人文科学とコンピュータ　40－8
（1998．10．23）

# 和歌データベースからの類似歌の自動抽出

山崎 真由美† 竹田 正幸† 福田 智子‡ 南里 一郎*
† 九州大学大学院システム情報科学研究科 ‡ 福岡女学院大学 * 純真女子短期大学

**要旨.** 和歌文学研究において，歌の類似性の抽出は重要である．歌の類似性に着目することにより，過去や同時代の歌人による作品への影響を明らかにすることができ，また歌人の個性や時代による特徴を獲得することができる．
従来，この類の研究は，任意の歌もしくは表現にまず注目し，次にその用例を収集するという方法で進められてきた．だがもし，大量の和歌のデータの中から類似歌を自動抽出することができれば，その類似歌の発見が契機となって新たな視点が得られ，研究の大きな進展につながることも期待できるのである．
本論文では，大量の和歌データを対象に，計算機による類似歌の自動抽出を目指し，そのために必要な類似性の指標を提案する．提案した指標は，最長共通部分列に基づく指標を改善したものである．本方式を用いて，古今集と新古今集からの類似歌抽出を試みたところ，類似度の高いものの多くは，実際に本歌取りであり，また，主な注釈書には指摘が漏れている本歌取りも指摘できることが判明した．

# Finding Similar Poems from Classical Japanese Poem Database

Mayumi Yamasaki† Masayuki Takeda† Tomoko Fukuda‡ Ichiro Nanri*
† Department of Informatics, Kyushu University
‡ Fukuoka Jo Gakuin College * Junshin Women's Junior College

**Abstract.** In this paper we consider a problem of automatically finding similar poems from a collection of classical Japanese poems. We propose two similarity measures, and show that they are superior to the similarity measure based on the longest common subsequence. We report successful results in finding similar poems between two imperial anthologies: KOKINSHŪ and SHINKOKINSHŪ.

## 1 まえがき

和歌文学研究において，歌の類似性の抽出は重要である．歌の類似性に着目することにより，過去や同時代の歌人による作品への影響を明らかにすることができ，また歌人の個性や時代による特徴を獲得することができる．

従来，この類の研究は，任意の歌もしくは表現にまず注目し，次にその用例を収集するという方法で進められてきた．そこでは，研究者がどのような歌もしくは表現に着目するかが，一つの鍵となる．だがもし，大量の和歌のデータの中から類似歌を自動抽出することができれば，上述のような研究を網羅的に行うことができる．また，類似歌の発見が契機となって新たな視点が生まれ，研究の大きな進展につながることも

期待できるのである．

一口に和歌の類似性といっても，歌と歌との類似の仕方にはさまざまなものが考えられる．著者らは，現在，次の3種類の類似性を考えている．

(1) 自立語が共通する．
(2) 付属語のなすパターンが共通する．
(3) (1),(2) と異なり，和歌を単なる文字の連鎖とみなした場合に共通した文字列を多く含む．

(1) の類似性については，古くから盛んに研究されてきた．しかし，(1) の類似性のみに偏った研究には，限界がある．例えば，紀貫之が桜の歌を多く詠んだからといって，「貫之は桜の歌を好んだ」と結論するのは，短絡に過ぎるであろう．というのは，当時は詠歌に際して題が与えられることがあり，歌人が歌語を任意に選択できない場合があったと考えられるからである．

そこで，著者らは，(2) の類似性に着目した [3]．例えば，「*ば*ざらましを*」などのパターンを扱うのである．これは反実仮想という表現技法に対応する．つまり，(2) の類似性は，表現技法上の特徴に対応している．そこで，最小記述長 (MDL) 原理に基づいた方法 [1] を用いて，歌集からの付属語パターン自動抽出を試みた．得られたパターンの歌集ごとの相違は，歌人の個性や時代の好みを反映しているようであり，研究者に非常に興味深い視点を提供してくれている [3]．

本論文では，(3) の類似性に焦点をあて，その自動抽出の問題を扱う．大雑把にいえば，(1) は意味，(2) は構造の類似性に対応しているが，(3) は単なる文字列としての類似性に対応する．この種の類似歌について，少し例を挙げてみよう．例えば，『古今集』巻第一春歌上，53番の在原業平の歌

