自然言語処理 91－4
（1992. 9. 17）

人名の読みからの検索法

高橋克巳　岩瀬成人
NTT 情報通信網研究所 メッセージシステム研究部

情報提供サービスの増加にともなって，データベースを検索する機会が増えている．その中でも人名を検索する機会は多い．本稿では人名をその読みから読みのゆれや誤りを解消して検索する方法について述べる．始めに日本人の人名データベースについて簡単な報告を行ない，そのデータを基に姓の読みのゆれの調査を報告する．その結果から，同等な姓を与える規則は，(1) 連接漢字先頭音の濁音化，(2) 漢字連接部の音の変形，(3) 同音の表記のゆれ，からなることを述べる．また類似した姓が発生する原因の考察も行なう．さらに実際に検索を行なう際に問題となる点について述べる．

Name Retrieval from Readings of *Kanji*

Katsumi Takahashi　Shigehito Iwase
NTT Network Information Systems Laboratories

Japanese names are represented by *kanji* and its readings and some *kanji* names have several reading variations. This paper proposes a name retrieval method from readings of *kanji*. It is suitable for finding a person's name from a database where keywords may have variations or errors in readings. Rules for variation of the readings are reviewed through a report of the Japenese person's name database. As a result of the research three rules are proposed for the family name equivalence. In addition to the equivalence, the family name similarity are also surveyed.

1 はじめに

データベースから人名を検索する際に，人名の「読み」を検索の条件として検索する手法について述べる．オンラインデータベースなどの増加によって，人名を検索する機会が増えている．人名は表現にゆれが存在するため，不正確な表現であっても目的とする人名が特定できる技術が必要である．

日本人の人名は「漢字」とその「読み」の二つの方法で表現される特徴があり，どちらを条件としても検索が可能なことが必要である．「読み」からの検索は，(1) 漢字に変換する手間がかからない，(2) 漢字が不明の場合でも検索が可能 [宮部 83]，などの利点がある．

かな読みから検索する場合，例えば「ナカタ / ナカダ（中田）」のようにひとつの姓に複数の読み方や表記法が存在するため，読みのゆれを解消して検索することが必要である．さらに誤りの訂正のために，「マツダ / マスダ」のような音の類似した姓まで拡張して検索する機能が必要である．

この研究の目的は，表現のゆれや誤りを解消した検索を行なうことである．文字列のゆれや誤りはランダムに起こるものであるが，対象を特に人名の読みに限ってやると，「同じ姓を表すもの」「似た姓を表すもの」の傾向が存在すると考える．これらを同値規則として収集し報告を行なう．同値規則から同等な文字列を検索する方法は，同値類を代表する要素に正規化を行なって検索することが一般的である [Hall80]．同値規則が決まれば，その規則から正規形への変換規則を機械的な操作で求めることができる [Knuth 70]．

そこで実際の電話帳のデータベースに基づいて，人名の姓のゆれの規則について報告する．

2 人名の読みのゆれ

日本人の人名は「漢字」と「読み」の二つの表現があり，特に「読み」のゆれが顕著である．

本研究の目的は，電話帳や文献検索システムなどにおいて，姓の「読み」を検索キーとして情報を検索する際，表現にばらつきが存在したり不正確な可能性のある問い合わせであっても，目的とする情報を提示することにある．そこで人名の読みのゆれがおこる事例を収集する必要がある．

大山らは [大山 91] にて，住所を読み上げて，他者がキーボードでかな入力する時に発生した誤りの分析を行なっている．

読みのゆれが起こる原因として，次のようなものが考えられる．

a) 一つの漢字表記に対して複数の読みが存在する
　ナカタ / ナカダ（中田）
　イワタニ / イワヤ（岩谷）

b) 同じ音の表記法が複数ある
　トウヤマ / トオヤマ（遠山）

c) 音の類似のため聞き違え
　マツダ / マスダ

d) キータイプミス，手書き文字認識誤りなど

a) は姓をかなで表現する際に最も起こる問題で，例えば「田」の読み方のように判断が難しいものから，単純な漢字の別読みとの混同まで存在する．c) は例えば電話などで名前を聞きとる時に生じる誤りであるが，そればかりではなく，聞き覚えの姓が誤って記憶されることの原因ともなり得る．d) であるが，特定のハードウエアの使用に付随して起こる問題で今回の検討では扱わない．

