人名のかな表記のゆれに基づく近似文字列照合法

高橋克巳[†] 梅村恭司[††]

日本人名のかな表記にゆれとよばれる変形が存在し，日本語情報検索システムの問題となっている．本論文では人名のかな表記にゆれが存在してももれのない検索を可能とする近似文字列照合法を提案する．ゆれの問題に対処するためには表記を統一して検索を行うことが一般的であるが，現在かな表記を統一する基準は明らかではなく，そのため統一すべきゆれが多種になった場合の対策も明らかになっていない．本文では日本人名約3,000万件を解析し，姓のゆれのデータを収集分析する．その結果，姓は9万種の姓のゆれ単位に分類できること，実データ上で58%の姓に何らかのゆれが存在すること，ゆれの原因は連濁などの接続部の変化が大部分を占めることを明らかにする．さらにこのゆれの関係に基づいた正規化による照合を提案する．すなわち，実際にすべてのゆれを21,276組の文字列の等式関係で記述し，そこから自動的に15,841の正規化規則を作成して照合する方法を提案する．この正規化規則を使った照合法を人名の分布にしたがった検索に適用し，再現率と適合率の観点から評価を行った．その結果，93%の適合率を達成したうえで，完全一致検索では1検索あたり15%存在していたゆれによる検索もれを解消した．人名についてかな表記のゆれが存在してももれのない検索が可能となった．

Approximate String Matching Based on *Kana* Variants of Names

KATSUMI TAKAHASHI[†] and KYOJI UMEMURA[††]

The existence of several *Kana* variants for many Japanese names is a major problem in constructing Japanese information retrieval systems. We propose an approximate string matching method for finding personal names when there may be *Kana* variants. Most methods for dealing with this problem normalize *Kana* variants for the name, but there is no normalizing standard and the best method for handling the many variants to be normalized is unclear. We collected name variants in 30 million Japanese names and then classified them into 90 thousand name-variant units. We found that 58% of the names have more than two variants. Our proposed matching method uses normalization based on the *Kana* variants, which are described by 21,276 equivalence relations that are automatically converted into 15,841 normalization rules. This matching method can retrieve all relevant records with 93% precision. Thus, it is superior to exact string matching, which overlooks 15% of the relevant records.

1. はじめに

日本人名のかな表記にはゆれとよばれる変形が多数存在している．かな表記のゆれとは，同一の語に対して複数のかな表記が許容されている現象をいう[1]．表記のゆれは一般に観察される現象であるが，特に人名には「なかだ-なかた(中田)」，「あめみや-あまみや(雨宮)」など多数の例が存在する．この現象は日常会話では見逃されることがあっても，情報検索では問題になる．電話番号検索や文献検索などのように人名から情報を検索する時，検索者は登録されている人名を正確に入力する必要がある．しかし対象に表記のゆれがある場合は，問い合わせの表現と登録情報の表現に不一致が生じ情報の検索もれが生じる場合がある．この問題を解決するために，検索を行う際に「完全に」一致したものばかりでなく「おおまかに」一致するかどうかを判定する処理，近似文字列照合 (approximate string matching) が必要である．

本論文では，かな表記を使う人名検索において，ゆれのある文字列の同一性を判断する近似照合法について述べる．近似照合を実現するための第1の課題は，どのような人名に，どのようなゆれが存在するかを明らかにすることである．日本の人名は種類が多く，検索のためにはゆれの類型化が必要である．そのため3,000万件の人名に対してゆれの解析を行い，ゆれを具

† NTT 情報通信研究所知的情報処理研究部
Intelligent Information Systems Laboratory, NTT Information and Communication Systems Laboratories
†† 豊橋技術科学大学情報工学系
Department of Information and Computer Sciences, Toyohashi University of Technology

体的に求めてみる．第2の課題は，求めたゆれの情報に基づいて，効率の良い検索ができることである．これに対しては，表記を統一して照合を行う正規化による照合法を行う．正規化規則を人手で作成する作業は一般に容易ではないために，本論文ではゆれの関係から機械的に正規化規則を求める手法を提案し，実際に正規化規則を作成する．このうえで，明らかにしたゆれに対して，検索もれがないことを保証し，無駄な名前を検索しないことを目標とする．この目標に対しては，提案する照合法を，現実に想定される問い合わせに対する検索の再現率（適合すると判断される総レコード中，実際に検索された割合）と適合率（検索されたレコード中，適合すると判断される割合）[2]の期待値から評価する．さらにこの照合法を10,000人のデータベースにおける検索処理に適用し，ゆれによる検索もれを解消したことを検証するとともに適合率に関する課題を論じる．

2. 背　　景

情報検索における課題として，キーワードの表記の不一致への対策の必要性が認識されている[3]．本研究も表記の不一致への対策を論じる．本章では関連する既存の研究を説明しながら，本研究の位置付けについて説明する．

2.1 人名のかな表記のゆれ

本研究で対象とするかな表記を検索キーとする検索法には，漢字に変換する手間がかからない，漢字が分からなかったり入力できない場合でも検索可能，などの利点がある．この検索では，音声の聞き違い，同一漢字の使用などの様々な程度や要因で表記の不一致が起きている．このなかでも「ゆれ」すなわち同じ名前を指しているにもかかわらず複数の表記が認められている現象を第一に解決する必要がある．

前章で第1の課題として，ゆれの起こる姓とその類型を明らかにすることをあげた．人名のかな表記に関して文献[4)~7)]の報告があり，かな表記の変形には，かなづかい，清音濁音，長音の扱いの違いなどがあることが明らかになっている．しかしその全体像は明らかになっていないため，3章で3,000万人のデータをもとにかな表記のゆれの分析を行う．

2.2 正　規　化

第2の課題である効率良い検索を行うために，正規化を使った照合を行う．一般に，文字列の正規化を使った検索は，問い合わせ文字列と登録情報の文字列の正規形を作って比較するために，1回の照合で複数の異なった表記のレコードを検索することができる．この性質はシステムの負荷を減じるので，電話番号検索のシステムなどで使用されている．

