統計数理 (2000)
第 48 巻 第 2 号 289–310

特集「“ ことば ”新研究」
［原著論文］

和歌データからの類似歌発見

九州大学* 竹田 正幸
福岡女学院大学** 福田 智子
純真女子短期大学*** 南里 一郎
九州大学* 山崎 真由美・玉利 公一

(受付 2000 年 5 月 2 日；改訂 2000 年 8 月 21 日)

要 旨

大量の古典和歌の集積から類似歌を抽出するための方法として，和歌間の類似性指標を定義し，その指標の値の大きい和歌の対を人手により検証する，といった方式が考えられる．このような方式においては，成功の鍵は，いかに類似性指標を定義するかにかかっている．しかし，多様な類似性を考慮すれば，有効な類似性指標が唯一つに定まるとは考えにくい．むしろ，研究者の視点に応じて指標を自由に変更し，その都度，類似度の値の高い対を確認していく，というシナリオが有効であろう．

本稿では，まず，類似性指標を自由に設計するための共通の土俵となる統一的枠組みを導入する．この枠組みでは，指標を，パターン集合とパターンにスコアを与える関数との対によって表し，二つの文字列間の類似度を，その共通パターンの最大スコアとして定義する．文字列間の類似性が共通パターンの形で陽に与えられるため，類似性を直感的に捉えやすい．次に，この枠組みのもとで，本歌取りの半自動抽出に適した三つの類似性指標を設計し，これを用いて『古今集』と『新古今集』の間の 200 万余りの組合せについて類似度を算出した．その結果，(1) 類似度の高い対の多くは，実際に本歌取りであること，(2) これまでに指摘のなかった本歌取りの例を，類似度の高いものとして拾うことができること，(3) 本歌取り以外にも，ある特定の詠歌状況下で用いられる表現や，伝来の過程で表現のバリエーションが生じた異伝歌，掛詞などの表現技巧が共通する歌などが抽出できること，が判明した．特に，共通パターンの生起頻度を考慮した指標では，既知の常套表現をできる限り排除した，より緊密な類似性をもつ歌の対を得ることができた．

キーワード：古典和歌，表現分析，類似性指標，類似歌，機械発見．

1. まえがき

平成 8 年，約 45 万首の古典和歌を収めた『新編国歌大観』CD-ROM 版が，角川書店より刊行された．本稿では，このような大量の和歌の集積から，類似歌を半自動的に抽出する手法について論じる．

*大学院システム情報科学研究科：〒812–8581 福岡市東区箱崎 6–10–1.
**人文学部：〒838–0141 福岡県小郡市小郡 2409–1.
***国文科：〒815–8510 福岡市南区筑紫丘 1–1–1.

一口に「類似歌」といっても，和歌の類似の仕方はさまざまである．もし，和歌に現れる語句の意味内容にまで踏みこんで類似性を扱うとすれば，自然言語理解の技法を適用することが考えられる．しかし，対象である古典和歌は，

- 現代語ではなく古代語であり，
- 日常文ではなく文学作品であり，しかも，
- 散文ではなく韻文 (詩歌) である．

したがって，古典和歌を相手に自然言語理解を行うためには，膨大な量の知識を集積し，組織化する必要がある．比較的浅いレベルの「理解」を目指す場合ですら，このような知識の集積・組織化は困難をきわめる．また，その労力の割に，和歌文学の最先端の研究に果たしてどれだけ寄与できるものか，はなはだ疑問である．

そこで本稿では，このような意味的処理を一切行わず，また，単語という概念すら捨ててしまって，和歌を単なる仮名文字の連鎖とみなし，和歌間の共通部分文字列に着目して類似性を考える．このような観点で類似した和歌は，本歌取り，すなわち，特定の歌を踏まえて新しい歌を作る手法によるものであることが少なくない．また，本歌取りではなくとも，先行歌と同様の発想で詠まれた類想歌や，一首の歌が伝来の過程で本文の微妙な違いを生じた異伝歌であることもある．したがって，このような類似歌を見出す有効な方法が得られれば，和歌文学研究への大きな寄与が期待できる．

類似歌の抽出法として，和歌間の類似性指標を定義し，その指標の値の大きい和歌の対を人手により検証する，といった方式が考えられる．このような方式においては，成功の鍵は，類似性指標をいかに定義するかにかかっている．そのための一つの方法として，類似歌対と非類似歌対のデータを集め，例からの学習 (learning by examples; e.g. Laird (1988)) の手法を用いて類似性指標を学習する方法が考えられる．

