日本の苗字の計量的分析

梅田 三千雄†

日本の苗字が備えている種々の性質を明らかにすることを目的として，苗字データベースを作成し，その計量的分析を行った．ここでは，より普遍的なデータの収集をねらいとして，約7.1万個から成る日本の苗字データベースを作成した．このデータベースをもとに，苗字に出現する文字の種類や頻度，文字位置とそこに出現する文字の種類など，文字と文字連接に関する統計データを求めた．これより，日本の苗字には文字位置によって，出現する文字の種類とその頻度に大きな偏りのあることが明らかになった．さらに，実際の使用頻度を考慮した分析として，市販の電話帳データベースを利用した検索により，苗字の使用頻度，苗字ならびに文字と文字連接のエントロピーなどを測定した．これより，苗字のエントロピーは英単語のそれにほぼ等しいことが明らかになった．また，ここで得られた苗字の諸性質は，宛名や個人情報の文字認識において，苗字部分の文字切り出しでの知識として利用したり，認識対象文字の種類を決定，限定したりするのに利用することが可能であり，認識精度の向上につながることが期待される．

Metrical Analysis of Japanese Family Names

MICHIO UMEDA†

In this paper, Japanese family names database is constructed and several characteristics of Japanese family names are extracted from this database to be utilized in the process of characters recognition. This database contains 71452 kinds of Japanese family names. For example, one to six characters are used in family names and 80% of names consist of two characters. All Japanese family names are composed of 3796 character categories. There are 1400 character categories which are used more than 10 times in the names. When 1000 character categories are selected in order of appearance frequency, the rate of those characters used in the names is to be 92%. The 84% of all the family names are perfectly constructed by high frequency 1000 characters. Furthermore, by accessing Japanese telephone numbers database, some characteristics of family names considered the usage frequency are extracted samely. From these metrical analysis, the lack of precision in the pattern recognition algorithm can be recovered by using such characteristics of Japanese family names.

1. はじめに

日本の苗字（姓氏）には，一千数百年に及ぶ長い歴史がある．しかし，江戸時代までは，苗字は一部の人たちのものであった．それが，明治8（1875）年の「苗字必称令」によって国民皆苗となり，幕末には3万種程度であったものが一気に3倍ほどに増大した．これが明治新姓である．現在では，読みの違いも含めると12万種以上の苗字が存在するとされている[1]．これは，数の多さにおいて世界最大である．ちなみに，同じ漢字を使用している中国では約500種，朝鮮ではその半分の約250種程度であり，いかに日本の苗字が多いかがよく分かる．欧米では，合わせても5万種前後と推定されている[2]．

日本の苗字は，1文字または複数文字の文字連接によって構成されている．本論文では，この日本の苗字について，文字の出現や文字連接に関するいくつかの規則を抽出し，苗字の持つ性質を明らかにすることを目的とする．具体的には，できるだけ普遍的なデータが得られることを目的として，できるだけ多くの苗字を収集した苗字データベースを作成する．次に，このデータベースを検索して，苗字に使用される文字やされない文字，苗字のある位置にだけ出現する文字，ある文字に連接する文字やしない文字など，苗字に使われる文字の諸性質を定量化する．さらに，市販の電話帳データベースを利用して，我が国で実際に使用されている頻度を考慮した苗字における文字の諸性質を明らかにする．

苗字や住所などの文字認識は，郵便物の宛名認識や

† 大阪電気通信大学情報工学部
Faculty of Information Science and Technology, Osaka Electro-Communication University

表 1 日本の苗字の長さと個数
Table 1 The number of Japanese family names of each character length.

苗字の長さ	苗字の個数
1	1798
2	55617
3	13583
4	439
5	12
6	3
合計	71452

表 2 苗字に使用される文字の種類
Table 2 Character categories used in Japanese family names.

文字の属性	個数
記号	3
平仮名	27
片仮名	16
第 1 水準漢字	2413
第 2 水準漢字	976
その他の漢字	361
合計	3796

個人情報の記載伝票の認識などへの応用が考えられる．しかし，たとえば郵便物の宛名認識では，常用手書きの漢字認識の技術的な難しさに加えて，

- まったくの自由書式である，
- 筆記用具が千差万別である，
- 認識対象文字の種類を限定できない，

などの難しさがあり，これに取り組んだ研究[3)~5)]でも必ずしも十分な認識精度を得るまでには至っていない．事実，郵便物の自動区分けでは，7 桁に拡張した数字認識により対処しているのが現状である．

ここでは，日本の苗字の計量的分析において，このような文字認識に組み込んで，精度向上の手助けになると推測される諸性質を求めることをあわせて目的とする．たとえば，自由書式の文字列から苗字の部分を有効に切り出すのに利用できる性質，認識対象文字の種類を決定したり，限定したりすることのできる性質などをデータベースの検索と分析によって抽出していく．