正規化を使った検索について，Hall と Dowling が同値性問題として述べている[8]．以下同値性問題について要約する．文字列の集合 S における関係 R が，文字列 $a, b, c \in S$ に対して，反射律，対称律，推移律すなわち，

（1） aRa

（2） $aRb \Rightarrow bRa$

（3） $aRb, bRc \Rightarrow aRc$

を満たすとき，R を同値関係と呼ぶ．このとき集合 S は R によって，同値類 $S_1, S_2, S_3, \cdots$，の直和に分解することが可能である．問い合わせ文字列 q に対して，同値関係にある文字列を検索する処理は，q の属する同値類 S_q に含まれるすべての文字列 $a \in S_q$ を求めることと言い換えることができる．このことは，同値類ごとに正規形と呼ばれる文字列を定め，同値類に属する文字列をその正規形に変換する正規化規則を与えることができれば，文字列 q の正規形と同じ正規形を持つレコードを探すことと定義することができる．これを人名の検索に適用したのが本研究である．4章では人名のかな表記のゆれから同値関係 R を求める．

さらに正規化処理をするためには，具体的に正規化規則を求めなければならない．まず現在行われている正規化手法を紹介する．

英語の代表的な類音誤りの正規化手法に，現在でもISODEのディレクトリ[9]など使われている Soundex システム[10]がある．このシステムはアルファベット26文字を定義した7つのグループに分類し，1文字ずつ各グループに対応したコード（**表1**）に変換したうえで（先頭文字はそのまま残す）照合を行う．

一方，日本語のシステムでとられている正規化の代

表1 Soundex codes
Table 1 Soundex codes.

0 A E I O U H W Y	4 L
1 B E P V	5 M N
2 C G J K Q S X Z	6 R
3 D T	

表2 番号案内におけるかな表記のゆれの対策の例
Table 2 Solutions for the *Kana* variants problems in the directory assistance system.

問題点	対策
濁音・半濁音の間違いやすさ	濁音・半濁音を清音に変換
拗音・促音の書き方	小書きを大書きに変換
「じ・ぢ」「ず・づ」の使い分け	「ぢ」を「じ」に「づ」を「ず」に変換

表は清音と濁音の対立などの扱いである．**表2**に番号案内のシステム[11),12)]で行われている対策を示す．表2以外にもゆれは存在するが，それらは漢字に依存するなどの理由で普遍的な規則を記述しにくい．実際に番号案内のシステムでは，人名ごとに随時派生テーブルを用意してゆれの対処を行っている．現在カタカナの異表記を正規化する手法が提案されているが[13)]，漢字のかな表記のゆれのための規則が求められている．

表2の正規化規則は単純であるが，精度を増すために，「むかいだ」「むかだ」「むこうだ」「こうだ」を等しく扱いたいなどと，正規化に盛り込む関係を増やすと規則を人手で作成することが困難になる．規則が正確に作成されないと，本来一つの形になるべきものが，異なる形に変形されてしまうことがある．ここで必要な処理は，文字列の等しいという関係から，等しい文字列に対してはその関係を保ったまま正規形への変換を行う正規化規則を求める処理である．本論文ではこの処理にKnuth-Bendixの完備化アルゴリズム[14)]を適用し，4章では求めたRから正規化規則を作成する．この正規化規則はもれのない検索を保証する．これはアドホックに正規形を求めた検索ではできない性質である．

もれのないことの副作用として，意図しなかった名前が同じ正規形に変換されること，すなわち適合率が低下する可能性がある．本手法は再現率を重視したものであり，適合率に関する問題は5章で実際のデータにより検証する．

3. かな表記のゆれの解析

3.1 調査の対象

本調査は電話帳に掲載されている漢字表記を持つ姓を対象として行った．日本人の姓は出版されたものだけでも13万種強報告されているが[15)]，電話帳に掲載されている漢字姓は16万4千種におよぶ．調査対象の母集団「姓データベース」は，全国すべての電話帳の掲載情報の中から姓の漢字情報とかな表記情報を取り出し，出現度数を付して作成した．1993年7月の電話帳に基づいている．このうち頻度が全国で3以上のものを調査対象とした（**表3**）．従って対象の「姓」の種類（漢字とかなの異なり語数）は約10万8千，のべ語数約3,200万の集合である．**表4**に例を示す．使用字種は「漢字」はJIS漢字（JIS X 0208）のみ2,957種，「かな表記」は「を」を除く直音すべて69音（「ぢ」「づ」を含む）と撥音「ん」の70種である．「ゐ」「ゑ」は含まれない．拗音，促音に用いられる「ゃ・ゅ・ょ・っ」は小書きされていない．

表3 姓データベースの度数
Table 3 Frequency of the family name database.

漢字長	異なり語数			のべ語数	姓の例
	漢字＋かな	漢字	かな		
1	2344	1464	1261	1098963	林，森，原
2	86413	60386	48074	29281304	佐藤，鈴木
3	19336	15047	15116	1313876	佐々木，長谷川
4	357	273	326	6164	勅使河原
全体	108450	77170	60283	31700307	

表4 姓データベースの例（上位10）
Table 4 Example of family name (Top 10).

