日本人の姓と名の分布

1. 姓名カナ漢字変換システムの誕生

昭和47年春頃から漢字について研究をはじめたが，情報処理に漢字を使うとしたらどのようなことが問題となるだろうかと当時考えてみた．このとき，そしていまも重要なテーマであるのは漢字の入力である．

そこでこの漢字の入力を少しでも簡単に処理できる方法を考えてみることにした．一般の事務部門で使われるとしたら，いままで処理している請求書やダイレクトメールなどの宛名，地名が中心であると考えられる．

地名は全国もれなく集めても10万件くらいであり，個々にはっきり定まっているから処理はあまり困らない．カナ文字の地名から漢字の地名に変換することもそれほど苦労はいらないと考えられる．地名を分析すると同音の地名はあるが，これも少し工夫をすることによって区別が行なえることが判明している．

しかし人名に関しては調査を行なってみたが統計的処理を施こしたものはなかなか見つからず，あるといえば佐久間英氏の佐久間ランキングと各名前の推定人口数があるだけであった．また，この推定についても十分な検定が行なわれたものでもなかった．また各会社で使用しているカナ文字を，漢字の名前へ変換することは誰も考えつかなかったような状況であった．

さらに調査を行なっているとたまたま東邦計算センターの役員である葛西隆三氏から，学習研究社でカナ文字漢字の変換が可能であると話しているということを聞いた．さっそく学習研究社の後藤榛男氏に会い，その内容を聞いてみた．

その原理はむずかしいものではなく，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，人間の介入によって最終的に「田中」を選択する方法であった．われわれはすべ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のである．

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わらないではないかという意見もあるが，数千の文字の中から2，3文字選ぶのと数個の中から1つ選択するのでは大変な労力の差がある．

また，コンピュータの中にファイルされている顧客情報は膨大な件数であり，これは生きものと同じで常に変化している．つまり訂正，削除，更新，新規登録が行なわれているので過去の申込書とか顧客名簿は参考になってもインプットの資料になりえない等という事情がある．

そこでわれわれはこの大変興味深い方法を研究することにした．このちょっとした契機から「姓名のカナ漢字変換システム」を作成することになったのである．

2. 姓名の種類と頻度

原理はわかっても，全国に姓名はいくつあるか，それはどのような頻度分布をしているのか，等の基本的事項はまったくわかっていない．柳田国男先生が約8～10万ぐらいあると書かれているのはあるが（全集20巻，p. 289），これも統計的な研究を行なって書いたもので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

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.

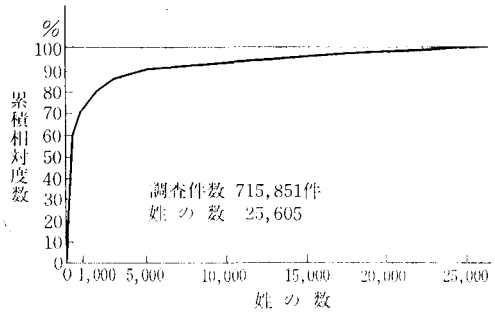


図 1　姓の頻度分布（第百生命）

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

表 2　姓の頻度分布

いファイルがないかと探してみたが，結局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と東邦生命のファイルを借用し，統計的処理を行なってみることにした．生命保険のファイルはほぼ全国を網羅しており，しかも姓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できるという利点もある．ただし，地域別，男女別，年齢別の偏りを除く処理を行なえばよいのであるが，「カナ漢字変換システム」を作成するための判断材料であるため厳密な調整は省略した．