2. 日本の苗字データベースの作成

日本の苗字データベースの作成に際しては，できるだけ普遍的なデータが得られることを目的として，できるだけ多くの苗字を収集することにした．日本の苗字を収集し，記載した書物は数多いが，ここでは最も豊富な日本ユニバック（編）「日本の苗字（表記編）」[2)]に基づいてデータベースを作成した[6)]．なお，この本で沖縄編に記載の苗字は，重複が多いので除外した．

データベースに登録した苗字の個数を**表 1** に示す．全苗字数は 71452 個である．長さ別では，2 文字のものが最も多く，約 8 割を占めている．平均の長さは 2.18 文字である．ちなみに，5 文字苗字には，

一尺八寸山　八月一日宮　八月十五日　十二月一日
十二月晩日　十二月晦日　十二月朔日　六月一日宮
次五右衛門　勘解由小路　釈迦牟尼仏　御措使河原

があり，6 文字苗字には，

一尺二寸五分　八月三十一日　左右衛門三郎

が存在した．

なお，この苗字データベースの容量は約 450 KB である．

3. 日本の苗字データベースの分析

3.1 苗字に出現する文字の種類

日本の苗字に出現する文字の種類を文字の属性別に分類して求めたものを**表 2** に示す．

日本の苗字に使用される文字は 3796 種である．属性別では，JIS 第 1 水準の漢字が最も多く，全体の約 64% を占めている．しかし，日常よく使用する文字として規定した第 1 水準の漢字でも，すべてが苗字に使用されるのではなく，約 550 種は苗字に使用されない漢字であることが分かる．苗字には，第 1 水準漢字の 81.4%，第 2 水準漢字の 28.8% が使用されている．

これにより，苗字の認識における対象文字の種類を限定することができる．なお，表において，その他の漢字とは，第 1，第 2 水準のいずれにも含まれない漢字であり，361 種のものが存在した．これらの漢字は，どのように符号化するか，表現するかなど，認識とは別の問題を引き起こすこともあるだろう．

3.2 苗字に使用される文字の出現頻度

苗字に使用される文字について，出現頻度の高い順に上位 100 位まで求めたものを**表 3** に示す．なお，ここでの出現頻度とは，苗字データベース全体において，その文字が出現する回数のことである．

この結果，日本の苗字に最も多く出現する，つまり使用される文字は「田」であり，出現頻度は 3559 回に及ぶことが分かる．これはデータベースの総文字数の 2.3% に相当する．次いで，「野」，「谷」，「井」，「川」，「木」，「山」，「小」の順である．これら 8 種類の文字で，合計の出現頻度は 1.6 万個余りとなり，総文字数の 10% 以上を占めている．

逆に，ただ 1 回しか出現しない文字は 1012 種が存在した．このうち，269 種は，第 1 および第 2 水準のどちらでもない，つまりその他の漢字である．これ

表 3 苗字に出現する頻度の高い文字
Table 3 High frequency character categories used in Japanese family names.

順位	文字	出現頻度	順位	文字	出現頻度
1	田	3559	51	保	536
2	野	2449	52	和	534
3	谷	1875	53	子	532
4	井	1822	54	生	528
5	川	1762	55	安	527
6	木	1735	56	吉	526
7	山	1706	57	水	521
8	小	1452	58	永	514
9	大	1328	59	加	507
10	原	1286	60	元	504
11	本	1169		倉	504
12	上	1090	62	家	503
13	中	1023		八	503
14	久	988	64	日	493
15	島	929	65	根	490
16	沢	909	66	名	483
17	村	895	67	神	480
18	内	889	68	波	479
19	屋	884	69	城	467
20	尾	881	70	金	465
21	下	880	71	寺	461
22	津	871	72	美	459
23	戸	856	73	志	443
24	部	844	74	道	432
25	三	791	75	馬	418
26	佐	782	76	東	416
27	岡	756	77	竹	414
28	崎	746	78	出	411
29	西	726	79	阿	410
	石	726		浦	410
31	見	725	81	富	408
32	高	699	82	土	405
33	瀬	692	83	橋	403
34	江	683	84	ノ	401
35	藤	681	85	宇	399
36	古	664		良	399
37	賀	644	87	北	397
38	間	642	88	海	395
39	松	639	89	塚	391
40	多	626	90	一	387
41	森	587	91	岩	380
42	口	583		林	380
43	地	574	93	之	379
44	伊	573	94	池	376
45	平	571	95	武	373
46	宮	567	96	新	372
47	坂	565	97	羽	371
48	矢	559	98	丸	366
49	河	547		前	366
	長	547	100	須	363

らの文字を含む苗字は，仮にこの文字だけが正しく認識できれば，苗字全体を決定することができる性質を持っていることになる．さらに，出現頻度が 2 回だけの文字は 420 種であった．この 2 つで，苗字に使用される文字の種類の 4 割近くに達している．