順位	漢字	かな表記	出現度数	順位	漢字	かな表記	出現度数
1	佐藤	さ／とう	486347	6	伊藤	い／とう	272248
2	鈴木	すず／き	430176	7	山本	やま／もと	271651
3	高橋	たか／はし	357501	8	中村	なか／むら	266947
4	田中	た／なか	337310	9	小林	こ／ばやし	256734
5	渡辺	わたな／べ	280624	10	加藤	か／とう	216225

3.2 ゆれの判断基準

ゆれの関係にある姓のかな表記は，同一の姓を指すものである．例えば「やまざき-やまさき（山崎）」は同一と判断し，「しようじ-とうかいりん（東海林）」や「ほんたに-もとや（本谷）」は別の姓という立場をとることとする．姓の同一か否かの判断を3,000万のデータすべてに対して行うための具体的な基準が必要である．そこで漢字単位に人名に用いられるかな表記を整理した．

漢字単位のゆれの判断の基準を初期には「漢和辞典の見出しで音と訓のように別項目扱いのものはゆれではない」と考えた．しかし常用漢字表[16)]を見ても「あめ，さめ（雨）」や「わける，わかれる，わかる，わかつ（分）」などの派生して生まれたものがそれぞれ見出しになっている例が存在した．また辞典に全く記載がないかな表記も存在し，慣用音まで含んだ音の体系化は容易ではない[17)]のが現状であり，独自の判断基準が必要になった．

そこですべての姓を漢字単位に分解して，漢字ごとにかな表記を集め，すべてのかな表記の対立を（1）変形位置，（2）音種，（3）形態の観点から35のカテゴリーに分類して，そのカテゴリーごとにゆれとみなせる関係が存在するかを調べた．その結果から漢字単位にかな表記をゆれのグループに分類する「漢字かな表記辞書」を作成し，これに従って姓のゆれの判定を行った．

3.3 姓のかな表記のゆれ

付録Aの処理を行ってすべての姓を「姓のゆれ単位」（以後NVUと略す）に分類した（**表5**）．NVUは

表 5 姓のゆれ単位の例
Table 5 Name-variant units.

漢字	かな表記
神代	くましろ
神代	こうしろ，こうじろ，こおしろ
神代	じんだい，しんだい
神代	かみしろ，かじろ，かしろ
神代	かくみ，かこみ
神代	かみよ
神代	こうだい
水谷	みずたに，みづたに，みつたき，みすたに
水谷	みずがい
水谷	みずや，みずのや
水谷	すいたに
角田	つのだ，つのた
角田	かくた，かくだ，かどた，かどだ
角田	すみだ，すみた，すまだ，すだ

表 6 姓のかな表記のゆれの原因
Table 6 Causes of the *Kana* name variants.

変形位置	種別	例	NVU数	比率(%)
連接部前項	母音	あめみや-あまみや(雨宮)	1498	8.9
	助詞	やまうち-やまのうち(山内)	753	4.5
	その他	かみざき-かんざき(神崎)	1186	7.1
連接部後項	連濁	やまさき-やまざき(山崎)	10725	64.0
	その他	ふじわら-ふじはら(藤原)	963	5.7
語頭	子音	おやま-こやま(小山)	786	4.7
	その他	しかの-かの(鹿野)	194	1.2
末尾	母音	じんぼ-じんぼう(神保)	813	4.9
	その他	たかはた-たかはたけ(高畠)	419	2.5
その他		やぎぬま-やなぎぬま(柳沼)	293	1.7
ぢ／じ，づ／ず		みづたに-みずたに(水谷)	410	2.4
おう／おお		とうやま-とおやま(遠山)	217	1.3
複数原因の混在		こうじろ-こおしろ(神代)	2747	16.4

比率はゆれのある 16,754 の NVU に対する百分率で示した．

「漢字表記」と「かな表記」の組を単位とする「姓」の集合で，1 つの NVU に属する姓は，漢字表記が同じで，かな表記が同一の名前をさす，すなわちゆれの関係にあるという性質を持つ．ゆれの判定は，前節で作成した「漢字かな表記辞書」に従った．

例えば「神代」には 13 種類のかな表記がある．このうちかな表記が「こうしろ」「こうじろ」「こおしろ」の姓を同じ名前と判断し 1 つの NVU に，同様に「かみしろ」「かじろ」「かしろ」を別の NVU に，つごう「神代」姓を 7 つの NVU に分類した．こうして姓全体 108,450 種を 87,630 の NVU に分類した．このうち 16,754 の NVU には 2 つ以上のかな表記が対応し，ゆれがあることが分かった．これは実データの 58%に何らかのゆれが存在することを示している．

次にゆれの起きた原因を**表 6** に示す．表 6 は NVU ごとに存在するゆれの原因を調べ集計した．連濁（連接部の後項語頭の清音が濁音化する，文献 18）に詳しい）が全 NVU の 6 割以上に存在する原因である．しかしそれ以外にも「あめみや-あまみや（雨宮）」といった母音のゆれや，「ふじわら-ふじはら（藤原）」といったの子音のゆれが存在している．これらは従来の正規化規則では対処できていないゆれである．

4. ゆれのある情報に対する文字列照合

本章では，目標とする近似照合を行うために，3 章で明らかにした姓のゆれ単位（NVU）から正規化規則を作成する．始めに「ゆれに基づく一致」を定義し，次に正規化に用いる同値類を定義し，その上で正規化規則を作成する．

4.1 ゆれに基づく一致

目標とする照合は，人名の検索において問い合わせまたは登録情報のかな表記にゆれがあっても，もれのない検索を行うことである．この検索を「ゆれに基づく一致」と呼ぶことにする．定義を次に示す．

姓のかな表記の集合を S とするとき，S における関係 V_s を，$a, b \in S$ に対して，a と b が同じ NVU に属すとき aV_sb が成り立つとする．すなわち，

$$\exists x, \exists u;\ (x, a) \in u \wedge (x, b) \in u \Leftrightarrow aV_sb.$$