しかし，本研究の目的は，和歌文学研究における「発見」の支援であり，そのためには，人手による従来研究を計算機になぞらせることではなく，まったく別の視点と手法を用いてこれまで看過されていた事実を見出すことこそが重要と考えている．すなわち，これまでに指摘された本歌取りの事例に基づいて同様の本歌取りの事例を指摘することよりも，人手では指摘しにくいタイプの本歌取りや類似歌を計算機ならではの手法によって指摘することを目指している．したがって，その意味では，学習に必要な訓練例は得られないものと考えざるを得ない．(実際，次章で述べるように，これまで指摘されてきた本歌取りの事例は，専ら自立語を中心としたものであり，言い回しに着目した研究は，あまり行われていない.) そこで，類似性指標は，機械学習によらず人手で設計することにする．

類似歌抽出に有効な類似性指標が唯一存在するとは考えられない．むしろ，研究者の視点に応じて指標を自由に変更し，その都度，類似度の値の高い対を確認していく，というシナリオに沿った研究が有効であろう．そして，そのような指標の設計と変更は，場当たり的に行うのではなく，ある共通の土台の上で，見通しよく行うべきである．

本稿では，まず，類似性指標のための統一的枠組みを導入する．この枠組みでは，指標を，パターン集合とパターンにスコアを与える関数との対によって表す．そして，二つの文字列間の類似度を，その共通パターンの最大スコアとして定義する．この枠組みは，

1. 代表的な非類似性指標である編集距離及びその変種をすべて表現でき，かつ，
2. 類似性が共通パターンとして陽に与えられるため指標を直感的に把握しやすい，

という利点をもつ．

次に，この枠組みのもとで，類似歌の半自動抽出に適した類似性指標を三つ提案する．第 1 の指標は，和歌を 5–7–5–7–7 の五句に分割し，句ごとに求めた類似度の総和を和歌間の類似度とするものである．句間の類似度は，上述の枠組みにおいて，パターン集合を正規パターン (regular pattern; Shinohara (1982)) の集合とし，各パターンのスコアを，パターン中の文字列の長さと個数に依存して定めるものである．また，第 2 の指標は，句に分割せずに，歌全体での共通部分文字列を求める．パターン集合としては，順序自由パターン (order-free pattern) の集合を用い，スコアはパターン中の文字列の長さに依存して与える．さらに，第 3 の指標は，パターン集合は第 2 の指標と同じであるが，パターンのスコアをパターンの生起頻度に依存して与えるものであり，稀少度が高いパターンを共通してもつ対ほど類似度は高くなる．

これら三つの指標を用いて,『古今集』と『新古今集』など，二つの和歌集の間のすべての対について類似度を算出し，類似歌の抽出を試みた．その結果,

- 類似度の高い対の多くは本歌取りであること．
- これまで指摘のなかった本歌取りの例を類似度の高いものとして拾うことができること．

が判明した．また，本歌取り以外にも，ある特定の詠歌状況下で用いられる表現や，伝来の過程で表現のバリエーションが生じた異伝歌，掛詞などの表現技巧が共通する歌などが抽出できた．特に，第 3 の指標を用いた場合には，その他の指標では類似度が下位になっていた歌の対が上位に浮上し，既知の常套表現をできる限り排除した，より緊密な類似性をもつ歌の対を得ることができた．

本研究の最終目標は，古典和歌における表現技法の系譜を明らかにすることである．本手法により，今まで見過ごされてきた表現の影響関係をいくつか見出すことができた．たとえば，親心の率直な吐露とのみ評価されてきた藤原兼輔の歌(『後撰集』1102 番)が，清原深養父(『古今集』585 番)の骨組みを利用した，いわば「替え歌」であることを発見した (福田 (2000a))．これにより，古歌を踏まえた歌作りの一面が明らかになった．また,『為忠集』の成立年代について，これまで鎌倉中期頃かといわれてきたが，表現の授受関係から，実は室町時代であることを実証した (福田 (2000b))．これは，表現研究が歌集の成立年代推定にまで発展した例である．

なお，実験には,『新編国歌大観』CD-ROM 版の句索引のデータファイルからもとの和歌を復元し，利用した．これにより，すべて清音表記された仮名文字列のデータを得ることができる．