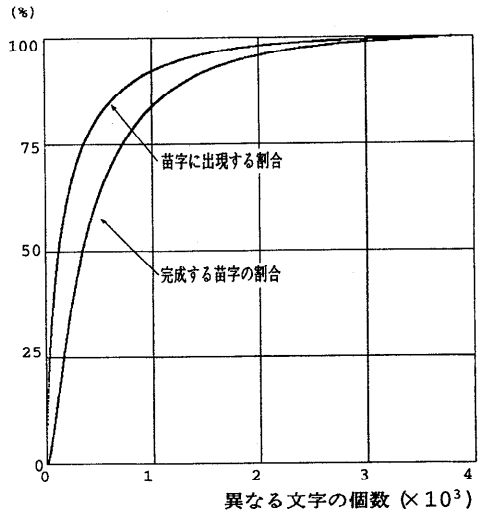

図 1 苗字に出現する割合と完成する苗字の割合
Fig. 1 The rates of characters appeared in family names and family names constructed by such characters.

なお，苗字のどこかに 100 回以上出現する文字は 346 種，50 回以上のものは 548 種，20 回以上は 973 種であり，10 回以上出現する文字は 1400 種であった．ここまでの文字で，苗字に使用される文字の 36.9% になる．

次に，苗字に使用される文字を出現頻度の高い順に並べ，上位から一定個数までをとったとき，これらの文字が苗字の中に出現する累積の割合を求めたものを図 1 に示す．これより，上位から 120 位までの文字をとれば累積出現割合は 50% に達し，500 位までで 81.6%，1000 位までとれば 92.3% になることが分かる．

3.3 文字の種類と完成する苗字の関係

出現頻度の高い順に上位から一定個数の異なる文字をとったとき，そこに含まれる文字の組合せによって苗字が完成する割合について検討する．

たとえば，最も出現頻度の高い文字である「田」だけをとると，これによってできる苗字は「田」の 1 個だけである．次に，第 2 位の「野」までとると，その組合せにより，新たに「野」,「野田」,「田野」の苗字ができる．つまり，上位 2 位までの文字により，合計 4 個の苗字が完成する．同様に，第 3 位の「谷」まで取ると，さらに「谷」,「谷田」,「田谷」,「谷野」,「野谷」,「谷谷」ができ，3 文字苗字では「田谷野」,「野田谷」ができる．つまり，上位 3 位までの文字により，合計 12 個の苗字が完成することになる．

このようにして，出現頻度の上位から一定個数の文

表 4 各文字位置に出現する文字の種類数
Table 4 The number of character categories used at each character position.

苗字長	全体	文字位置					
		1	2	3	4	5	6
1	1798	1798					
2	3477	2899	2505				
3	1743	1166	1147	784			
4	384	134	178	180	143		
5	34	8	7	8	11	7	
6	16	3	3	3	3	3	3

字をとったとき，それらの文字の組合せによって完成する苗字の割合を求めたものを図 1 に示す．

これより，日本の苗字は，上位から 200 位までの文字によって 30.8% が完成し，500 位までで 63.2%，1000 位までとれば 83.9% のものが完成する．逆に，苗字全体の 90% を完成させるためには 1331 種類の文字を選択しておく必要があり，95% を完成させるには 1852 種の文字が必要であることが分かる．

3.4 文字位置と文字の種類の関係

苗字の長さによって，あるいは同じ長さでも苗字内の位置によって，そこに出現する文字は異なることが予想される．たとえば，1 文字苗字に使用される文字，あるいは 2 文字や 3 文字苗字の先頭または末尾にだけ使用される文字などが存在するであろう．

苗字の長さごとに，苗字内の各文字位置に出現する文字の種類数を求めたものを**表 4** に示す．

まず，1 文字苗字に使用される文字は 1798 種であり，これは使用される文字の 47.4% に当たる．つまり，1 文字苗字に使用される文字の種類は，全体の半分弱である．

次に，2 文字の苗字において，その先頭位置に出現する文字は 2899 種であり，末尾に出現するものは 2505 種である．つまり，先頭に出現する文字は，末尾に出現するものに比べて，400 種近くも多くなっている．また，2 文字の苗字全体で使用される文字は 3477 種である．それゆえ，2 文字苗字の先頭にだけ使用される文字は 972 種，末尾にだけ使用される文字は 578 種であり，残りの 1927 種の文字は両方に使用されていることになる．なお，2 文字苗字には，苗字に用いられる文字の 91.6% が使用されている．

3 文字の苗字には，1743 種の文字が使用されている．これは，1 文字苗字に使用される文字の種類より少ない．また，先頭と中間の位置において使用される文字の種類がほぼ同じであるのに対して，末尾位置に使用される文字は 784 種であり，前の 2 つに比べて 400 個近くも少ないことが分かる．

表 5 各文字位置における出現頻度の高い文字
Table 5 High frequency character categories used at each character position of 2- & 3-length family names.