ただし x はある姓の漢字表記，u はある NVU．

このとき「ゆれに基づく一致」は，問い合わせ文字列を q とすると，q に対して，

$$\forall x \in A(q),\ \text{ただし}\ A(q) = \{x | x \in S, qV_sx\}$$

なる x を求めることである．

$A(q)$ は，NVU から具体的に求めることができる．例えば問い合わせが「かくた」であれば，「角田」「額田」などの NVU から，

A（“かくた”）
＝{かくた，かくだ，かどた，かどだ，がくた}

となる．この集合は問い合わせに対してゆれの関係にある姓のかな表記を網羅している．この集合を次の集合と比べてみよう．

X（“かくた”）
＝{かくた，かくだ，かどた，かどだ，がくた，
すだ，すまだ，すみた，すみだ，つのた，
つのだ，ぬかた，ぬかだ，ぬがた}

X（“かくた”）は，“かくた”の「漢字表記」と同じ「漢字表記」を持つ「かな表記」を集めたものであるが，ゆれを求めるためには精度が低い．NVU を明らかに

したことによって「ゆれに基づく一致集合」$A(q)$が求められ，ゆれの関係をとりだすことが可能になった．以下$A(q)$に基づいて検索を行う方法について述べる．

4.2 正規化に用いる同値類の定義

姓のかな表記の集合Sにおける関係Vを次のように定義する．

$\exists h_1, h_2, \cdots, h_n \in S;$

$aV_s h_1 \wedge h_1 V_s h_2 \wedge \ldots \wedge h_n V_s b \Leftrightarrow aVb.$

Vは推移律を満たすので同値関係である．またこの定義から$aV_s b \Rightarrow aVb$であることが分かる．

関係Vから次の検索が定義できる．問い合わせ文字列をqとすると，

$\forall x \in B(q)$，ただし$B(q)=\{x|x \in S, qVx\}$.

なるxを求める．この$B(q)$はSにおけるqの同値類であり，2章で述べた同値性問題の枠組で正規化処理を行うことができる．

同値類集合$B(q)$と，ゆれに基づく一致集合$A(q)$の間には$B(q) \supseteq A(q)$なる関係があるため，正規化をして検索を行った場合，ゆれの関係にあるかな表記はすべて検索される．ただし同じNVUに属さないかな表記も検索される可能性がある．このことについては5章で分析する．

4.3 正規化規則の作成

NVUは数万あるので，NVUから正規化の変換規則を正しく作成する作業は容易ではない．それゆえ正当性がある操作で，変換規則が具体的に求められることが必要である．

同値関係から変換規則を求めるアルゴリズムはKnuth-Bendixの完備化アルゴリズムとして知られている．このアルゴリズムの基本は，1つの文字列が異なる2つの文字列に書き換えられる場合，この2つの文字列が等しくなるような規則の追加を繰り返すことである．この完備化アルゴリズムを使って，姓のかな表記のゆれのすべての変形から正規化規則を作成した．

2章で述べたとおり，提案する正規化法は，姓のかな表記をVによる同値類に分けることであり，同じ同値類のかな表記が1つの正規形に変換される変換規則を求めることが必要である．扱いたいゆれの関係は，3章で明らかにしたすべてのNVU（姓のゆれ単位）の任意の2かな表記を関係Vで結んだものである．例えば，表5の2番目のNVUは，次のように表現される．

“こうしろ“V”こうじろ”，
“こうしろ“V”こおしろ”，
“こうじろ“V”こおしろ”

こうしてすべてのゆれの関係21,276組から正規化規則を求める処理を行った．その結果，もともとKnuth-Bendixの完備化アルゴリズムは停止性が保証されていないが，Vについては処理は停止し，ゆれの正規化規則15,841組を得た．処理する際に1つの文字列全体を1つのシンボルとして扱った（“こうしろ”と“こうじろ”を別個のシンボルと扱う）．与えたゆれの関係の集合を**図1**に，完備化アルゴリズムによって作成したゆれの正規化規則を**図2**に示す．

{ “あいいそ”V“あいそ”, “あいうら”V“あいのうら”, “あいかさ”V“あいがさ”, “あいかみ”V“あいがみ”,
“あいかわ”V“あいがわ”, “あいかわ”V“あわかわ”, “あいかわ”V“かいがわ”, “あいか”V“あきか”,
“あいか”V“あきしか”, “あいきよう”V“あゆきよう”, “あいき”V“あいぎ”, “あいくち”V“あいぐち”,
“あいさか”V“あいざか”, “あいさき”V“あいざき”, “あいさわ”V“あいざわ”, “あいざわ”V“あゆさわ”,
“あいざわ”V“あゆざわ”, “あいしま”V“あいじま”, “あいず”V“あいつ”, “あいず”V“あいづ”,
“あいず”V“かいづ”, “あいせき”V“あいぜき”, “あいせん”V“あいぜん”, “あいた”V“あいだ”,
“あいた”V“かいた”, “あいた”V“かいだ”, “あいだ”V“かいた”, “あいだ”V“かいだ”,
“あいつき”V“あいづき”, “あいつ”V“あいづ”, ...
}

図1 かな表記のゆれの等式集合
Fig. 1 Equivalence relations.

{ “あいのうら”⇒“あいうら”, “あいがさ”⇒“あいかさ”, “あいがみ”⇒“あいかみ”, “あいがわ”⇒“あいかわ”,
“あわかわ”⇒“あいかわ”, “かいかわ”⇒“あいかわ”, “かいがわ”⇒“あいかわ”, “かゆかわ”⇒“あいかわ”,
“かゆがわ”⇒“あいかわ”, “あきか”⇒“あいか”, “あきしか”⇒“あいか”, “あゆきよう”⇒“あいきよう”,
“あいぎ”⇒“あいき”, “あいぐち”⇒“あいくち”, “あいざか”⇒“あいさか”, “あいざき”⇒“あいさき”,
“あいざわ”⇒“あいさわ”, “あゆさわ”⇒“あいさわ”, “あゆざわ”⇒“あいさわ”, “あいじま”⇒“あいしま”,
“あいつ”⇒“あいず”, “あいづ”⇒“あいず”, “かいず”⇒“あいず”, “かいつ”⇒“あいず”,
“かいづ”⇒“あいず”, “あいぜき”⇒“あいせき”, “あいぜん”⇒“あいせん”, “あいいそ”⇒“あいそ”,
“あいだ”⇒“あいた”, “かいた”⇒“あいた”, ...
}

図2 かな表記のゆれの正規化規則
Fig. 2 Normalization rules.