3 日本人の姓のデータベース

本章では日本人の姓のデータについて簡単な報告を行なう．

電話帳掲載情報の中から，姓の「漢字」と「読み」の 2 項目を組みとして抽出し姓データベースを作成した．データはほぼ全国からで，母集団数は約 2900 万，抽出は 1991 年 8 月である．

ここで以下の種類の姓が抽出された．外国の姓と考えられるものなどは機械的に取り除いてある．

漢字＋読み	158008 種
漢字	97966 種
読み	81497 種

表 1: 姓データベースの規模

始めに代表的な姓を示す．「漢字」＋「読み」を統計の単位としている．姓はかなり偏った分布を見せており，上位 10 種の姓の出現類計で全体の 10% を越えている．

佐藤	サ / トウ	1.62	吉田	ヨシ / ダ	0.66
鈴木	スズ / キ	1.44	山田	ヤマ / ダ	0.65
高橋	タカ / ハシ	1.18	斎藤	サイ / トウ	0.62
田中	タ / ナカ	1.03	佐々木	サ / サ / キ	0.59
渡辺	ワタナ / ベ	0.93	山口	ヤマ / グチ	0.51
伊藤	イ / トウ	0.89	松本	マツ / モト	0.49
中村	ナカ / ムラ	0.84	井上	イノ / ウエ	0.47
小林	コ / バヤシ	0.83	木村	キ / ムラ	0.47
山本	ヤマ / モト	0.82	林	ハヤシ	0.43
加藤	カ / トウ	0.71	清水	シ / ミズ	0.42

表 2: 姓データベース（上位 20，数字は全体に占める出現累計の割合 %）

次に姓の種類とその全体に占める割合を示す．上位約 300 の姓までの累計で全体の 50% を，同約 2500 の姓で 80% を占めている．

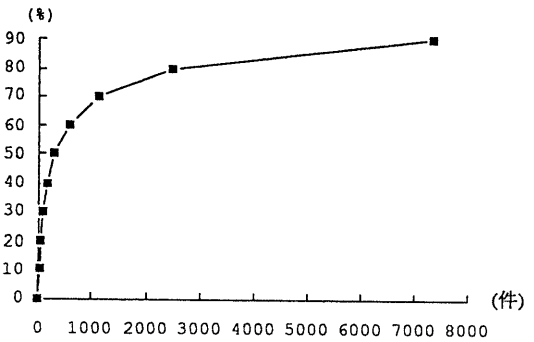

図 1: 姓の種類と全体に占める割合

4 姓の読みのゆれの調査

読みかたが 2 通り以上ある姓は 35,000 種を越える．これは姓（漢字）の総数の 3 分の 1 以上に相当する．

本章では，上位 90% をしめる代表的な姓 7350 種について行なった読みのゆれの調査について報告する．

4.1 同字異音姓の分析

2 章で指摘した読みのゆれが起こる原因のうち，a) の漢字の読みの問題の事例は，姓データベースで漢字が同じで読みが異なる同字異音姓を解析することで収集することができる．また，b) の表記法についても同じ調査である程度知ることができる．

ここで同字異音姓の調査方法について説明する．

1. 調査対象は，累計で全体の 90% を占める姓 (図 1) とする．これは漢字で 6505 種存在する．
2. この 6505 種に対して同字異音姓を探す．
3. 同字異音姓は出現比が 10:1 までのものとする．
4. 読みが部分的に一致する姓は（ナカタ / ナカダ）ゆれの起こる漢字単位（タ / ダ）で，出現位置 (先頭か，末尾か，それ以外か) とともに収集し，そうでない姓は 姓単位で収集する．