世中にたえてさくらのなかりせば
春の心はのどけからまし

であるが，これが『土佐日記』の中で引かれるときには，第3句が次のようになっている．

世の中にたえて桜のさかざらば
春の心はのどけからまし

多少の異同があっても，本来は同じ歌である．では，こうした異同が生じたのは，偶然だろうか．それとも，何か理由があるのだろうか．『土佐日記』でこの歌が見えるのは，満開の桜の前を通りかかった場面である．「なかりせば」が「さかざらば」になっているのも，このような状況に合わせた，意図的な改変であったかもしれない．ここに，1首の和歌が，後世どのように享受されていくかという一面が見えてくる．

また，『兼盛集』68番

白波ののどけきうらの姫松は
千とせのかずぞそひて見えける

と，『能宣集』458番

しらなみののどけきはまのひめまつは
ちとせのかげぞそひてみえける

との場合には，別の問題提起が必要となる．すなわち，これらの歌も，下線部の異同はあるものの，元来は同一歌であったと判断されるが，それでは，なぜその同じ歌が，兼盛と能宣という，二人の歌人の個人歌集に載るのか．どちらかが，作者を誤ったのか．問題は，それぞれの歌集の性格や成立の問題にまで発展する [5]．

一方，別々に詠まれた歌であっても，一方が他方の歌を踏まえると，表現の類似性をもつことがある．『山家集』187番を例にとると，

郭公きかで明けぬる夏の夜の
浦島の子はまことなりけり

は，『拾遺和歌集』122番

夏の夜は浦島の子が箱なれや
はかなくあけてくやしかるらん

を踏まえて詠まれていることが指摘できる．この場合，「浦島の子が箱」は「あ (開) けてくやし」いものであり，その「あ (明) けてくやし」いのが「夏の夜」であるという『拾遺和歌集』の歌を知らなければ，『山家集』の下の句，「浦島の子はまことなりけり」の意味が理解できない．1首の解釈という，和歌研究の最も基礎的でかつ重要な目的の一つが，類似歌を考慮することにより，達せられることもある．

このように，表現（文字列）が一致あるいは類似する歌は，和歌文学研究において，興味深

い視点を提供してくれる．すなわち，(3) の類似歌は，同一の歌であったものが，伝来の過程で本文が改変されたものであるかもしれない．そこで，伝来の過程で異同を生じた理由を追求することにより，和歌史的考察にまで発展することが期待できる．あるいは，このような類似歌は，本歌取りといって，古歌を踏まえて新たに歌を詠んだものであるかもしれない．この場合には，本歌取りにおいてどのように古歌を踏まえているかを分析することを通じて，表現技法に関する知見を得ることができよう．いずれにせよ，類似歌の発見を契機にして，新たな視点が得られ，研究の大きな進展につながると期待できる．なお，ここでいう本歌取りとは，俊成・定家によって確立された表現技巧を指すのではなく，より広義に，先行歌を踏まえた作歌手法を指すものとする．

本論文では，(3) の類似歌を計算機により自動抽出することを目指す．このためには，和歌間の類似度をいかに定義するかが成功への鍵となる．本論文で提案する類似性の指標は，最長共通部分列に基づく指標を改善したものである．本方式を用いて，古今集と新古今集からの類似歌抽出を試みたところ，類似度の高いものの多くは，主な注釈書において本歌取りと指摘されていることが確認できた．また，それらの注釈書に指摘のないものの中には，本歌取りと指摘されてしかるべき歌も含まれていた．すなわち，本方式は，従来の研究成果に加えて，これまで看過されていた本歌取りをも指摘できる可能性をもっている．

## 2 和歌の類似度

この章では，研究の第 1 段階として，文字列間の最長共通部分列 (longest common subsequence)[2] の長さに基づいて和歌の類似度を定義し，その性質を考察する．