5. 検索処理への適用

5.1 検索の定義

4章で求めた正規化規則を使った検索を以下のように定義する．ゆれの正規化規則集合を C とする．

ゆれの正規化規則を使った検索 問い合わせ文字列の C による正規形に対して登録情報の文字列の C による正規形が一致するものを探す．

正規形 文字列 w_1 に対して変換規則の集合 R を適用して変換文字列 w_2 を得る．この操作を w_2 に適用する規則がなくなるまで繰り返す．w_2 を w_1 の R による正規形と呼ぶ．

5.2 評 価

ゆれの正規化規則を使って，レコード数3,000万件の姓データベースを検索する実験を行った．本節では，実験結果を適合率と再現率の期待値から評価する．

問い合わせの入力は，データベースに存在するかな表記全てが，その出現頻度に応じて現れる（例えば「さとう」は全問い合わせの1.5%になる）と仮定した．文字列 q が入力される確率を P_q，ある照合一致基準が定義された時，問い合わせ q に対する検索レコード数を M_q とすると，検索レコード数の期待値は次の式で表される．

$$E=\sum_q P_q M_q.$$

ただし，入力に関して前記の仮定をおいたので，P_q は q をかな表記とするレコードのデータベース中の存在比率として求めることができる．

実験は照合一致基準として，(1)提案する「ゆれの正規化法」，(2)表2に示した濁音の清音化などを行う「従来の手法」，(3)問い合わせと同じ文字列のみを照合する「完全一致法」について行った．その結果，検索レコード数の期待値は，順に274，222，216（10万人あたりの検索に換算）となった．

この結果に対して，ゆれによる見落としを防いだ効果と，それに伴って無駄なレコードが検索された程度の2点を調べる．この2点は，「問い合わせに対してゆれの関係にあるものであれば，適合する(relevant)」と判断する場合，再現率，適合率となる．この場合，前章で求めた「ゆれに基づく一致集合 $(A(q))$」は，問い合わせ q に対して，「適合する」と判断される集合である．従って，入力 q に対して，それぞれの照合一致基準による検索レコードの集合を $R(q)$ とすると $R(q)$ と $A(q)$ から，

$$\text{再現率}=\mathrm{N}(R(q)\cap A(q))/\mathrm{N}(A(q))$$
$$\text{適合率}=\mathrm{N}(R(q)\cap A(q))/\mathrm{N}(R(q))$$

として計算することができる（ただし $\mathrm{N}(S)$ は，集合 S の要素のデータベース中の存在数とする）．これらの期待値は，検索レコード数の期待値と同様の計算で求めることができる．以上の結果を**表7**に示す．完全一致法では再現率は85%であり，適合するレコードの15%が検索にもれていたことがわかる．検索もれの解消は，再現率を91%まで上げている従来の手法でも不十分であり，本手法の有効性を確認した．この時の適合率は93%であった．実際のシステムにおいては，いったん検索したレコードから不要なものを落とす方法はたくさんある．一方，一度落ちてしまった情報については対処の方法がない．それゆえ適合率より再現率を重要視する方法が有用である．

表7 ゆれの正規化規則を使った検索の期待値
Table 7 Expectations of the retrieval using the variants normalization rules.

照合一致基準	検索レコード数	適合率	再現率
ゆれの正規化法	274	0.93	1
従来の手法	222	0.99	0.91
完全一致法	216	1	0.85

検索レコードは10万人あたりの件数に換算して示す．

5.3 姓のかな表記の同値類

正規化は姓のかな表記の集合を互いに交わりのない同値類に分解する処理である．この処理で全60,283種の姓のかな表記が44,442の集合に類別された．「やまもと」，「あおき」など34,500種の姓は単独で同値類を作っており，ゆれに対して安定した姓であるといえる．次に「やまさき」と「やまざき」などのように15,042種の姓が，2つのかな表記で同値類を作っている．一方10を越えるかな表記が連結してできた同値類が81種，1,564語存在した．かな表記数が最大となった正規化の同値類は正規形が「ゆ」となる集合で表記数は109である（**図3**）．これらの存在が適合率を下げる原因となっている．

5.4 実データベースへの適用

提案する照合法を実際のデータベースに適用する時の影響を述べる．姓データベースから，姓の実データ上の重みづけを行った上で無作為に抽出して，レコード数10,000，かな表記の異なり語数3,540種の試験データベースを作成し，これを用いてゆれの正規化規則を使った検索の実験を行った．本電話帳データベースの場合，統計的には母集団が10,000を越えると1検索あたりの検索レコード数の期待値が20を越えるため住所などの他の条件を付加して対象を少なくとも10,000以下にする必要がある．検索の例を**表8**に示す．