その結果，約 4 分の 1 の姓に別読みが存在し，1876 種の姓の読みのゆれのパターンを収集した．

代表的な姓（漢字）	6505 種
別読みの存在した姓	1517 種

表 3: 別読みの存在

なお，全データ 16 万件を走査しても全く別読みが存在しない姓が 1755 種あったが，その例を以下に示す．

高橋	田中	斎藤	坂本	岡本	中野	原	今井
菊地	松尾	野村	松井	杉本	北村	矢野	山中
浅野	松下	中西	森本				

表 4: 別読みのない姓（左から右へ頻度順）

収集した例を示す.

種類	読み 1	読み 2	出現位置 *
103	タニ	ヤ	1
101	タ	ダ	1
51	カワ	ガワ	1
48	シマ	ジマ	1
48	サキ	ザキ	1
39	サワ	ザワ	1
30	キ	ギ	1
30	オ	コ	0
22	ウエ	カミ	0
20	ハタ	バタ	1
17	ズ	ヅ	1
16	ハラ	バラ	1
16	カナ	カネ	0
15	ツカ	ヅカ	1
14	トミ	ドミ	1
14	ツ	ヅ	1
14	ウエ	ガミ	1
13	シン	ニイ	0
13	コ	フル	0
13	カミ	ガミ	1
12	ジ	チ	1
12	コシ	ゴシ	1

* 出現位置は, 0: 先頭 1: 末尾 2: それ以外

表 5: 姓の漢字単位の読みのゆれ

収集した読みのゆれを，互いの関連性を考慮して分類を試みる.

発音，表記に関するものとしては,

(y1) **2 漢字以上の姓において 2 つめ以降の漢字の先頭音で連濁が起こるもの**．(696 件 /1876 件中)
ナカダ / ナカタ (中田)
これは特定の漢字を使う姓に極めて特徴的に見られる現象である．代表としては，末尾に「田」「崎」「川」「島」「沢」などを使う姓がある.

(y2) **発音の便宜上から，漢字の連接部で音が変化するもの**．(104 件 /1876 件中)
カナマル / カネマル (金丸)
カジワラ / カジハラ (梶原)
これも特定の漢字の使用に見られる現象で，先行する漢字では「金」「成」「立」などが，後続の漢字では「原」「幡」などに見受けられる.

(y3) **最後の音または先頭の音が微妙に変化するもの** (39 件 /1876 件中)
クマガイ / クマガヤ (熊谷)
カケイ / カケヒ (筧)
例えばケンモチ / ケンモツ (剣持) の様に動詞の活用語尾に対応して変形するものが多数見受けられるが，原因は明らかでない．先頭音の変化ではアガツマ / ワガツマ (我妻)1 例のみ見つかった.

(y4) **二重母音，ou/oò の混同**．(12 件 /1876 件中)
トオヤマ / トウヤマ (遠山)
遠いの訓は「とおい」であるが，トオヤマ / トウヤマは 5 対 3 の割合で使われている.

(y5) **ジとヂ，ズとヅの混同**．(44 件 /1876 件中)
イイヅカ / イイズカ (飯塚)
(y1) と組み合わせると，特にジ＝ヂ＝チとズ＝ヅ＝ツの関係がある.

(y6) **音の挿入または欠落**．(132 件 /1876 件中)
シカノ / カノ (鹿野)
イノモト / イモト (井元)
この例も互いの関連が強い．前者は漢字の読みの一音が欠落したもの，後者は元々の漢字の読みにない「ノ」が挿入されている.

それ以外のものは漢字の読みの違いの程度によって次のようになる.