苗字の長さ	文字位置					
	1		2		3	
	文字	出現頻度	文字	出現頻度	文字	出現頻度
2	大	744	田	1336		
	小	545	野	952		
	中	438	川	847		
	高	402	谷	846		
	三	387	山	829		
	山	348	本	740		
	田	332	原	704		
	上	322	井	687		
	西	316	木	677		
	古	296	村	582		
3	小	749	野	622	田	1276
	大	487	田	389	谷	626
	三	310	ノ	379	川	510
	伊	272	之	342	野	476
	久	261	久	297	井	413
	上	245	ケ	259	原	397
	佐	208	井	249	木	385
	下	202	木	233	山	381
	中	188	津	221	部	335
	加	187	賀	212	内	292

3.5 文字位置と文字の出現頻度の関係

2 文字と 3 文字の苗字について，各文字位置でみた出現頻度の高い文字の種類を**表 5** に示す．いずれも上位 10 種の文字を示した．

これより，2 文字および 3 文字苗字の先頭位置には，「大」,「小」,「中」,「高」,「上」,「三」,「古」など，状態を表す文字が多く使用されていることが分かる．これに対して，末尾位置には,「田」,「谷」,「野」,「川」,「山」,「原」,「井」など，物を意味する文字が多く使用されている．また，3 文字苗字の中間位置では,「ノ」,「之」,「ケ」など，文字と文字の接続に用いられる特殊な文字の出現頻度が高い．

また，2 文字苗字の末尾位置において，表 5 の 10 種類の文字のどれかが出現する割合は 14.7% であり，先頭位置における割合のほぼ 2 倍である．3 文字の苗字でも，表に示した 10 種類の文字が苗字に出現する割合は，末尾位置において最も高く，37.5% に達している．つまり，3 文字苗字の末尾では，3 つに 1 つはこれら 10 種類の文字のどれかであることになる．このように，2 文字および 3 文字苗字の末尾位置には，少数の種類の文字が集中して使用される性質のあることが分かる．

この結果より，苗字などの文字認識において，仮に苗字の場所が確定できたとすれば，まず末尾位置の文字を認識し，その結果をもとに連接する可能性のある

表 6 苗字の範囲を決定できる文字の個数
Table 6 The number of character categories which can determine the range of Japanese family name.

苗字長	文字位置					
	1	2	3	4	5	6
1	192					
2	581	432				
3	39	46	20			
4	0	2	1	2		
5	0	1	0	0	0	
6	0	0	0	0	0	0

表 7 範囲が決定できる苗字の個数
Table 7 The number of Japanese family names those range can be determined by some specific character.

苗字長	文字位置					
	1	2	3	4	5	6
1	192					
2	960	568				
3	47	156	25			
4	–	2	1	2		
5	–	1	–	–	–	
6	–	–	–	–	–	–

文字の種類を候補文字と限定して，他の位置の文字を認識していけば，認識精度の向上に結び付けていけることが期待される．

3.6 苗字範囲を決定できる文字の種類

日本の苗字において，特定の長さの特定の文字位置でだけ使用されるという性質の文字が存在すれば，逆に，その文字を手がかりとして，正確な苗字の範囲を決定することができる．つまり，この文字を手がかりとして，どの文字からどの文字までが苗字の範囲であるかを知ることが可能となる．

たとえば,「蝦」という文字は，2 文字苗字においてだけ使用され，かつその先頭位置においてだけ使用される．他の長さの苗字や 2 文字苗字の末尾には，決して使用されることがない．したがって，この文字を認識することができれば，それを手がかりに，この文字から 2 文字苗字が始まり，次の文字までが苗字の範囲であると知ることができる．同様に,「袈」という文字は，3 文字苗字の先頭位置でしか使用されないため，この文字から始まる 3 文字が苗字の範囲として決定できる．

このように，ある特定の長さの苗字の特定の位置でだけ使用されるという性質を利用して，苗字の範囲を決定することのできる文字の種類を求めたものを**表 6** に示す．この性質を備えている文字は 1316 種が該当し，苗字全体に使用される文字の種類の約 35% に達している．つまり，1/3 以上の文字は，ある長さの苗字の，しかもある特定の位置でしか使用されないことが分かる．

さらに,「蝦」という文字によって，苗字の範囲を決定できるものには「蝦原」,「蝦名」,「蝦塚」,「蝦夷」,「蝦島」,「蝦田」の 6 個の苗字がある．同様に,「袈」によって,「袈裟栗」,「袈裟丸」,「袈裟寺」の 3 個の苗字の範囲を決定することができる．このようにして，苗字の範囲を決定できる文字を含む苗字の種類を求めたものを**表 7** に示す．