つぎにこの結果を整理したものから個々の名前を分析してみる．

第百生命ファイル

調査件数：715,815

姓の種類：　25,605

この調査で姓の読み方が25,605件と割合少ないことがわかった(**表 1** 参照)．さらにこれを頻度順に並べて分析すると**表 2** の結果が得られた．

表2からわかるように約100個の姓で37%，さらに約5,000個の姓で92%を占めている．これ以後

表 3 頻度上位60の姓（第百生命）

順位	姓	件数	比率(%)	順位	姓	件数	比率(%)
1	サトウ	11,407	1.59	31	エンドウ	2,291	0.32
2	スズキ	11,380	1.58	32	オカダ	2,160	0.30
3	タカハシ	8,687	1.21	33	ムラカミ	2,155	0.30
4	ワタナベ	8,098	1.13	34	ゴトウ	2,106	0.29
5	タナカ	7,341	1.02	35	ナカジマ	2,083	0.29
6	イトウ	6,608	0.92	36	フジタ	2,051	0.28
7	コバヤシ	6,272	0.87	37	キクチ	2,048	0.28
8	サイトウ	6,198	0.86	38	ヤマシタ	2,040	0.28
9	ヤマモト	5,717	0.79	39	アオキ	2,038	0.28
10	ナカムラ	5,711	0.79	40	コンドウ	1,973	0.27
11	カトウ	4,926	0.68	41	アライ	1,971	0.27
12	ヨシダ	4,795	0.66	42	カネコ	1,970	0.27
13	ヤマダ	4,764	0.66	43	タケウチ	1,887	0.26
14	ササキ	3,950	0.55	44	サカモト	1,877	0.26
15	マツモト	3,575	0.49	45	オオタ	1,869	0.26
16	ヤマグチ	3,350	0.46	46	マエダ	1,779	0.24
17	イノウエ	3,336	0.46	47	オノ	1,761	0.24
18	アベ	3,223	0.45	48	ナカノ	1,758	0.24
19	キムラ	3,204	0.44	49	タケダ	1,754	0.24
20	ハヤシ	3,004	0.41	50	ナカガワ	1,753	0.24
21	シミズ	2,744	0.38	51	スギヤマ	1,685	0.23
22	イケダ	2,669	0.37	52	ウエダ	1,670	0.23
23	モリ	2,591	0.36	53	タムラ	1,660	0.23
24	イシカワ	2,497	0.34	54	フクダ	1,660	0.23
25	ヤマザキ	2,489	0.34	55	フジイ	1,659	0.23
26	ハシモト	2,447	0.34	56	ミウラ	1,638	0.22
27	オガワ	2,429	0.33	57	オカモト	1,637	0.22
28	サカイ	2,324	0.32	58	ニシムラ	1,616	0.22
29	ハセガワ	2,321	0.32	59	マスダ	1,596	0.22
30	イシイ	2,303	0.32	60	モリタ	1,588	0.22

表 4 頻度上位60までの姓（第百生命）

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

は姓の件数が増えても全体の割合はあまり増加しないという状況がはっきりつかめる．これをグラフにあらわしてみると**図 1**のようになる．これは累積百分率をグラフにまとめたものである．

この結果，カナ文字の姓から漢字への変換は容易であることがわかる．参考までに頻度上位60の姓をあげてみると**表 3**のようになる．また，これを代表的な漢字の姓であらわしたものが**表 4**である．