2. 研究の背景

古典和歌における表現の類似性の分析は，これまで，もっぱら名詞や動詞を中心とする自立語に着目して行われてきた．これらの語は，表現の素材となり,「梅に鶯」「紅葉と鹿」のように，特定の組合せで用いられる．本歌取りという作歌手法を考察する際にも，自立語を中心とするきらいがある．特定の歌をもとにして新たな歌を作るという，この作歌法は，いわば「替え歌」作りである．もとの歌と「替え歌」とに共通する自立語は，比較的指摘しやすい．自立語は，主題や情景等と直接結びつくために，記憶に残りやすいのである．

しかし，自立語に偏した従来の研究は，片手落ちの誹りを免れない．なぜならば，自立語と自立語を連繋させ一首の和歌にまとめあげるという重要な役割を担う付属語 (助詞・助動詞) が，ここではまったく度外視されているからである．本歌取りにおいても，先行歌のどの部分を踏まえるかはさまざまである．したがって，自立語とともに，付属語の共通性をも視

野に入れることで，より作歌の実際に近づくことができると考えられる．また，5–7–5–7–7 という，音数律に制約のある和歌においては，共通する音を把握することにも，また大きな比重をおいてよい．一例を挙げよう．

人のおやの/心はやみに/あらねども/子を思ふ道に/まどひぬるかな

『後撰集』1102 番

この歌は，三十六歌仙のひとり，藤原兼輔 (877–933) の代表作で，子を思う親の心情をストレートに表現した，ほとんど無技巧な歌である，という共通理解を得てきた．ところがこの歌，実は，次の先行歌を踏まえて作られたものと見られるのである．

人を思ふ/心はかりに/あらねども/くもゐにのみも/なきわたるかな

『古今集』585 番

この『古今集』歌と先の兼輔歌とを比べてみると，

ひと ··· /こころは ··· に/あらねども/··· /··· るかな

という一首の輪郭が共通する．そしてさらに，第二句の「やみ」(兼輔)と「かり」(古今集)は，ともに [a][i] という，母音が共通する語である．このように，兼輔歌は，共通する自立語こそ多くはないものの,『古今集』585 番歌と，きわめて高い類似性を示している．

すると，先の兼輔歌は，単に無技巧といって片づけられないことになる．彼の念頭には，上の『古今集』の恋歌があった．兼輔は，そこに詠まれた恋人への一途な愛情を，我が子に向けて「替え歌」に仕立てたのであろう．このような有名な歌でも，付属語や音の共通性を考慮することで，これまで忘れ去られていた一面が発見されることもある．

このような付属語重視の発想に基づいて，著者らは，付属語や用言の活用語尾などの作るパターンである**ふし** (節) を表現技法を特徴づけるモデルとして提案した (竹田 他 (1999), Yamasaki et al. (2000))．たとえば,「∗せば∗ざらましを∗」は反実仮想という表現技法に対応する．竹田 他 (1999), Yamasaki et al. (2000) では，最小記述長原理 (Rissanen (1978)) に基づいたパターン抽出法 (Brāzma et al. (1996)) を用いて，歌集からのふしの自動抽出を試みた．得られたふし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おり，研究者に非常に興味深い視点を与えるものであった．

本稿においては，類似歌の半自動抽出の問題を扱うが，その際，和歌を単なる仮名文字の連鎖とみなし，共通する部分文字列に着目して類似度を定義する．すなわち，自立語・付属語の区別なく類似性を考慮するのである．先に述べたように，本歌取りに関しても，これまではもっぱら自立語を中心に考察されており，本手法はそれを補うものといえよう．

実際，上で兼輔の歌の本歌として掲げた『古今集』585 番歌は，この手法を用いて『古今集』と『後撰集』を比較した際に見出したものである．

3. 方法

3.1 文字列としての和歌

文学作品を計算機で処理する際には，通常，テキストに何らかの統語情報を埋め込んだいわゆるタグ付きコーパス (tagged corpus) を用いることが多い．たとえば，村上・今西 (1999) は，8 年もの歳月を費やして源氏物語の本文を単語に分割し，各々に品詞等のタグを付与する作業を行い，作成したコーパスをもとに助動詞の計量分析を試みている．和歌の類似表現を計算機で抽出する場合にも，このようなタグ付きコーパスを用いることが考

えられる．もし,『新編国歌大観』に収められた 45 万首すべてを正確に読み解き，単語に分割する基準を一貫させてコーパスを作成することができるならば，これほどすばらしいことはない．だが現実には，1 語で二つの意味を担う掛詞や複合語の扱いをどうするかなど，問題は数多い．一首の歌の解釈に行き詰まることすらある．たった一つの歌集に対する作業さえ容易でないことは，著者らを含め，実際に試みたことのある者が共通してもつ実感であろう．