(y7) **漢字 1 文字の読みの違い**．(608 件 /1876 件中)
イワタニ / イワヤ (岩谷)

(y8) 全ての漢字の読みの違い．(241 件 /1876 件中)
ホンタニ / モトヤ (本谷)
ショウジ / トウカイリン (東海林)

4.2 音の類似した姓の調査

2 章で指摘したゆれの原因のうち残る，

c) 音の類似のため聞き違え

は誤り訂正を精度良く行なう問題である．姓の聞き違えが起こる条件に以下のものがあると考え，類似した姓を収集を行なっている．

1. 音節数が同じであること
2. 1 音のみが異なり，かつその音も類似している
3. アクセントが同じであること

今回上記のうち 1. と 2. のルールを使って，読みが 4 文字と 3 文字の姓について類似した姓の組を作成した．その結果対象とした 5233 種の姓 のうち 2151 種に類似した姓が 存在した．

類似姓数	4 字姓	3 字姓
2	568	773
3	23	80
4	0	7

表 6: 類似音による類似姓の発生 (単位: 件数)

4 文字姓で類似姓数 3 の組は，

- ヒノハラ, シノハラ, イノハラ ((a) から)
- ハラオカ, ハナオカ, ハタオカ ((c) から)

など 23 組があり， 3 文字姓の類似姓数 4 の組は

- ウヤマ, クヤマ, スヤマ, ツヤマ ((i) から)
- イムラ, キムラ, シムラ, チムラ ((a) から)

などの 7 組が存在した．

またこの規則でそれぞれの姓に類似した姓を派生させたところ，派生数の上限は 4 文字姓で最大 5 つ (イシヤマ 1 例)， 3 文字姓で最大 7 つ (タワタ, イシマ 2 例) でありこのような派生を行なわせて検索を行なったとしても実用的に可能であることがわかった．以下に例を示す．

イシヤマ {ヒシヤマ, ウシヤマ, イイヤマ, イチヤマ, イシハマ}
カキハラ {サキハラ, タキハラ, カイハラ, カシハラ}
タワタ {ナワタ, カワタ, サワタ, タバタ, タハタ, タワダ, タワラ}
イシマ {チシマ, エシマ, イイマ, イシバ, イシワ, イシガ, シマ}
ハナイ {アナイ, ヤナイ, ハタイ, ハライ, ハマイ, ハナキ}

実験の方法を紹介する．

1. 累計で全体の 90% を占める姓 (図 1) から， 読みが 4 文字のもの (3046 種)， 3 文字のもの (2187 種) を集める．
2. [石野 91] に紹介されている国語音韻相通例をもとにして， 母音の無声化 [前川 89] などを参考にして， 類似した音の組を 14 決める (表 7).
3. 各姓の先頭から 1 文字ずつ下記の規則を適応して， 1 文字違いでかつ違いの範囲が下記の規則に入る姓の組を 同じ姓の集合から探す．

(a)	「イ」「ヒ」「シ」「チ」「キ」	ヒライ / シライ
(b)	「マ」「バ」「ワ」	カマタ / カワタ
(c)	「ナ」「ダ」「タ」「ラ」	タカダ / タカナ
(d)	「カ」「タ」「サ」	カワモト / サワモト
(e)	「ハ」「ア」「ワ」「ヤ」	アベ / ヤベ
(f)	「マ」「ナ」「ガ」	ホンナ / ホンマ
(g)	「ホ」「オ」「ゴ」「コ」	アカホ / アカオ
(h)	「ビ」「ミ」「ニ」	ミナガワ / ニナガワ
(i)	「ウ」「フ」「ス」「ツ」「ク」	ウシダ / クシダ
(j)	「イ」「エ」	ホリエ / ホリイ
(k)	「ウ」「ユ」「イ」	ウリノ / ユリノ
(l)	「ウ」「オ」	トウヤマ / トオヤマ
(m)	「ケ」「テ」	ケヅカ / テヅカ
(n)	先頭の「イ」「(無声音)」	イシダ / シダ