第百生命のファイルだけでは一般性がないのではないかという不安からさらにつぎの東邦生命のファイルについても同様の調査を行なってみた．

表 5　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
カキクケコ	2,850	7.03	157,192	5.93
	1,107	2.73	65,642	2.48
	1,207	2.98	68,296	2.58
	172	0.42	1,343	0.05
	1,657	4.09	114,420	4.32
カ行計	6,993	17.26	406,893	15.37
サシスセソ	1,257	3.10	173,697	6.56
	1,778	4.39	90,820	3.43
	709	1.75	88,053	3.32
	475	1.17	21,787	0.82
	347	0.85	9,296	0.35
サ行計	4,566	11.27	383,635	14.49
タチツテト	2,007	4.95	191,411	7.23
	430	1.06	12,095	0.45
	921	2.27	36,432	1.37
	373	0.92	13,861	0.52
	1,381	3.41	43,049	1.62
タ行計	5,112	12.62	296,848	11.21
ナニヌネノ	1,311	3.23	123,452	4.66
	689	1.70	43,225	1.63
	160	0.39	3,964	0.15
	129	0.31	4,772	0.18
	500	1.23	23,439	0.88
ナ行計	2,789	6.88	198,942	7.51
ハヒフヘホ	1,392	3.43	94,808	3.58
	1,052	2.59	49,091	1.85
	1,135	2.80	83,335	3.14
	112	0.27	1,512	0.05
	685	1.69	40,030	1.51
ハ行計	4,376	10.80	268,776	10.15
マミムメモ	1,105	2.72	94,750	3.58
	3,394	8.38	79,657	3.00
	505	1.24	30,638	1.15
	105	0.25	1,121	0.04
	602	1.48	46,418	1.75
マ行計	5,711	14.10	252,584	9.54
ヤユヨ	1,052	2.59	121,772	4.60
	315	0.77	6,839	0.25
	635	1.56	61,407	2.32
ヤ行計	2,002	4.94	190,018	7.17
ラリルレロ	54	0.13	228	0.00
	165	0.40	1,791	0.06
	37	0.09	143	0.00
	21	0.05	38	0.00
	75	0.18	390	0.01
ラ行計	352	0.86	2,590	0.09
ワ	479	1.18	46.911	1.77
ワ行計	479	1.18	46,911	1.77
合計	40,499	100.	2,647,011	100.

東邦生命ファイル

調査件数：2,647,011

姓の種類：　40,499

アイウエオ順の統計は表 5 のようになる．調査件数が増加したため姓の種類は多くなっているが全体的な割合は第百生命の姓の分布とほぼ同様である．さらに，これを累積百分率でグラフを作成すると図 2 のようになる．この図と第百生命の図とを重ねあわせてみるとほぼ一致することがわかる．つまり調査件数が増加しても，全体の傾向は変わらず，珍しい姓の種類が増えただけだということになる．

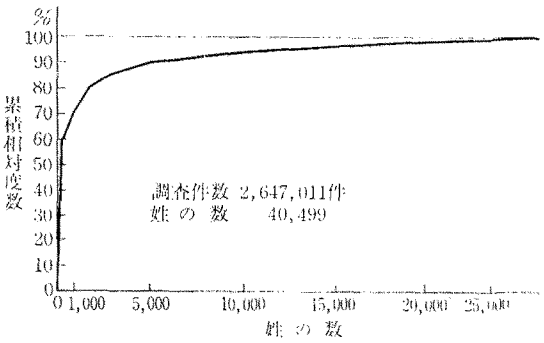


図 2　姓の頻度分布（東邦生命）

出現件数	姓の数
10,000以上	23
5,000～9,999	48
1,000～4,999	373
500～ 999	394
400～ 499	166
300～ 399	275
200～ 299	447
100～ 199	1,112
90～ 99	233
80～ 89	274
70～ 79	326
60～ 69	446
50～ 59	535
40～ 49	745
30～ 39	1,125
20～ 29	1,972
10～ 19	8,569
9	506*
8	1,324
7	657*
6	2,177
5	984*
4	3,987
3	1,703*
2	8,661
1	

表 6　姓の頻度分布（東邦生命）

（注）　被保険者と契約者の両方を調査に加えているため，これらが同一姓であることが多いため偶数が多くなり，奇数が少なくなる．

これを出現頻度順の表にすると表 6 のようになる．これからみても 5,000から 10,000の姓のよび方でほぼ全体を網羅していることがわかる．

参考までに頻度60位までのカナ文字の姓（表 7）と，漢字になおした姓（表 8）をあげておく．この

表 7　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	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	7,564
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	オノ	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	11,951	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	6,289
20	ハヤシ	10,911	50	マツダ	6,257
21	キクチ	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	エンドウ	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	8,595	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