これより，ある特定の文字によって，範囲を決定することのできる苗字の種類は，わずかに 1954 種であり，苗字全体の 2.7% にすぎないことが分かる．範囲を決定できる文字が全体の 1/3 以上であったのに対して，それを含む苗字の割合はきわめて少ない．つまり，特定の文字位置でだけ使用されるという性質を持つ文字は，日本の苗字全体においても使用されることの少ない，特殊な文字であるといえよう．

なお，本章で得られた性質は，苗字に限定したものであり，郵便物の宛名認識のように住所と氏名が混在した場合には，他の部分にも同じ文字列が現れる可能性がある．

4. 使用頻度を考慮した日本の苗字の分析

これまでの分析は，実際に使用されている頻度をまったく考慮しない，日本の苗字そのものに出現する文字の性質に関するものであった．しかし，実際には，苗字によって使用されている頻度が大きく異なっている．たとえば,「佐藤」や「鈴木」を苗字とする人は非常に多いが，一方では，他にはだれもいない苗字の人もいるであろう．ここでは，このような使用頻度を考慮した日本の苗字について分析する．

1 人 1 人の苗字を正確に把握し，データベース化することは，きわめて大規模な作業が必要であり，ほとんど不可能である．本研究では，全国規模の電話帳を取り上げ，そこでの出現件数が近似的な日本の苗字の使用頻度を反映しているものとして，頻度を考慮した分析を行った．具体的には，市販されている電話帳検索ソフトウェアを利用して，前述のデータベースに登録されている苗字を入力して検索し，その出現件数を求めて苗字の使用頻度とした．

4.1 使用頻度を考慮した苗字の規模

このようにして検索できた苗字の種類と使用頻度を**表 8** に示す．異なる苗字の種類は 54704 個であった．前述のデータベースに登録のものより，かなり少なくなっている．これは，表示できない漢字が適当に統合されていることにもよるが，今日では使用されていな

表 8 苗字の種類と使用頻度
Table 8 The number of Japanese family names and those total usage frequency.

苗字の長さ	苗字の種類	使用頻度
1	1383	1089541
2	44039	28785419
3	9114	1214636
4	167	4820
5	1	1
6	0	0
合計	54704	31094417

表 9 使用頻度の高い日本の苗字
Table 9 High usage frequency Japanese family names.

順位	苗字	使用頻度
1	佐藤	481980
2	鈴木	426804
3	高橋	353911
4	田中	334073
5	渡辺	276257
6	伊藤	270047
7	山本	269344
8	中村	264448
9	小林	254807
10	加藤	214905

い苗字がかなり存在することを示している.

検索できた日本の苗字の総数は約 3100 万個で，そのうちの 92.6% が 2 文字苗字である．頻度を考慮した平均の苗字長は 2.00 文字となり，これからも 2 文字苗字に集中していることが分かる．なお，検索できた 5 文字苗字は「勘解由小路」のみで，使用頻度も 1 個だけであった．6 文字苗字は存在しなかった．

また，検索できた苗字に含まれている異なる文字の総数は 2638 種であった．前述のデータベースのものより，かなり少なくなっている．これより，実際の苗字を対象とする文字認識では，認識対象をかなり限定することも可能であることが示唆される．

4.2 日本の苗字の使用頻度

日本の苗字について，その使用頻度の高い 10 種と頻度を求めたものを**表 9** に示す．

これより，最も多く使用されている苗字は「佐藤」であり，全体の 1.55% を占めている．ほぼ 65 人に 1 人がこの苗字である．次いで,「鈴木」,「高橋」の順である．上位 10 種はすべて 2 文字苗字であり，ここまでで日本の苗字の 10.1% を占めている．10 人のうち 1 人は，これらの苗字のどれかを使用していることになる．なお，頻度の高い 1 文字苗字は,「林」が 20 位に，次いで「森」が 27 位に出現する．3 文字苗字では,「佐々木」と「長谷川」が各々13 位と 25 位に位置している．

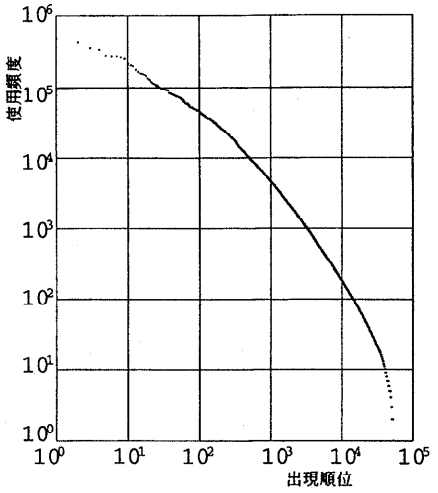

図 2 苗字における出現順位と使用頻度の関係
Fig. 2 The relation of appearance order to usage frequency in Japanese family names.