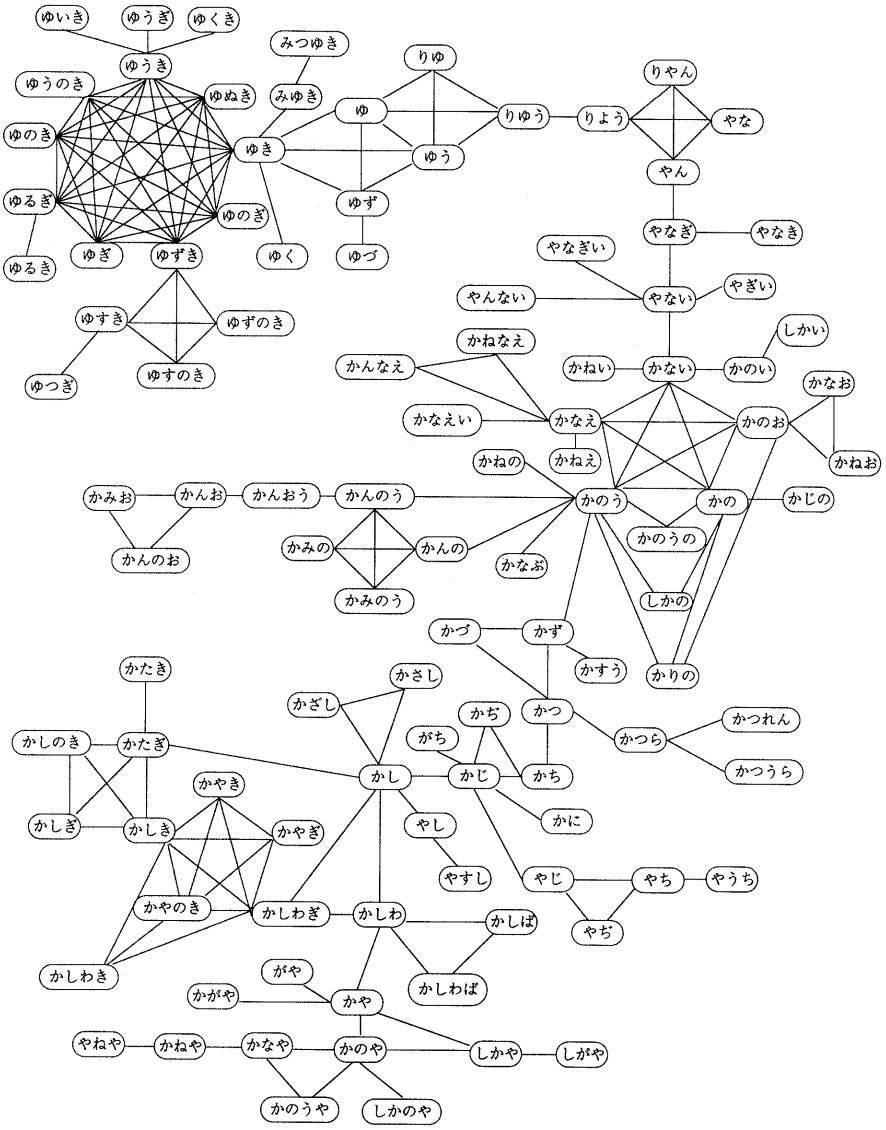


図3 ゆれの最大の同値類
Fig. 3 The maximum equivalence set of the *Kana* variants.

表8 検索例
Table 8 Example of retrievals.

入力	出力
おはら	おおはら，おはら，おばら，こはら，こばら，こばる
とみやま	とうやま，とおやま，とみやま，とやま
わたなべ	わたなべ，わたべ
あがづま	あがつま，あずま，わがつま
かしはら	かしはら，かしわばら，かじはら，かじわら

このように検索結果は4.1節で求めたゆれに基づく一致集合に該当するかな表記をすべて含み，ゆれによる検索もれを解消した．この検索の検索レコード数の期待値は27.6件で，完全一致の場合は20.1であり，この差7.5件の中にゆれの存在によって検索からもれていたレコードが含まれている．照合一致されるゆれの文字種が多種へ拡張されているだけでなく，「かしはら」に対する「かしわばら」などの長さの変化を伴うゆれへの対処も可能となっている．

最後に正規化による悪影響の可能性について検索レコードの数と種類の観点から述べる．最大の検索レコード数を示すものは正規形が「さとう」となる検索で232件である．これは表記が「さいとう」なる「佐藤」姓の存在を介して「さとう」(153件) と「さいとう」

(79件)が合流したためである．一方，最多種の姓の合流は図3で示した正規形が「ゆ」となる検索で，実際に検索された姓は24種，検索レコード数は49件にすぎない．この他の例でも多数の合流による検索レコード数の爆発的増加はおきなかった．もしこれらの正規化による合流を避けたいならば，頻度の少ないかな表記を持つ姓に対して，データベース上で代表的なかな表記を別名（alias）として登録しておき，その姓にかかわる変換を正規化規則から外すことが可能である．こうすることによってデータベースの規模は大きくなるものの正規化規則数および望ましくない合流を減じることができる．これは具体的なシステム設計時の最適化の課題である．

6. おわりに

人名のかな表記のゆれによって生じる日本語情報検索における問題点を解決するために，姓のかな表記のゆれを解析し，それに基づいた近似文字列照合法を提案した．

人名のゆれを明らかにするために総数3,000万件の姓を使い，かな表記の対立を変形位置，音種，形態の観点から35のカテゴリーに分類してゆれの判定を行い，ゆれを具体的に求めた．その結果姓は約9万種の姓のゆれ単位に分類でき，実データ上で58%の姓に何らかのゆれがあることが分かった．ゆれの原因は連濁などの漢字接続部の音の変化が大部分を占めることが分かった．

このゆれの関係に基づいた照合を行うために，ゆれの関係を21,276組の文字列の等式関係で記述し，そこからKnuth-Bendixの完備化アルゴリズムを使って15,841の正規化規則を作成する方法を提案した．この規則を使ってゆれの関係にある文字列を正規形に変換し評価する正規化処理による照合法を提案した．

この検索方法を3,000万件のデータベースの検索に適用して，再現率，適合率の観点から評価を行った．再現率の観点からは，文字列の完全一致による検索では平均15%存在していたゆれによる検索もれを解消することができ，さらに平均93%の適合率を得た．これらの正規化規則は日本人3,000万人のレベルで考え得るすべてのゆれを包含しており，稀な表記のゆれがあっても1回の検索でもらさず検索することができる．これは網羅性を必要とする検索に対して重要な性質である．

今後は，利用者による実験等により本手法の操作性の観点からの評価と，明らかにしたゆれの関係の表記の対立をさらに解析し，漢字姓にとどまらずより広い検索対象に適用できる近似文字列照合法の確立することが課題である．

謝辞 日頃議論していただく情報案内方式研究グループの方々に感謝します．また検索に関する数々の問題点を教示していただいたNTTの104番号案内業務（ANGELシステム）にたずさわる方々に感謝します．

