下さった東邦生命、第百生命の方々、またこのシステム作成に協力して下さいた多くの人々に深く感謝します。

(昭和 50 年 1 月 10 日受付)