次に，頻度を考慮した日本の苗字について，苗字を使用頻度の高い順に並べ，その出現順位と使用頻度との関係を求めたものを図 **2** に示す．

日本語を対象とした適当な例はないが，ブラウン・コーパスを題材とした英文における単語の出現順位と頻度では，互いに反比例の関係にあるとするジップの法則が成り立つことが知られている[7]．つまり，もし図 2 に記入すれば，英単語の使用頻度は右下がりの直線上にきれいに並んで分布することになる．

これに対して，本結果は上に凸の分布をしており，日本の苗字では，頻度の高いものが集中して出現し，逆に頻度の低いものの出現は極端に少なく，必ずしもジップの法則には従わないことが分かる．ちなみに，頻度の順に上位から 500 位までの苗字をとると，累積の頻度は 60.1% になり，1000 位までとると 70.9%，さらに 2000 位までは 80.2% に達する．つまり，1% 弱の苗字で全体の 6 割をカバーするほど集中したものになっている．

4.3 日本の苗字のエントロピー

情報の曖昧さ，無秩序さを表す尺度にエントロピー（平均情報量）がある．エントロピーは，1 つの事象が生起するごとに得られる情報量の期待値で定義され，完全事象系 E における個々の事象の生起確率を p_i とすれば，

$$H(E) = -\sum p_i \log_2 p_i$$

によって与えられる．

シャノンは，文字の出現頻度を用いて英語のエントロピーを推定することを試みた[8]．まず，文字が等確

表 10 苗字に使用される頻度の高い文字と 2, 3 文字の文字連接
Table 10 High usage frequency character categories and 2- & 3-character connections in family names.

順位	1 文字	苗字数	出現頻度	2 字組	出現頻度	3 字組	出現頻度
1	田	2862	4644263	佐藤	482037	佐々木	179096
2	藤	606	2259886	鈴木	426932	長谷川	92716
3	山	1404	2197074	高橋	353911	五十嵐	40238
4	野	2030	1681694	田中	335179	久保田	37398
5	川	1456	1626348	渡辺	276257	大久保	33928
6	本	1040	1550116	伊藤	270054	小野寺	23494
7	村	790	1484042	山本	269345	小笠原	23152
8	井	1537	1451489	中村	264546	佐久間	21774
9	中	877	1437582	小林	254811	長谷部	7319
10	木	1380	1425415	加藤	215608	宇佐美	6883
11	小	1162	1226978	山田	209629	阿久津	6557
12	原	1076	1045894	吉田	208797	小野田	6518
13	大	1005	926830	佐々	184376	宇都宮	6366
14	松	552	873838	々木	179603	大和田	6359
15	佐	574	848855	斎藤	166634	波多野	6324
16	高	589	841779	山口	160384	海老原	6169
17	谷	1644	826828	松本	158588	小久保	5685
18	島	749	709340	井上	152489	日比野	5443
19	岡	680	706799	久保	149539	宇田川	5183
20	上	885	706715	木村	146018	日下部	4882
21	橋	403	667709	清水	133789	小山田	4204
22	崎	742	660718	山崎	121649	大河原	3974
23	口	577	647049	小野	117979	小山内	3911
24	沢	895	619799	池田	115481	小田島	3811
25	石	724	615370	阿部	114872	小宮山	3722
26	西	723	598364	橋本	112041	加賀谷	3470
27	吉	523	531033	長谷	111023	竹之内	3425
28	林	376	514263	山下	108760	仲宗根	3230
29	森	586	494432	谷川	107140	海老沢	3224
30	内	881	472428	石川	104193	大河内	3189
総個数	62323736			31229331		1224418	
文字組数	2637			48757		9372	
エントロピー	7.26			11.12		8.75	
等確率エントロピー（ビット）	11.36			15.57		13.19	

率で出現するとしたときの 1 文字あたりのエントロピーが 4.7 ビットであるのに対して，3 連接までの出現確率を考慮すると，そのエントロピーは 3.3 ビットになることを示した．この結果をそのまま用いれば，1 単語あたりのエントロピーは，単語の平均長が 4.5 文字であるので，14.85 ビットになる．しかし，これには単語の使用頻度が考慮されていない．そこで，彼は単語の使用頻度をジップの法則で近似して，英語 1 単語あたりのエントロピーは 11.82 ビットであると推定した．これを 1 文字あたりのエントロピーに換算すると，2.62 ビットになる．

これに対して，日本の苗字 54704 種が等確率で出現するとしたときの 1 苗字あたりのエントロピーは，15.74 ビットである．一方，各苗字の使用頻度をもとにして，1 苗字あたりのエントロピーを算出すると 11.21 ビットになる．このエントロピーは，英語 1 単語あたりのそれにほぼ等しくなっている．しかし，苗字の平均長が 2.0 文字であるので，1 文字あたりのエントロピーに換算すると，5.60 ビットになる．つまり，苗字の 1 文字あたりのエントロピーは，英単語の 1 文字あたりのエントロピーよりはるかに大きくなっていることが分かる．