一方，計算機によって文の単語分割を行い，品詞等の統語情報を付与するいわゆる形態素解析の研究は，自然言語処理の分野で古くから行われている．しかし，解析の過程で大量の曖昧さ (ambiguity) が発生するため，単語の分割や付与すべき品詞を一意に決定できないことが知られている．この曖昧さを除くために，意味的制約を用いる方法が提案されているが，この制約の作成は容易ではなく，また，意味に関する曖昧さが新たに発生するという問題がある．そこで，現在では，統計情報を用いて曖昧さの除去を行う方法が主流となっている．ところが，この統計情報は，上述のタグ付きコーパスより得ることが一般的であるから，いずれにしても，人手によってある程度の量のコーパスを作成する作業は免れない．

それでは，比較的少量のタグ付きコーパスを作成し，それから得た情報をもとに精度の高い形態素解析器を開発し，それによって，大量の古典文学作品に形態素解析を施す，というシナリオを検討してみよう．このようなシナリオは，あるコーパスにおいて成立した何らかの統計的性質が，別の文の集合においても成立すること，すなわち，対象の均質性を前提としている．ところが，古典和歌における用語の選択はたいへん保守的であるとはいえ，作品の成立は幅広い時代にわたっており，言語的には等質でないと考えなければならない．したがって，統計的手法を用いるために十分な量のコーパスを得ようとすると，極端にいえば，ほとんどすべての作品にタグ付け作業を行うことになるかもしれない．このように，完全に近いコーパスを得ようとするなら，形態素解析器による労力の軽減は，ほとんど望めない．(一方，現代文を対象とする場合には事情は異なる．すなわち，機械可読な大量の文書が溢れている状況においては，ある程度の精度の低さに目をつぶってでも自然言語処理技術によって大量データを捌くことの意義は決して小さくはない.)

形態素解析器によって単語分割や品詞決定を行ってタグ付きコーパスを作成する場合，どれだけ改良を重ねたとしても，その精度には限界がある．そのようなコーパスをもとに類似性やパターンの抽出の研究を進めた場合，その抽出結果が解析器の使用しているアルゴリズムの性格や辞書・文法規則等に影響を受けるため，問題の所在が不明瞭となる懼れがある．こう考えると，形態素解析などは施さず，単語という概念さえも捨ててしまって，和歌を単なる文字の連鎖とみなして処理を行う方が，処理の透明性は保証されるといえる．

以上の理由により，著者らは，一切の自然言語処理を放棄し，また，人手によるタグ付け作業なども極力避け，和歌を単なる文字の連鎖と扱う立場で研究を進めている．同様の立場をとるものに，近藤 (1999, 2000) がある．これは，単語や単語列の代わりに n グラムを用いてその統計をとり，詠作者の性差による語彙や表現の特徴を抽出しようとするものである．近藤 (1999) は，n グラムを用いたアプローチには，以下のような利点があることを指摘している．

- タグ付け作業を必要としないこと．
- 単語に分割してしまうと見落としがちな，掛詞に関する分析が見落としなく行えること．
- 文字表記の統一が容易であること．
- n の値を大きくとれば，長い文字列の慣用的表現や類句の抽出が可能なこと．

しかし，これらの利点は，n グラムを用いることの利点というよりは，和歌を文字列とみなしたアプローチの利点というべきである．和歌を文字列とみなした立場で取りうる方法は，何も n グラム統計ばかりではない．文字列は最も基本的なデータ構造であって，形式言語理論・オートマトン理論，文字列パターン照合，テキストデータ圧縮，計算論的学習理論など，理論計算機科学の幅広い分野にわたって，文字列に関するさまざまな処理方式やその効率的実現について古くから盛んに研究されており，その一連の成果は，文字列学 (stringology) とでもよぶべきものとして結実している．その研究の蓄積の上に立てば，n グラム統計以上の，より有効な処理方式を見出すことが可能である．本稿では，その一端を示すことになる．

3.2 機械学習手法の適用の検討

類似歌の抽出の問題に，例からの学習 (learning by examples; e.g. Laird (1988)) の手法を適用することを考えよう．すなわち，類似歌である歌の対 (正例) とそうでない歌の対 (負例) が与えられ，これらの訓練例から正例・負例を判別するための規則を学習し，その規則を用いて訓練例にない和歌対に対し正か負かを判定する．このような方式が成功するためには，次の二つが必要である．