### 2.1 最長共通部分列

文字列$\xi$の長さを $|\xi|$ で表し，$\xi$の $i$ 番目の文字を$\xi[i]$ で表す．$\xi$と$\tau$を長さ 1 以上の文字列とする．$1 \leq i \leq |\xi|$ を満たす任意の $i$ に対して，$\xi[i] = \tau[k_i]$ となるような整数の単調増加列 $k_1, \ldots, k_{|\xi|}$が存在するとき，$\xi$を$\tau$の**部分列** (subsequence) という．文字列 $x, y$ に対し，$\xi$が両方の部分列となっているとき，$\xi$を $x, y$ の**共通部分列** (common subsequence) という．文字列 $x, y$ に対する最長の共通部分列を $x, y$ の**最長共通部分列** (longest common subsequence; LCS) という．例えば，$x = abcbdda, y = badbabd$ とするとき，文字列 $abbd$ は $x, y$ の LCS であり，その長さは 4 である．

LCS は，文書処理におけるミススペルの修正や，ゲノム情報処理における塩基配列やアミノ酸配列の近似的照合においてよく用いられる編集距離 (edit distance) [2] とも密接な関係をもつ．実際，許容する編集操作 (edit operation) を文字の挿入と削除に限定したとき，編集距離は，二つの文字列の長さの和から LCS の長さの 2 倍を引いたものに一致する．

### 2.2 LCS に基づく類似度

最も単純には，二つの和歌の LCS の長さを類似度とする方法が考えられる．しかし，本歌取りなどの場合には，対応する句の位置が変化することが多い．そこで，$5! = 120$ 通りの句の対応づけのうちで，対応する句の間の LCS の長さの総和を最大にする対応付けを考え，そのときの値を**類似度**とする．次の例に示す 2 首では，『古今集』147 番歌の句に対して，『新古今集』216 番歌の句が，1,4,5,2,3 と対応している．

**例 1**
ほととぎす ながなくさとの あまたあれば
猶うとまれぬ 思ふものから (古今集#147)
ほととぎす 猶うとまれぬ 心かな
ながなく里の よその夕ぐれ (新古今集#216)

句ごとに求めた LCS の長さを表 1 に示す．この 2 首の類似度は，句ごとの LCS 長を合計した 21 ということになる．

### 2.3 実験とその結果

LCS に基づく類似度の有効性を評価するためには，十分な量の類似歌および非類似歌のデータが必要である．そこで，慈円の『拾玉集』の

表 1: 句ごとのLCS長（例1）

| 古今集#147 | 新古今集#216 | LCS長 |
|---|---|---|
| ほととぎす | ほととぎす | 5 |
| なかなくさとの | なかなくさとの | 7 |
| あまたあれは | よそのゆふくれ | 1 |
| なほうとまれぬ | なほうとまれぬ | 7 |
| おもふものから | こころかな | 1 |

一部を利用することを考えた．すなわち，『拾玉集』の3,472番から3,571番までの100首は，『古今集』の歌を踏まえて詠まれたものであり，題詞にその本歌が示してあるため，これを類似歌の例として用いるのである．古今歌を踏まえたとはいっても，必ずしも文字列としてよく似ているとは限らないが，他に適当なデータがないのでこれを用いることにした．

『拾玉集』の上述の100首とその本歌から成る100対を，類似歌の例，すなわち正例 (positive examples) とする．また，この100首と本歌以外の古今歌とから成る9,900対を，非類似歌の例，すなわち負例 (negative examples) とする．この正例・負例のそれぞれについて類似度を算出したところ，正例の96%は類似度が10以上であり，負例の96%は類似度が10以下であった．そこで，閾値を10付近にとり，類似度が閾値以上のとき類似歌，閾値未満のとき非類似歌とする判定法が考えられよう．実際，