参考文献

1) 国立国語研究所：現代表記のゆれ，国立国語研究所報告75，秀英出版（1983）．
2) Cleventon, C. W.: Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems, *Aslib-Cranfield Reseach Report*, Cranfield (1962).
3) 藤澤浩道，絹川博之：情報検索における自然言語処理，情報処理，Vol. 34, No. 10, pp. 1259-1265 (1993).
4) 田中康仁：姓名のカナ漢字変換システム，情報処理，Vol. 16, No. 3, pp. 230-238 (1975).
5) 田中康仁：カナ文字のデータ•チェック—特に名前のエラーチェックについて，ユニバック研究会，SYSTEMS No. 122, pp. 41-47 (1975).
6) 高橋克巳，岩瀬成人：人名の読みからの検索法，情報処理学会自然言語処理研究会資料，Vol. 92, No. 72, 92-NL-91-4, pp. 25-32 (1992).
7) 宇土行良，遠山 潤，森 茂樹：日本人著者の名前についての調査，図書館学会年報，Vol. 39, No. 1, pp. 13-24 (1993).
8) Hall, P. A. V. and Dowling, G.: Approximate String Matching, *Comput. Surv.*, Vol. 12, No. 4, pp. 381-402 (1980).
9) Robbins, C. J. and Kille, S. E.: The ISO Development: User's Mannual Volume 5: QUIPU, p. 35 (1991)
10) Odell, M. K. and Russell, R. C.: U. S. Pat, 1-261-167 (1918), 1-435-663 (1922).
11) 宮部 博，大山 実，本郷郁夫：名義検索システム—電話番号案内業務への適用—，情報処理学会論文誌，Vol. 24, No. 4, pp. 421-428 (1983).
12) 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, pp. 841-850 (1990).
13) 獅々堀正幹，津田和彦，青江順一：片仮名異表記の生成，電子情報通信学会論文誌，Vol. J 77-D-II, No. 2, pp. 380-387 (1994).
14) Knuth, D. E. and Bendix, P. G.: *Simple Word Problem in Universal Algebras, Computational Problems in Abstract Algebra,* pp. 263-297, Pergamon Press (1970).
15) 丹羽基二，日本ユニバック：日本姓氏大辞典，角川書店（1985）．
16) 文化庁国語課，国語研究会編集：国語表記事務必

携[改定版], ぎょうせい (1992).
17) 湯沢質幸: 漢字の慣用音, 漢字講座 3, 漢字と日本語, pp. 82-111, 明治書院 (1987).
18) 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), pp. 223-265, 明治書院 (1989).
19) 藤堂明保編: 漢和大辞典, 学習研究社 (1978).
20) 小松英雄: 日本語の音韻, 日本語の世界 7, 中央公論社 (1981).
21) 国立国語研究所: 読みの実験的研究—音読にあらわれた読みあやまりの分析—, 国立国語研究所報告 9, 秀英出版 (1955).
22) 国立国語研究所: 語彙の研究と教育(上)(下), 大蔵省印刷局 (1985).
23) 国立国語研究所: 文字・表記の教育, 大蔵省印刷局 (1988).

付　　録

以下では，3章で述べた，姓を「姓のゆれ単位」(NVU) に分類する手順について述べる．

A.1 手　　順

(1)漢字とその語中の出現位置単位にすべてのかな表記を使用度数 (使われている姓の種類) を付与して集める．(2)漢字単位にかな表記間の対立を変形位置，音種，形態のカテゴリーに分類する．(3)カテゴリー単位にゆれの関係が含まれているかを調べる．(4)漢字単位のかな表記のゆれ関係を導く(「漢字かな表記辞書」)．(5)姓を NVU に分類する．

A.2 漢字単位のゆれの判定

(1)から(3)までの調査の結果を**表 9** に示す．以下で説明する基準に基づき，判定の項目に○印を付したカテゴリーにゆれの関係が存在すると判定した．

それぞれのカテゴリーでゆれと判断した基準を具体的に説明する．まず漢字の基本的な音訓の枠組，訓の活用の有無を，常用漢字表や漢和辞典[19]などで判断した．その上でゆれはかなづかいの対立[1]や音韻の変化[20]，読みあやまり[21]などの典型的パターンが定着したものであると考え，ゆれが起こるケースとして次の3項を設けた．(1)かなづかいの変化．(2)末尾の変化．(3)漢字接続部の変化．

A.2.1 かなづかい

表9のカテゴリー31，32(以下C31などと記する)．かなづかいのゆれとして「ぢ/じ，づ/ず」と二重母音「おう/おお」の書き分けが存在する．この2つは語中のあらゆる位置で現れた．

A.2.2 末尾の対立

C1-6, C19-24．これらは語幹が同じでありこのカテゴリーに属する関係をゆれと考えた．ただしC3に分類される「き-こ (木)」の例は語幹が同じではないが母音交替として知られた現象であるのでこれもゆれと扱う．

表 9 かな表記の対立のカテゴリーとゆれの判定
Table 9 Difference categories and judgements on the *Kana* variants.