表 7: 類似音の組と類似姓の例

5 検索法

前章までで，姓の読みにさまざまなゆれが存在することを述べた．本章ではそのゆれの性質から，姓の読みの同等性，類似性について指摘する．

5.1 姓の読みの同等性

姓の読みは，もともと各個人に対しては唯一に特定でき，ゆれは存在しない．しかし実際上は読みの小さな差があってもおなじ姓とあつかうことが多い．ここで姓の読みの同等性を「多くの場合同一の姓を表す読み」と定義し，姓の同等性を与える同値規則について考える．

前章の結果から，以下に与えられる3つの規則を姓の同等性とする．同等性の規則として選定した理由は以下の通りである．

1. 漢字表記が同じであること．
2. 音の違いは1音までで，その違った音も似ていること．
3. 音節数が変わらないこと．
4. 高い頻度で双方が使われていること．

姓の読みの同等性規則

(E1) **特定の漢字が2文字目以降に使われるとき，その読みの先頭音で連濁が起きやすい．この清音と濁音を同じとあつかう．**
連濁 [佐藤 89] が起こる漢字は，田, 崎, 川, 島, 沢など108種に存在する．

(E2) **2漢字以上の姓で特定の漢字が使われるとき，発音の便宜上から漢字の連接部で音が変化することがある．この読みを同じとあつかう．**
具体的には，先行漢字の音の変化では，「金(カナ / カネ)」「成(ナリ / ナル)」「立(タチ / タツ / タテ)」など12種，後続漢字では「原(ワラ / ハラ)」「幡(ハタ / ワタ)」の2種を規則として収集する．

(E3) **表記法のゆれ．二重母音オウ / オオなどや，濁音のジ / ヂ，ズ / ヅを同じとあつかう．**
2重母音は，オウ / オオ音に4種(王, 逢, 近, 扇)，トウ / トオ(遠)1種，コウ / コオ音(興, 郡)，コウ / ゴウ音(郡)の8種に存在する．ジ / ヂの関係は，地, 近の2種に，ズ / ヅは津, 塚, 妻など10種で起こる．

(E4) **その他音の小さな変化**
最後の音の変化としてガイ / ガヤ(谷)，モチ / モツ(持)，など6種，先頭の音の変化として，アガ / ワガ(我)1種を規則として収集する．

5.2 同等な姓の検索法

前節で指摘した姓の同等性規則を使えば効率良く同等な姓を検索することができる．本節では具体的な手法にふれ問題点を述べる．

規則は「漢字」と「読み」と「漢字の出現位置」および「漢字中の変形位置」によって記述されている．したがって，最も単純かつ正確な検索方法は，データベースの検索キー作成時に該当する姓に対して，(E1)から(E4)の規則を使って複数の読みを派生させ，あり得る同等な良みを前もって網羅しておく方法である．しかしこの方法は冗長性が高く，特に大規模なデータベースの処理には不向きである．

同値関係の文字列を処理するには，同値類を代表する文字列へ正規化を行なって検索を行なうことが知られている [Hall80]．現在番号案内用のデータベースもいくつかの変換規則を定めて正規化処理を行なっている [宮部 83][戸部 90]．正規化処理は「同じとあつかう文字列は，同じ文字列へ変形される」ことを保証した変形を，検索文字列とデータ中の文字列に同等に行ない，検索を行なうものである．正規文字列への変換規則は同値規則から決定でき，それは機械的な操作でも求めることができることが知られている [Knuth 70]．

正規化で問題となる点は，かな文字の検索条件の解析の難しさである．かな文字の解析が不正確で単純な正規化処理を行なうと，中野(ナカノ)と長野(ナガノ)が混同されるなどの問題が起こるので，かな文字の解析手法の検討が必要である．

5.3 姓の読みの類似性と検索法

前章の読みのゆれの分類のうち，残るものから類似性の規則を考える．前にも指摘した通り，その姓が(漢字を)見て覚えたか，聞いて覚えたかで類似性は大きく変わる．それぞれに対応して，同字異音の分析と類似音によるゆれの調査を報告した．