$$Score_{\mathrm{P}} = \frac{\text{類似度が閾値以上の正例数}}{\text{正例数}},$$

$$Score_{\mathrm{N}} = \frac{\text{類似度が閾値未満の負例数}}{\text{負例数}}$$

とおき，その相乗平均

$$Score = \sqrt{Score_{\mathrm{P}} \cdot Score_{\mathrm{N}}} \qquad (1)$$

を評価関数として用いて得た最良の閾値は，10.1であり，そのときの評価関数の値は，

$$\sqrt{0.9200 \cdot 0.9568} = 0.9382$$

であった．

## 2.4 問題点

ここでの類似度の定義は，単純にLCSの長さを類似度としており，LCSの各文字が連続している場合とそうでない場合とを区別していない．以下の例をみてみよう．

表 2: 句ごとのLCS長（例2）

| 古今集#315 | 拾玉集#3528 | LCS長 |
|---|---|---|
| やまさとは | やとさひて | 2 |
| ふゆそさひしさ | あきのしらつゆ | 1 |
| まさりける | そてにそのこる | 1 |
| ひとめもくさも | ひとめもくさも | 7 |
| かれぬとおもへは | かれぬれは | 4 |

**例 2**

山里は 冬ぞさびしさ まさりける
人めも草も かれぬと思へば（古今集#315）

やどさびて 人めも草も かれぬれば
袖にぞのこる 秋のしら露（拾玉集#3528）

句ごとのLCSの値を表2に示した．表より，「やまさとは」と「やとさひて」の「や」と「と」で2文字，「ふゆそさひしさ」と「あきのしらつゆ」の「し」で1文字，「まさりける」と「そてにそのこる」の「る」で1文字，それぞれ加点されてことがわかるが，これはほとんど無意味な加点であるようにみえる．これに対し，「かれぬとおもへは」と「かれぬれは」の間での「かれぬ」「は」で4文字一致した，というのは，意味のある加点であろう．

議論を明確にするために，二つの句の間の共通パターンを考えよう．例えば，「やまさとは」と「やとさひて」の共通パターンとしては「や*と*」があり，一方，「かれぬとおもへは」と「かれぬれは」の共通パターンとして，「かれぬ*は」がある．LCSとは，二つの文字列間の共通パターンにおける文字の個数に着目し，その数を最大にするような共通パターンを考えたものであるといえる．これに対して，パターン中の文字の個数が同じでも，文字が連続している場合には，より高い値を与えるようにすれば，上に示した問題点は解消されると考えられる．すなわち，パターン$*ab*$とパターン$*a*b*$があるとき，前者により高い値を与えるようにすればよい．

## 3 新しい類似度の指標

この章では，パターン中の文字の連続性を考慮した，パターン評価関数$\Phi$の定義を与える．まず，3.1 節では，そのために必要ないくつかの定義を行う．次に，3.2 節と 3.3 節で，パターン評価関数の具体的な定義を二つ与え，それぞれ評価を行う．

### 3.1 準備

$\Sigma$を文字の有限集合とし，$* \notin \Sigma$を間隙記号（gap symbol）とする．$\Sigma^*$の要素を**文字列**とよび，$(\Sigma \cup \{*\})^*$の要素を**パターン**とよぶ．パターン$\pi$中の$*$の生起を，それぞれ，任意の文字列で置き換えて得られる文字列全体の集合を，パターン$\pi$の生成する言語といい，$L(\pi)$ で表す．文字列$x, y$とパターン$\pi$に対して$x \in L(\pi)$ かつ $y \in L(\pi)$ であるとき，$\pi$は$x, y$の**共通パターン**であるという．いま，パターン全体の集合$(\Sigma \cup \{*\})^*$から実数全体の集合$\mathcal{R}$への写像$\Phi$を考え，これを**パターン評価関数**とよぶことにする．この関数をもとにして，任意の文字列$x, y \in \Sigma^*$に対して，$x, y$の類似度を次のように定義する．

$$\mathrm{SIM}_\Phi(x, y) = \max\{\Phi(\pi) | x, y \in L(\pi)\} \quad (2)$$

上式において，$\Phi(\pi)$ を，パターン$\pi$中の文字の個数と定めれば，$\mathrm{SIM}_\Phi(x, y)$は，文字列$x; y$のLCSの長さに一致する．類似度$\mathrm{SIM}_\Phi$の「良い」定義を与えるためには，パターン評価関数$\Phi$をどのように定義するかが鍵となる．