	変形位置	音種別	形態	例	判定
1	連接部	子音	交替	かわ-かは(川-)	○
2	前項末尾		脱落/添加	ほう-ほく(北-)	○
3		母音	交替	あめ-あま(雨-)	○
4			脱落/添加	そう-そ(曽-)	○
5		その他	交替	こえ-こし(越-)	○
6			脱落/添加	かわ-か(川-)	○
7	連接部	子音	交替	はら-わら(-原)	○
8	後項先頭		脱落/添加	おん/のん(-音)	○
9		母音	交替	ざい-ぜい(-西)	×
10			脱落/添加	うみ-み(-海)	○
11		その他	交替	かん-じん(-神)	×
12			脱落/添加	みなみ/なみ(南)	○
13	先頭	子音	交替	しち-ひち(七-)	○
14			脱落/添加	わが-あが(吾-)	○
15		母音	交替	きん-こん(金-)	×
16			脱落/添加	いで-で(出-)	○
17		その他	交替	はぎ-おぎ(萩-)	×
18			脱落/添加	むこう-こう(向)	○
19	末尾	子音	交替	ばた-ばな(-端)	○
20			脱落/添加	かけい-かけひ(筧)	○
21		母音	交替	たて-たち(-館)	○
22			脱落/添加	のう-の(-能)	○
23		その他	交替	かど-かく(-角)	○
24			脱落/添加	たいら-たい(-平)	○
25	内部	子音	交替	つむら-つぶら	○
26			脱落/添加	いおり-いほり(庵)	○
27		母音	交替	えびす-えべす(胡)	○
28			脱落/添加	おぎ-おおぎ(扇)	○
29		その他	交替	はりのき-はんのき(榿)	○
30			脱落/添加	やぎ-やなぎ(柳)	○
31	任意	かなづかい	ぢ/じ，づ/ず	ふじ-ふぢ(藤-)	○
32			おう/おお	とうり-とおり(通)	○
33	上記以外の1文字共通		先頭文字が共通	あがり-あげ(上)	○
34			その他	こく-くろ(黒)	×
35	共通文字列なし		さん-やま(山)，たか-こう(高)		×

A.2.3 連接部の対立

C1-6，C7-12．連接部後項の変化は語幹の変化が起こる．このうちC7，C8，C10，C12には語幹の対立のうち一般語彙でも存在する変化であり順に説明する．

連接部の音の交替としては，先頭の子音が清音から

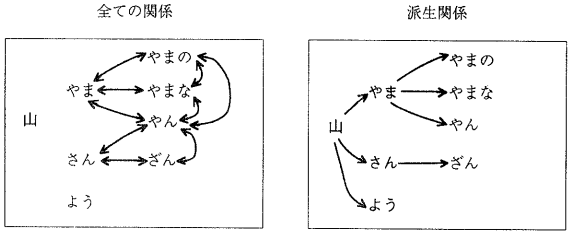


図4 派生関係の抽出
Fig. 4 Finding derivative relations.

濁音に変化する連濁が知られており，ゆれの代表的なものである（C 7 の一部）．それ以外の子音の交替では派生の認められるものに「は，わ，ば，ぱ」（はら-わら（原））の対立などがある．他方辞書で別個の音としてあつかわれているものに「ぶ-む（武）」などがあるが，カテゴリー全体としてはゆれを含んでいる．同様のことが子音の脱落/添加にもあてはまる．この連濁以外の C 7 と C 8 を条件つきでゆれに分類する．しかし C 9，C 11 には派生関係が見い出せなかった．

同じく音の脱落/添加には「の・が・つ・な・て・ん」の添加が多く見受けられる．これらは助詞の役割を果たしていると考えゆれとみなす（C 12 の一部）．その他の脱落/添加にも純粋な音の省略が含まれており C 10，C 12 の残りを条件つきでゆれとみなす．

なおこれらの連接部の対立は転音，音便，音韻の添加，脱落，融合などの複合語の変音現象として国語研究所の文献[22),23)]などに詳しい．

A.2.4 その他の対立

先頭部の対立は語幹の変化であるため原則としてゆれと認めない（C 15，C 17）．ただし清音と濁音の対立をゆれと認めた（C 13 の一部）．他は連接部後項の先頭と同様に考えて，C 13 の残りと C 14，C 16，C 18 を条件つきでゆれとみなす．

それ以外の関係では，語幹が変化しない内部での 1 文字の交替，脱落/添加のすべて（C 25-30）と，頭部で 1 文字以上の共通文字列があるもの（C 33）のみをゆれと考え，他はすべて除外した（C 34，C 35）．

A.3 漢字かな表記辞書

前節で判断した関係に次の規則を入れて実際に派生関係として機能しているものをとりだした．この規則は，「漢字には複数の代表的な表記があり，それから音の変化や脱落，活用などで派生が生まれている」と仮定して作成した．派生関係を取り出す処理を**図4** に示す．

1. 使用度数の多いものから少ないものへ派生する．
2. 2 種類以上の親表記から派生を受けることはない．
3. 2 種類以上の親表記の候補が存在する時は，語幹が変わらないもの，使用度数の多いものを優先する．

「やま」「さん」「よう」のように派生の先頭に位置するものが，漢字の代表的な表記であり（見出し表記と呼ぶことにする），それ以外の表記は見出し表記に対する派生と扱った．このことから**表 10** に示す漢字かな表記辞書を得た．

表 10 漢字かな表記辞書
Table 10 *Kanji-Kana* dictionary.

漢字	位置	表記
田	先頭	た，たの，たん，たな，たつ，だ でん とう，とお
上	末尾	がみ，かみ うえ，え，う じょう，しょう あげ，あがり

下線を付した語は見出し表記

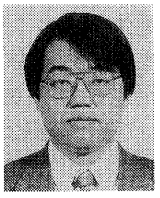
A.4 姓のかな表記のゆれ

漢字かな表記辞書を使って任意の姓がゆれの関係にあるかを調べることができる．すなわち同一の漢字を持つ姓のうちで，かな表記を漢字単位に見出し表記に変換したとき，見出し表記が同じになるものがゆれの関係にある．この結果得られた NVU を本文の表 5 に示す．

（平成 6 年 6 月 10 日受付）
（平成 7 年 4 月 14 日採録）

高橋 克巳（正会員）

1964 年生．1988 年東京工業大学理学部数学科卒業．同年日本電信電話株式会社入社．以来番号案内システムの研究開発に従事．NTT 情報通信研究所知的情報処理研究部研究主任．ACM 会員．



梅村 恭司（正会員）

1959 年生．1983 年東京大学大学院情報工学専攻修士課程修了．1983 年より 1995 年まで NTT 研究所所属．1995 年 4 月より豊橋技術科学大学情報工学系助教授．工学博士．記号処理，プログラミング言語，統計言語処理などに興味をもつ．ACM，ソフトウェア科学会，電子情報通信学会，計量国語学会各会員．