類似度にも様々な程度が存在するが，次のように分類する．

(S1) 漢字単位で読みのゆれが起こる
タニ / ヤ (谷)，オ / コ (小)，ウエ / カミ (上) のように人名に良く使われ，高い確率で読みのゆれが起こる特徴的な読みを収集すること，およびイデ / デ (出)，ミズ / ミ (水) などの音の欠落が起きている読みを収集することが効果的である．

(S2) 漢字の連接部での「ノ」「ガ」「ツ」「ナ」の挿入
「ノ」は，イ (井，猪)，イチ (一) などの直後，「ガ」はヤ (谷) の直前に挿入されることが多い．

(S3) 姓全体での読みのゆれ
ショウジ / トウカイリン (東海林) などの姓を収集する．

(S4) 音の類似性により起こる聞き違え
4.2節参照．

(S2) および，(S1)(S4) が有効な規則であると考える．類似検索を行なう際には，4.2節で指摘したように，類似な姓の範囲がどの程度まで広がるかを正確に知っておく必要がある．

6 むすび

日本人の人名データベースについて簡単な報告を行ない，そのデータに基づいた，姓の読みのゆれについて行なった調査の報告を行なった．同等な姓を与える規則は，(1) 連接漢字先頭音の濁音化，(2) 漢字連接部の音の変形，(3) 同音の表記のゆれ，からなることを述べ，具体的な漢字と読みに関する規則として示した．また，類似した姓が発生する原因についても考察した．

本検討は姓の統計情報を調べることによって，読みのゆれの規則を推測し網羅することを試みた．さらにはこの結果をふまえた実験により，誤用などの例を収集してやる必要がある．

なお本検討は姓名のうち姓に対象を限って行なったが，名前固有の問題や姓と名の組み合わせによる問題は別に検討したい．

日本人の人名の問題は，漢字と読みの問題と深いつながりがある．人名は呼称でもあり，口に出されながら永い年月を経てあるものは発音が容易になるように，またあるものは他との区別を明確にするように変化してきたものと考えられる．この漢字と読みの問題を正視して検討を続けたいと考える．

また本研究の成果を ANGEL(電子番号案内システム) に反映し，さらに検討を続ける予定である．

謝辞

電話帳データは ANGEL システムから提供を受けました．また 104 番のコミュニケータの皆さんほか ANGEL システムで働く多くの人に，問題点や事例を聞くことができました．ここに感謝します．日頃討論いただく情報案内方式研究グループの皆さん，NTT 基礎研究所の梅村恭司さん，NTT 情報通信網研究所の千田昇一さんに感謝します．

参考文献

[Hall80] Patrick A.V. Hall, Geoff R. Dowling. Approximate String Matching. Computing Surveys, Vol.12, No.4, December 1980

[Knuth 70] D. E. Knuth, P. G. Bendix. Simple word problem in universal Algebras, Computational problems in abstract algebra. Pergamon Press, page 263-297, 1970.

[石野 91] 石野博史．放送における類音語．日本語学 1 月号, 1991

[大山 91] 大山芳史，今村賢治．住所入力支援方式の検討．電子情報通信学会春季全国大会，D-106, 1991

[佐藤 89] 佐藤大和．複合語におけるアクセント規則と連濁規則．日本語と日本語教育，第 2 巻 日本語の音声・音韻 (上)，明治書院，1989

[戸部 90] 戸部美春, 武藤信夫, 山本康二. 高付加価値型番号案内システム (CUPID) の電話帳検索方式. NTT R&D, Vol.39, No.6, 1990

[前川 89] 前川喜久雄. 母音の無声化. 日本語と日本語教育, 第2巻日本語の音声・音韻(上), 明治書院, 1989

[宮部 83] 宮部博, 大山実, 本郷郁夫. 名義検索システムー電話番号案内業務への適用ー. 情報処理学会論文誌 Vol. 24 No. 4, 1983