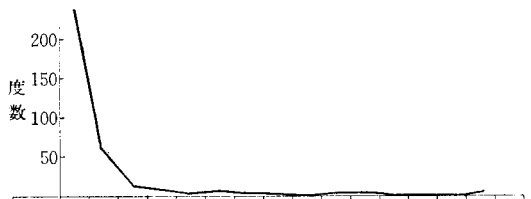
表 8　頻度上位60までの姓（東邦生命）

順位	姓	順位	姓
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	後藤	58	西村
29	石川	59	福田
30	小川	60	岡本

表を第百生命の表と見くらべてみると，「サトウ」「スズキ」「タカハシ」までは同じ順位であるが，それ以後はかなりのくい違いがある．しかし，それも上下数位にずれがあるだけで，あまり大きな変動はない．

これだけの調査で姓の数を推定することは不充分なので別に 623件の名前を調査し，同音異字の姓の発生頻度を調査してみた．これによると一つのよび方について約 1.7個の異なった姓があることがわかる．東邦生命のカナ姓の件数４万件と掛け合わせると68,000件ということになる．このことからみてどんなに少なくみても全国には６万以上の姓があるといって間違いない．

実際には日本人の姓を集めてみると約13万種類の姓があった．単純な推定と実際の間には２倍のくい違いがある．これはカナ姓のデータが人口１億人全部を網羅したものでないし，同音異字のデータ規模が小さかったことによるものと思われる．参考までに同音異字の姓の頻度表を掲げておく．（図 3）



同音異姓の度数	1々	2々	3々	4々	5々	6々	7々	8々	9々	10々	11々	12々	13々	…々	20々	
種類数	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

図 3　同音異字の姓の発生頻度

表 9　名前の分布（第百生命）

	男		女		男		女	
	名前の数	%	名前の数	%	件　数	%	件　数	%
ア	408	2.05	125	0.62	15,915	2.27	7,206	1.02
イ	554	2.78	151	0.75	13,728	1.96	2,648	0.37
ウ	194	0.97	52	0.26	930	0.13	812	0.11
エ	253	1.27	88	0.44	6,693	0.95	5,248	0.74
オ	184	0.92	101	0.50	2,782	0.39	191	0.02
ア行計	1,593	8.01	517	2.60	40,048	5.71	16,105	2.30
カ	854	4.29	234	1.17	34,478	4.96	9,461	1.35
キ	807	4.05	236	1.63	19,166	2.73	12,824	1.83
ク	313	1.57	71	0.35	6,229	0.88	2,113	0.30
ケ	364	1.83	60	0.30	16,871	2.40	3,825	0.54
コ	471	2.36	146	0.73	14,605	2.08	976	0.13
カ行計	2,809	14.13	747	3.75	91,349	13.04	29,199	4.17
サ	534	2.68	160	0.80	12,791	1.82	7,324	1.04
シ	1,298	6.52	320	1.60	46,324	6.61	8,032	1.14
ス	307	1.54	129	0.64	5,919	0.74	4,085	0.58
セ	393	1.97	64	0.32	10,349	1.47	3,061	0.43
ソ	187	0.94	49	0.24	1,759	0.25	358	0.05
サ行計	2,719	13.67	722	3.63	77,142	11.01	22,860	3.26
タ	992	4.99	224	1.12	45,308	4.67	6,413	0.91
チ	411	2.06	186	0.93	4,036	0.57	6,639	0.94
ツ	357	1.79	132	0.66	8,586	1.22	2,470	0.35
テ	300	1.50	67	0.33	9,731	1.38	2,728	0.38
ト	853	4.29	210	1.05	32,691	4.66	11,417	1.63
タ行計	2,913	14.65	819	4.11	100.352	14.33	29,667	4.23
ナ	362	1.82	111	0.55	3,872	0.55	2,379	0.33
ニ	97	0.48	22	0.11	304	0.04	24	0.01
ヌ	12	0.06	5	0.02	18	0.01	56	0.01
ネ	15	0.07	1	0.01	48	0.01	1	0.—
ノ	228	1.44	64	0.32	14,038	2.00	4,088	0.58
ナ行計	714	3.59	203	1.02	18,281	2.61	6,548	0.93
ハ	330	1.66	122	0.61	8,206	1.17	5,639	0.80
ヒ	581	2.92	153	0.76	42,340	6.04	9,053	1.29
フ	426	2.14	142	0.71	7,590	1.08	6,939	0.99
ヘ	85	0.42	6	0.03	866	0.08	11	0.01
ホ	106	0.53	33	0.16	235	0.03	46	0.01
ハ行計	1,528	7.68	456	2.29	59,237	8.45	21,742	3.10
マ	550	2.76	163	0.81	41,325	5.90	9,050	1.29
ミ	540	2.71	245	1.23	21,243	3.03	17,531	2.50
ム	151	0.75	33	0.16	1,481	0.21	514	0.07
メ	37	0.18	22	0.11	287	0.04	235	0.03
モ	264	1.32	53	0.26	3,639	0.51	751	0.10
マ行計	1,547	7.78	516	2.59	67,975	9.70	28,081	4.01
ヤ	289	1.45	93	0.46	11,531	1.64	4,047	0.57
ユ	256	1.28	84	0.42	14,132	2.01	6,084	0.86
ヨ	503	2.53	134	0.67	33,372	4.76	9,945	1.42
ヤ行計	1,048	5.27	311	1.56	59,035	8.43	20,081	2.86
ラ	20	0.10	12	0.06	36	0.01	59	0.01
リ	345	1.73	83	0.41	6,448	0.92	1,895	0.27
ル	12	0.06	13	0.06	37	0.01	337	0.04
レ	56	0.28	21	0.10	526	0.07	1,695	0.24
ロ	38	0.19	4	0.02	388	0.05	30	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