### 3.2 パターン評価関数の定義 (1)

パターン評価関数$\Phi$を，例えば，パターン$*ab*$とパターン$*a*b*$に対して

$$\Phi(*ab*) > \Phi(*a*b*)$$

となるように定めたい．そのための自然な定義として，以下のような準同型写像が考えられる．

$$\begin{cases} \Phi(c) & = & 1 \quad (c \in \Sigma) \\ \Phi(*) & = & -s \\ \Phi(\pi_1 \cdot \pi_2) & = & \Phi(\pi_1) + \Phi(\pi_2) \\ & & \quad (\pi_1, \pi_2 \in (\Sigma \cup \{*\})^*) \end{cases} \quad (3)$$

ここで，$0 < s < 1$とする．すなわち，この定義は，パターン中の$\Sigma$の文字についてはLCSの場合と同様1点ずつ加点するが，パターン中に$*$が出てくる度にペナルティとして$s$だけ減点するものである．こうすることにより，パターン中の文字列が細切れであるものには低い値が付与されることになる．

$x, y$を文字列とし，$0 \le i \le |x|$，$0 \le j \le |y|$なる整数$i, j$に対して，

$$S_{i,j} = \mathrm{SIM}_\Phi(x[1..i], y[1..j])$$

とおく．明らかに，$S_{0,0} = 0$，$S_{i,0} = S_{0,j} = -s$ $(0 < i \le |x|, 0 < j \le |y|)$ である．また，$0 < i \le |x|$ かつ $0 < j \le |y|$ のとき，$S_{i,j}$の値は，次の三つのうちの最大値と一致することが示せる．

(1) $S_{i-1,j} - s \cdot \delta(x[i-1], y[j])$.

(2) $S_{i,j-1} - s \cdot \delta(x[i], y[j-1])$.

(3) $S_{i-1,j-1} + 1$ ($x[i] = y[j]$ のとき)，
$S_{i-1,j-1} - s \cdot \delta(x[i-1], y[j-1])$ (そうでないとき).

ここで，便宜上，$x[0] = y[0] \notin \Sigma$としておき，$\delta(a, b)$は文字$a, b$が等しいとき1，そうでないとき0を返す関数とする．以上より，$\mathrm{SIM}_\Phi(x, y)$の値は，$x, y$のLCSの長さを求める場合と同様，$O(|x| \cdot |y|)$ の領域を用いて，$O(|x| \cdot |y|)$ 時間で求めることができる．

さて，ペナルティである$s$の値を定めるために，$s$の値を変化させて，2.3 節で述べた正例と負例に対して1式の評価関数の値を最大にする閾値とそのときの最大値とを求めた．その結果，$s = 0.9$付近で評価関数の値は最大値

$$\sqrt{0.9600 \cdot 0.9528} = 0.9564$$

をとり，そのときの閾値は6.6であった．

### 3.3 パターン評価関数の定義 (2)

上で述べたパターン評価関数の定義のほかにも，条件を満たす定義は可能である．

パターン中の連続文字列の長さに注目しよう．例えば，$\pi = *a*bc*d*$においては，左から，1, 2, 1である．正整数全体の集合$\mathcal{N}$から正実数全体の集合$\mathcal{R}^+$への写像$f$を仮定し，パターン評価関数の値を$\Phi(\pi) = f(1)+f(2)+f(1)$のようにすることを考えよう．ここで，特に，$f(\ell) = \ell$ $(\forall \ell \in \mathcal{N})$ とすれば，$\Phi(\pi)$ は，$\pi$中の文字の個数に一致する．文字が連続している場合に大きい値を与えるようにするためには，任意の正整数$n, m$に対して

$$f(n+m) > f(n) + f(m) \qquad (4)$$

でなければならない．この条件を満たす$f$は無限に存在する．まず，$f(\ell)$を$\ell$の1次関数に限定して考えよう．明らかに定数倍は無視してよいから，$f$は，