4.4 苗字における文字の連接頻度

日本の苗字における 1 文字および 2，3 文字組の文字連接の使用頻度を求めたものを表 10 に示す．各々で使用頻度の高い上位 30 種を示した．この表で，1 文字欄は，その文字を含む苗字の個数と苗字全体での使用頻度を求めたものであり，2 および 3 文字組欄は，苗字のなかに各々の長さの文字連接が使用される頻度を求めたものである．したがって，2 文字組には，1

文字苗字は含まれていない．3文字組には，1文字および2文字の苗字は含まれていない．

まず，1文字でみると，「田」が最も多く使用され，次いで「藤」，「山」，「野」，「川」の順である．使用頻度の高い文字は，表3でも出現頻度の高かったものである．つまり，多くの文字と連接して苗字になる文字はその使用頻度も高いことが分かる．しかし，これには例外もあり，たとえば，第2位の「藤」は表3では35位である．これは，使用頻度10位までの苗字に「佐藤」，「伊藤」，「加藤」が存在するなど，藤原氏に端を発する歴史的な背景に基づくものであろう[1]．逆に，表3での出現頻度が第3位であった「谷」の使用頻度は17位に低下している．

2文字連接では，頻度の高い文字組は2文字苗字の使用頻度にほぼ一致する．これは，苗字には2文字のものが圧倒的多数を占めていることによる．しかし，「佐々」と「々木」，「長谷」と「谷川」など，使用頻度の高い3文字苗字の2文字連接も上位に出現している．3文字連接では，2文字以下の苗字が対象に含まれないこと，4文字以上の苗字の数がきわめて少ないことなどにより，3文字苗字の使用頻度にほぼ完全に一致している．頻度そのものもきわめて少ない．

これらの苗字における文字連接の頻度情報を積極的に活用して，たとえば，N-gramに基づく認識手法[8]に応用するなど，文字認識の精度向上に結び付けていくことができよう．ちなみに，使用頻度を考慮した苗字1文字のエントロピーが7.26ビットであるのに対して，2文字連接のエントロピーは11.12ビットになり，これからも連接情報を利用して認識することに効果のあることが示唆される．

4.5 範囲を決定できる文字と苗字

頻度を考慮した日本の苗字について，3.6節と同様に，特定の長さの特定の文字位置でだけ使用されるという性質を利用して，苗字の範囲を決定できる文字の種類を求めたものを**表11**に示す．

この性質を備えた文字は699種であり，文字の種類全体の26.5%になる．これは，前述の苗字データベースの分析結果よりかなり少なくなっており，特に2文字苗字の末尾位置における該当文字数の減少が目立っている．これより，検索できなかった苗字，つまり電話帳に出現しなかった苗字には，特殊な文字連接のものが多く存在しているといえよう．

次に，これらの文字によって，苗字の範囲を決定することのできる苗字数を求めたものを**表12**に示す．この性質を満たす苗字はわずかに1323種で，全体の2.4%にしかならず，前述のデータベースの分析結果と同様に，きわめて少数であることが分かる．

表11 頻度を考慮して苗字範囲を決定できる文字の個数
Table 11 The number of character categories which can determine the range of family names when the usage frequency is considered.

苗字長	文字位置					
	1	2	3	4	5	6
1	160					
2	420	102				
3	4	10	3			
4	0	0	0	0		
5	0	0	0	0	0	
6	0	0	0	0	0	0

表12 頻度を考慮して範囲が決定できる苗字の個数
Table 12 The number of family names which can be determined by some specific character when the usage frequency is considered.

苗字長	文字位置					
	1	2	3	4	5	6
1	160					
2	830	125				
3	4	201	3			
4	–	–	–	–		
5	–	–	–	–	–	
6	–	–	–	–	–	–

表13 範囲が決定できる苗字の使用頻度
Table 13 The usage frequency of Japanese family names those range can be determined by the character.