3.　名前の種類と頻度

さらに名前について同様の統計を行なった．名前は世襲のものでなく個人にとって1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．

表10 名前の頻度分布

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
～19,885	100.00

第百生命ファイル

調査件数：700,201

名前の種類： 19,879

名前の種類は姓の種類よりも少ないのである．

つぎにアイウエオ順，男女別に整備した**表9**を示す．この表からわかるように生命保険のファイルは男性が多く女性が少ないため補正を行なって比較をしなければならない．さらに頻度順に名前を並べ名前の累積百分率を作成すると**表10**になる．約3,000の名前で93%，10,000で98%になる．親たちは子供に立派になってほしいと知恵をしぼるにもかかわらず，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．

さらにこれをグラフにあらわしてみると**図4**のようになる．このグラフを姓のグラフと比較してみると名前のグラフのほうがカーブが急であることがわかる．東邦生命のファイルについて行なってみたが結果はほぼ同じものであった．

名前を実際に集めてみると約15万種類あった，まだ未整理のデータが15万件程度あるので約30万近くになる予定である．名前の同音異名の発生頻度は平均10件程度と思われる．

以上紹介してきた統計からわれわれの研究の第1段階として日本人の姓名についてカナ文字→漢字変換はかなり規則的に行なえることがわかる．しかしこれまでの研究はカナ漢字変換システムの実用化までの約1割程度のものである．

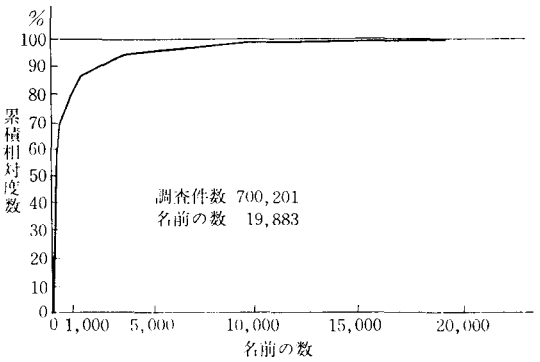


図 4 名前の頻度分布（第百生命）

4. 姓名カナ漢字変換システムの実用化

つぎに「姓名カナ漢字変換システム」実用化までを簡単に述べてみる(**図 5**)．

第2段階として考えなければならないことはデータの収集と整理である．これは大変な忍耐のいる作業である．さらに大変な作業は漢字システムへ入力されたデータの校正である．