$$f(\ell) = \ell - s \qquad (0 < s < 1) \qquad (5)$$

とおくことができる．パラメータ$s$を変化させて1式の評価関数を最大にする閾値の値とそのときの最大値を求めたところ，$s = 0.8 \sim 0.9$付近で最大となり，その値は，

$$\sqrt{0.9600 \cdot 0.9608} = 0.9604$$

であった．$s = 0.9$のとき，この最大値を与える閾値は，8.9であった．LCSに基づく指標を用いた場合と比べ，評価関数の値が増大していることがわかる．

$f$として，2次関数やより高次の多項式関数を用いれば，パターン中の連続文字列の長さが長くなるに従って，類似度が急激に増加するようにできる．しかし，本論文では，句ごとに類似度を計算するために，句の長さは高々5もしくは7であり，問題となる連続文字列はあまり長くならず，顕著な効果は期待できない．また，パラメータが増える分，使用したデータに過剰適合（overfitting）する危険性がある．

以上のことから，$f$として4式の1次関数を用いることにし，パラメータ$s$の値は，$0.8 \sim 0.9$とする．

## 4 類似歌の発見

LCSに基づく指標を「指標A」とよび，3.2節で定義した指標を「指標B」，3.3節で定義した指標を「指標C」とよぶことにしよう．前章で示したように，指標Bと指標Cは，いずれも，指標Aに比べ，1式の評価関数の値を大きくする．このことは，これらの指標の有効性を示す一つの根拠である．

そこで，今度は二つの歌集間で，すべての歌の組合せについて類似度を算出し，用いた指標による順位の変動を調べるとともに，順位の高いものについては，実際に本歌取りになっているかを調査することにした．実験に用いる歌集としては，まず，これまで和歌文学研究者によって盛んに研究され，本歌取り等についての研究の蓄積のあるものとして，『新古今集』を取り上げ，これを『古今集』と比較する．

『古今集』1,111首と『新古今集』2,005首の間の220万を超える組合せの各々について，三つの指標による類似度の値を計算した．類似度の度数分布を表3に示す．

まず，指標Cによる順位と，指標Aによる順位を比較した．表に示したように，指標Aによる類似度の値が19以上のものは13対あるが，この13対については，順位の変動はせいぜい14程度であり，大きな変動は見られなかった．しかし，指標Aによる類似度が17, 18のものについては，大きいもので数十から数百の順位の変動が見られた．順位が大きく下降したものは，共通パターン中の文字が非連続であったために類似度が下がったものである．このうち，20位以上順位が下降したものについて実際に和歌を読んで判断したところ，指標Aの値が高い割にはあまり「似ていない」ものが多かった．すなわち，この順位の下降は妥当なものであったといえる．このことは，指標Cが指標Aに比べ有効であることを示すものである．指標Bについても同様の作業を行い，同様の結果を得た．

次に，指標Cの値が高いものについて，岩波新日本古典文学大系『新古今和歌集』[6]の脚注における本歌取り等の情報と，人手による突合せを行った．すると，指標Cの値が13以上である15対に関しては，そのうちの13対までが，対応する古今歌を本歌として挙げてあった．す

表 3: 類似度の分布

指標 A

| 類似度 | 対の数 | 累積 |
|---|---|---|
| 23 | 1 | 1 |
| 22 | 0 | 1 |
| 21 | 3 | 4 |
| 20 | 3 | 7 |
| 19 | 6 | 13 |
| 18 | 25 | 38 |
| 17 | 51 | 89 |
| 16 | 108 | 197 |
| 15 | 271 | 468 |
| 14 | 878 | 1346 |
| 13 | 3239 | 4585 |
| 12 | 12643 | 17228 |
| 11 | 49183 | 66411 |
| 10 | 160296 | 226707 |
| 9 | 390872 | 617579 |
| 8 | 631971 | 1249550 |
| 7 | 592016 | 1841566 |
| 6 | 292200 | 2133766 |
| 5 | 68372 | 2202138 |
| 4 | 7037 | 2209175 |
| 3 | 330 | 2209505 |
| 2 | 5 | 2209510 |