苗字長	文字位置		
	1	2	3
1	7855		
2	75498	3842	
3	158	10422	40

さらに，これらの苗字の使用頻度を求めたものを**表13**に示す．範囲が決定できる苗字の頻度は全体でも10万個弱であり，総使用頻度の0.3%にしかすぎないことが分かる．そのうち，約77%は2文字苗字の先頭位置において，この性質を備えた文字を持っているものであった．

なお，使用頻度がただ1個である文字は，その文字を手がかりとして苗字を決定できるだけでなく，その苗字を使用している個人をも決定できるという性質を持っていることになる．この性質を有する文字として「快」，「棋」，「奉」など92種が存在した．ただ，その種類は非常にわずかである．

5. おわりに

日本の苗字データベースを作成し，その計量的分析

を行った．ここでは，より普遍的な統計データの収集を目的として，71452 個の苗字を登録した．

このデータベースの分析により，日本の苗字には次のような特徴のあることが明らかになった．

(1) 日本の苗字には，1 文字から 6 文字のものまであるが，その 8 割は 2 文字苗字である．平均の苗字長は 2．18 文字である．

(2) 苗字には，3796 種の文字が使用され，その 6 割は第 1 水準漢字である．しかし，第 1 水準漢字もすべてが使用されるのではなく，約 550 種は苗字に使用されない．

(3) 出現頻度が 100 回以上の文字は 346 種，10 回以上のものは 1400 種である．逆に，ただ 1 回しか出現しない文字は 1012 種である．また，頻度の上位 120 位までの文字をとれば出現頻度は過半数に達し，1000 位までとれば 92% になる．

(4) 頻度の上位から 500 位までの文字によって苗字の 63% が完成し，1000 位までとれば 84% が完成する．逆に，苗字全体の 95% を完成させるには 1852 種の文字が必要である．

(5) 文字位置によって，使用される文字と頻度には大きな偏りがある．たとえば，2 文字および 3 文字苗字の末尾に出現する文字は，他の位置に比べて 400 種近くも少ない．また，文字によっては，苗字の範囲を決定できるものがあり，約 35% がこの性質を満たす．しかし，これらの文字によって範囲を決定できる苗字は全体の 2.7% にすぎない．

また，電話帳データベースによる苗字の使用頻度を考慮した分析では，日本の苗字に次のような特徴のあることが明らかになった．

(6) ここで使用されている文字および苗字の種類は，上記のデータベスのものよりかなり少ない．

(7) 実際に使用されている苗字は，ごく少数のものに集中している．たとえば，わずか 1% 弱の苗字で使用頻度は全体の 6 割に達する．

(8) 使用頻度を考慮した日本の苗字のエントロピーは 11.2 ビットである．これは，シャノンの求めた英単語のエントロピーにほぼ等しいが，1 文字あたりに換算すると，苗字でのそれの方がはるかに大きい．

(9) 範囲が決定できる苗字の使用頻度は，10 万個弱であり，全体の 0.3% にしかすぎない．しかし，文字連接の頻度情報は，うまく利用すれば認識処理に効果をもたらすであろう．

これらの分析によって得られた日本の苗字の性質を知識として，たとえば，郵便物の宛名認識や個人情報の記載伝票の認識において，そこでの苗字位置を確定したり，認識すべき対象文字の種類を限定したりすることに利用し，認識精度の向上に結び付けていくことが可能となろう．しかし，住所にも苗字と類似の文字列が多く存在するため，今後は苗字だけでなく，住所や名前の分析にも拡張し，全体としての諸性質を定量化していく必要がある．

謝辞 本研究を進めるに際してご指導，ご鞭撻をたまわりました特定領域研究「人文科学とコンピュータ」の「データベース」研究班代表者・小澤一雅大阪電気通信大学教授に深く感謝します．

参考文献

1) 丹羽基二：姓氏の語源，角川書店 (1981)．
2) 丹羽基二（監修），日本ユニバック（編）：日本の苗字（表記編），日本経済新聞社 (1978)．
3) 鈴木雅人，孫 寧，阿曽弘具：キー文字駆動型地名推論による手書き宛名認識アルゴリズム，電子情報通信学会技術研究報告，PRU 95-5 (1995)．
4) 井野英文，孫 寧，根元義章：ストローク幅に着目した閾値可変型手書き宛名の切出し・認識アルゴリズム，電子情報通信学会技術研究報告，PRU 95-109 (1995)．
5) 梅田三千雄，濱裕治郎：2 段階文字認識と後処理を融合した手書き住所文字列の認識，電気学会論文誌（C），Vol.117-C, No.11, pp.1607–1614 (1997)．
6) 梅田三千雄：日本の苗字データベースの作成と分析，電子情報通信学会技術研究報告，PRU 92-5 (1992)．
7) 北 研二，中村 哲，永田昌明：音声言語処理—コーパスに基づくアプローチ，森北出版 (1996)．
8) 中川聖一：確率モデルによる音声認識，電子情報通信学会 (1988)．

(平成 10 年 8 月 24 日受付)
(平成 11 年 1 月 8 日採録)

梅田三千雄（正会員）

昭和 20 年生．昭和 43 年大阪大学基礎工学部卒業．同年日本電信電話公社（現 NTT）入社．平成元年大阪電気通信大学工学部教授．現在，同大学情報工学部教授．工学博士．文字認識，画像処理，認知科学等の研究に従事．電子情報通信学会，電気学会，映像情報メディア学会，画像電子学会，計量国語学会，日本鑑識科学技術学会各会員．