第3段階として問題になることは同音異字の姓や名前を使いやすくするため，これらの中で使用頻度の高い順序に整えることである．25万人の名の頻度を調べたものを紹介しよう．たとえば「アキオ」を取りあげてみると，

順位	読み	漢字	頻度	順位	読み	漢字	頻度
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	89	21		明郎	4
6		昭雄	88	22		暁雄	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つぎのようになる．

安起郎，安喜男，安喜夫，暉雄，暁生，暁男，暁夫，顕男，顕夫，顕郎，晃夫，晃郎，皓男，皓郎，秋尾，穐雄，昭大，昭尾，昭朗，照夫，昌雄，晶男，晶雄，彰百，彰郎，彰朗，璋夫，楢男，旦夫，哲男，哲郎，斌夫，斌郎，明応，爺夫，燿雄，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片寄りがあることは興味のあることである．

第4段階の問題点はこの変換方式を実際に試行し多くの人々を説得し，しかも直接入力方式よりも安く，早く，正確に，取り扱いやすくできることを示すことである．

さらにこの方式で行なうユーザを見つけることであった．このような説得と宣伝がこのシステムをつくる最大の仕事であり，分析作業，実際のシステム作りは大きな問題ではなかった．

第5段階としてはシステム開発，変換精度の向上，効率よいシステム改良である．この方式が役立つとわかるとはじめは変換率がせいぜい80～90％であれば充分といっていた人がもっと精度を上げるように要求してくるものである．このような人たちの要求，心理的変化にも気を配らなければならない．しかも，80～90％から95～98％へ性能を上げるにはいままでに使った労力とほぼ同じ程度の努力をしなければならないのである．