指標 B

| 類似度 | 対の数 | 累積 |
|---|---|---|
| 18∼19 | 1 | 1 |
| 17∼18 | 1 | 2 |
| 16∼17 | 1 | 3 |
| 15∼16 | 1 | 4 |
| 14∼15 | 4 | 8 |
| 13∼14 | 10 | 18 |
| 12∼13 | 19 | 37 |
| 11∼12 | 18 | 55 |
| 10∼11 | 43 | 98 |
| 9∼10 | 47 | 145 |
| 8∼ 9 | 95 | 240 |
| 7∼ 8 | 206 | 446 |
| 6∼ 7 | 483 | 929 |
| 5∼ 6 | 1007 | 1936 |
| 4∼ 5 | 2408 | 4344 |
| 3∼ 4 | 6191 | 10535 |
| 2∼ 3 | 17957 | 28492 |
| 1∼ 2 | 56998 | 85490 |
| 0∼ 1 | 168277 | 253767 |
| -1∼ 0 | 391833 | 645600 |
| -2∼ -1 | 572532 | 1218132 |
| -3∼ -2 | 386292 | 1604424 |
| -4∼ -3 | 80124 | 1684548 |
| -5∼ -4 | 138 | 1684686 |

指標 C

| 類似度 | 対の数 | 累積 |
|---|---|---|
| 16∼17 | 2 | 2 |
| 15∼16 | 1 | 3 |
| 14∼15 | 4 | 7 |
| 13∼14 | 8 | 15 |
| 12∼13 | 26 | 41 |
| 11∼12 | 30 | 71 |
| 10∼11 | 78 | 149 |
| 9∼10 | 133 | 282 |
| 8∼ 9 | 330 | 612 |
| 7∼ 8 | 1038 | 1650 |
| 6∼ 7 | 3145 | 4795 |
| 5∼ 6 | 9946 | 14741 |
| 4∼ 5 | 34946 | 49687 |
| 3∼ 4 | 132667 | 182354 |
| 2∼ 3 | 430687 | 613041 |
| 1∼ 2 | 873750 | 1486791 |
| 0∼ 1 | 722719 | 2209510 |

なわち，このことは，類似度の高いもののほとんどが，実際に本歌取りの指摘になっていることを示すものである．一方，本歌として挙がっていない2対は，本歌取りというよりはむしろ，「春霞 たなびく山の」や「*のみどりぞ いろまさりける」という慣用表現が共通しているために類似度が高くなったものであった．このようなものの類似度を下げるためには，句や表現の生起確率を考慮したモデル化を行う必要があろう．

類似度が下がるにつれ，脚注に指摘がないものが多くなるようであるが，その中には，本歌取りと考えてしかるべきものも含まれていた．例えば，以下の1対である．

**例 3**

あふ事を ながらのはしの ながらへて
こひ渡るまに 年ぞへにける (古今集#826)

ながらへて 猶君が代を 松山の
まつとせしまに 年ぞへにける (新古今集#1636)

この2首の指標 A, B, C による類似度は，それぞれ，16, 13.3, 11.5 であり，その順位は，表から，それぞれ，197, 18, 71 位以内ということになる．上述の岩波新日本古典文学大系の脚注には，別の古今歌，すなわち，

かくしつつ とにもかくにも 永らへて
君が八千代に 逢ふよしもがな (古今集#347)

を本歌として挙げてあるが，先に示した826番の古今歌も，併せて本歌とすべきものと考えられる．347番の古今歌との類似度は，指標 A, B, C に対して 13, 4.9, 6.7 と，いずれも低い値であった．ちなみに，新潮日本古典集成『新古今和歌集』[4] にも，826番の古今歌の指摘はない．注釈書の中でも，最新かつ，もっとも普及しているこれらの本に指摘のない歌を拾えたということは，和歌文学研究において，本方式が有効であることを示しているといえよう．

現在のところ，注釈書に挙げてある本歌取りその他の情報に関する機械可読データがないので，十分な評価ができないでいる．そこで，上述の岩波新日本古典文学大系と新潮日本古典集成の2冊を用いて，その注から本歌や類歌などの情報を取り出し，入力する作業を進めている．この作業が完了すれば，古くから蓄積されてきた研究成果との突合せにより，本方式の有効性の十分な検証を行うことができよう．さらに，この方式を，『古今集』や『新古今集』だけではなく，これまで研究がほとんど行われていな

かった歌集にも適用することにより，注釈作業を円滑に行うことはもとより，個々の歌人の作風の特色や，時代の好尚を探る端緒となりうるのである．

## 5 むすび

本論文では，和歌の文字列としての類似性に着目した類似歌の自動抽出を目指し，そのために必要な類似性の指標を提案し，その有効性を示した．

しかし，この類似度にも以下のような問題点があるようである．例えば，二つの句「たてるやいづこ」「たてるはみやこ」は，直感的に，非常に良く似て感じられる．この二句は，「たてる*こ」という共通パターンにおいて「たてる」と「こ」が離れているため，提案した指標Cでの類似度は低くなってしまう．それでは，これらの二つの句は，なぜ似ていると感じられるのであろうか．その理由として，まず，同じ7拍であり，「たてる」「こ」が二つの句で同じ位置に表れていることが挙げられよう．さらにいえば，*に合致する「やいづ」と「はみや」についてみると，「や」と「は」，および「い」と「み」は，どちらも母音が同じである．同様な二句の対応として，「ゆきはふりつつ」と「ゆきやふるらん」，「きみやこむ」と「きみはこず」などが挙げられる．このような現象を捉えるためには，文字列を母音・子音といった単音にまで分割し，音から見た句の成り立ちをモデル化することによって，新しい類似度を定義する必要があろう．なお，これらの例は，いずれも本論文で扱った『拾玉集』から得たものであるが，このような現象は，『古今集』と『新古今集』の間ではあまり見られなかったことも，興味深い事実である．

もう一つの問題は，5拍の句と7拍の句が対応している場合である．例えば，「ひとしれぬ」「ひとにしられぬ」は，共通パターン「ひと*し*れぬ」において文字列が非連続であるため指標Cでは類似度が下がるが，これを類似性の高いものとして扱いたい．このような例としては，他にも，「いはにさく」「いはほにもさく」，「まきのとを」「まきのいたとも」，「なかりせば」「なきよなりせば」などがあり，いずれも，意味の改変は最小限にとどめつつ，拍数を増減させたかのように見えるのである．このような2句を類似したものとして扱うためには，両者に，同音の拍がどのように配置されているか，という規則性を見出し，それに基づいて類似度を定義する必要がある．

従来の和歌文学研究は，専ら勅撰集，特に，『古今集』と『新古今集』とを中心に行われてきた．本論文で提案した手法を用いることにより，比較的手薄であった私家集の分野にも広く光を当てることができる．そうして得られた知見を比較検討することを通じて，新たな，統合的視点が得られることになろう．そこから新しい研究領域が切り開かれるものと著者らは信ずる．

## 参考文献


[1] A. Brāzma, E. Ukkonen, and J. Vilo. Discovering unbounded unions of regular pattern languages from positive examples. In *Proc. 7th International Symposium on Algorithms and Computation (ISAAC'96)*, pp. 95–104, 1996.

[2] R.A. Wagner and M.J. Fischer. The string-to-string correction problem. *Journal of the ACM*, Vol. 21, No. 1, pp. 168–173, 1974.

[3] M. Yamasaki, M. Takeda, T. Fukuda, and I. Nanri. Discovering characteristic patterns from collection of classical Japanese poems. In *Proc. 1st International Conference on Discovery Science (DS'98)*, 1998. (to appear).

[4] 久保田 淳 (校注). 新日本古典集成『新古今和歌集』. 新潮社, 1979.

[5] 福田 智子. 藤原兼家六十賀和歌をめぐって—正保版歌仙家集本『兼盛集』と西本願寺本『能宣集』—. 国語国文, Vol. 62, No. 12, pp. 1–14, 1993.

[6] 田中 裕, 赤瀬 信吾 (校注). 新日本古典文学大系 *11*『新古今和歌集』. 岩波書店, 1992.