われわれの変換精度はつぎのとおりである．

	マスター件数	変換率	用紙に表示する件数
姓	13万件	98％	5件
名	30万件	90～95％	15件

変換に使う用紙に表示できる件数に制限があるためこれ以上の変換率は用紙の表示件数を変える等の方法をとらなければならないであろう．

このように姓名について，いろいろな観点から調査を行なってみると，おもしろい結果が得られる．今後は男女の名の統計を年代別にとって時代

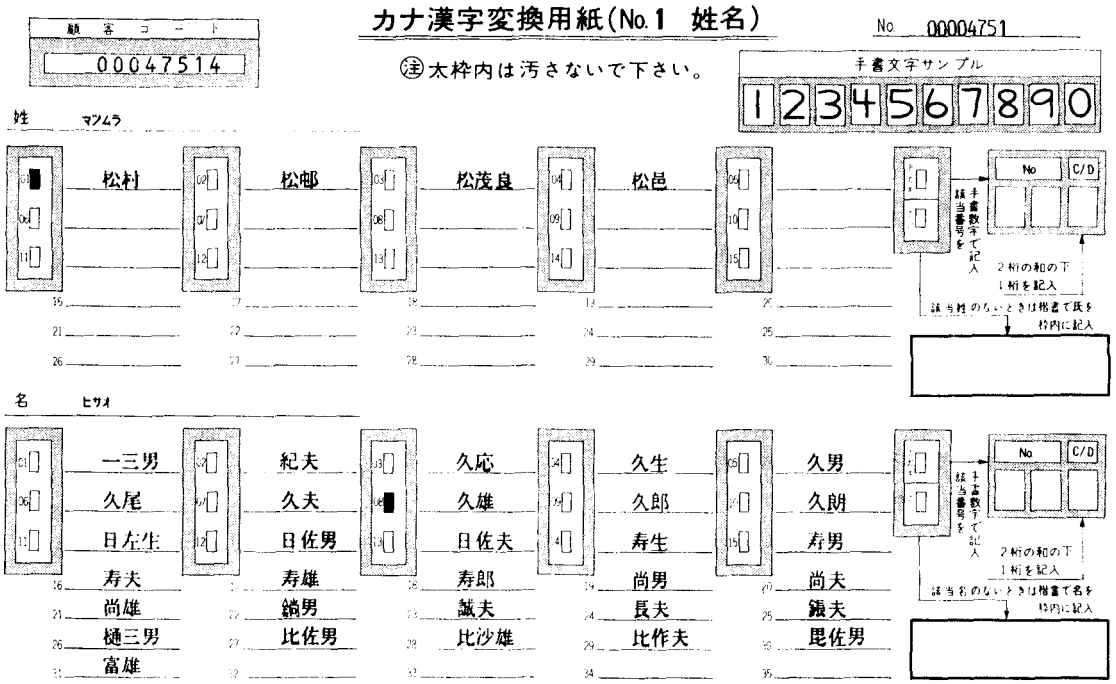


図 5　カナ漢字変換用紙

的特徴があるかないか調べたり，姓の地域的分布を調べたりしていきたいと思う．

同時に姓名についてのより詳細な研究がなされていくことを期待したい．姓名の研究といえば民族的研究や占いの対象としての研究しかなかったが，数量的に取り扱うことにより科学の分野への道を開くことができる．

最後に，この研究について支援してくださった国立国語研究所林大所長，またファイルを提供してくださった第百生命，東邦生命の方々に深く感謝したい．また，ORに興味をもたれる方々が一見なんでもないと思われている事柄を調査・研究し新しい分野を進まれることを期待する．

参考文献

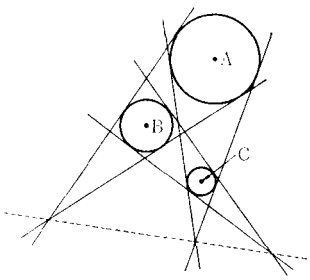
田中康仁，日本人の姓と名に使われる漢字
日本ユニバック(株)，姓・名に使われる漢字
昭和51年10月．

たなか・やすひと　1939年生
1962年　岡山大学教育学部卒　数学専攻
1962年　日本ユニバック入社
最近の仕事，漢字システムの研究・開発

フォーラム

数理パズルを楽しもう (8)

問題　半径が異なる3つの円A，B，Cを図のようにかき，円Aと円B，円Aと円C，円Bと円Cのそれぞれについて，共通外接線の交点を求めました．すると，3つの交点はピタリと一直線上にのっているようです．偶然ではないようなのでその理由のうまい説明を考えてみてください．



〔5月号（337ページ）の解答〕　一般に，7で割ったときの余りをa，11で割ったときの余りをb，13で割ったときの余りをcとし，

$$n=715a+364b+924c$$

を計算する．このnを1001で割った余りが，花子さんの考えた数である．5月号の出題では，$a=3$，$b=2$，$c=1$であるから，$n=3797$となり，花子さんの考えた数は794となる．つまり，タネになる3つの数は，715と364と924だったのである．

この理由は，以下のようである．7，11，13はすべて素数であるから，もちろん互いに素である．よって，たとえば7については，

$$143u+7v=1,\quad (143=11\times 13)$$

を満たす整数u，vの組が存在する．uの最小の正整数は5で，143に5を掛けた数が715である．この作り方から，715を7で割ると1が余り，11と13で割ると割り切れる．同様にして，364は11で割ると1が余り，7と13では割り切れる数，924は13で割ると1が余り，7と11では割り切れる数である．$7\times 11\times 13=1001$であるから，$n=715a+364b+924c$ を1001で割った余りが求める数となるのである．

なお，この種の数当て遊戯は百五減算といって，その発祥は東洋にあるとされていた〔1〕．ところが，筆者の調査によると，それより古い西洋の数学書〔2〕に，すでに同種の問題が紹介されていた．

〔1〕平山諦，東西数学物語，恒星社，1973.
〔2〕Bachet de Meziriac, C. G., *Problèmes plaisants et délectables*, 1612, (Albert Blanchard 社から，1959年に複刻版が出ている)．

（中村義作　信州大学工学部）

FORUM