1. 類似歌か否かを判定するための規則が，設定された仮説空間の中に存在していること．
2. 訓練例を含めたデータ全体が均質であり，訓練例で成立する規則が，そのまま全体に敷延できること．

1 の規則の自然な表現形式として，和歌間の類似性指標と類似度の値に関する閾値によって正・負を判定する方式が考えられよう．すなわち，類似性指標によって和歌対の正例らしさを数量化し，その値が閾値を超えるか否かによって正・負を判定するのである．この場合，類似性指標と閾値とを訓練例から学習することになるが，そのためには，類似性指標の属すクラスを仮説空間としてあらかじめ設定しなければならない．文字列の類似性を扱う際によく用いられる指標として，重み付き編集距離 (weighted edit distance) がある．仮にこのクラスを仮説空間とするならば，指標の学習の問題は，各編集操作に関わる重みを学習する問題となる．しかし，たとえば置換操作に関わる重みの値は，文字と文字の対の数だけ定める必要があるなど，学習すべきパラメータの数は少なくない．よって，大量の訓練例が必要となる．また，元来，この種の指標は，タイプミスの半自動修正などを目的としたものであり，和歌の場合にそのまま有効であるとは考えにくい．

一方，2 の均質性は，自然現象に関する科学的観測データにおいては，ある程度保証されよう．しかし，和歌の類似の仕方はさまざまであるから，例をできるだけ多く，しかも偏りなく得ることが望ましい．ところが，訓練例の作成は人手を要するため，あまり大きなサイズの訓練例は作成できない．たとえば，1000 首程度の二つの歌集の間において，すべての組合せは 100 万対にもなるため，その各々について正・負の判別を人手によって行うことは不可能に近い．実際，のちにふれる『新古今集』の注釈書には，いくつかの歌について，先行歌を添えて本歌取りであることを指摘してあるが，それが本歌取りのすべてを尽くしてはいない．また，重要なものは網羅している，という訳でもない．

このように，類似歌の抽出の問題に例からの学習の手法を適用しようとする際，量と質の両面から十分な訓練例を得ることができない．不十分な量の訓練例しかない場合，それに依存して学習を行うことは非常に危険である．このため，訓練例に依存しない方式をとらざるを得ない．

また，先に述べたように，人手によって指摘されてきた本歌取りは自立語に偏したもの

になっている．したがって，それを忠実に学習することよりも，それとは別の観点から類似した歌の対を抽出することが重要である．

4. 類似性指標の統一的枠組み

前章で述べたように，著者らは，和歌を単なる仮名文字の連鎖とみなし，和歌の類似性を文字列の類似性として扱うことにする．また，類似性指標については，訓練例から学習するというスタイルをとることができないため，専門的知識を取り込みつつ，有効な指標を人手によって設計するという方法をとらざるを得ない．そこで，このような指標の設計を，場当たり的でなく，見通しよく行うための土台として，類似性指標のための統一的枠組みを導入する．

この章では，はじめに，既存の類似性 (非類似性) 指標を概観し，次に，それらの指標を統一的に扱う枠組み (Tamari et al. (1999)) を導入する．この枠組みは，以下の特長をもつ．

- 代表的な非類似性指標である編集距離およびその変種をすべて表現できること．
- 文字列間の類似性が共通パターンとして陽に与えられるため，指標を直感的に把握しやすいこと．

この枠組みのもとで，和歌のみならず，さまざまな応用場面において，問題に適した指標を見通しよく設計できる．たとえば，MIDI データなどを対象に主旋律の類似性を扱う場合にも有効である (門田 他 (2000))．

4.1 既存の指標

類似性と非類似性とは，双対的な概念である．文字列間の類似性に関しては，類似性指標よりも，むしろ，非類似性指標がよく知られている．多くの非類似性指標は距離の公理を満たしており，理論的観点からは扱いやすい．しかし，実用的観点から有用な指標が，必ずしも距離の公理を満たしているとは限らない．三角不等式はおろか，対称律すら満たさないこともしばしばである．この節では，代表的な非類似性指標である編集距離 (edit distance)，およびその変種について概観する．

編集距離は，一方の文字列を他方に変換するために必要な編集操作の回数の最小値として定義される．編集操作としては，通常以下の三つが用いられる．

- 文字の挿入 (insertion).
- 文字の削除 (deletion).
- 文字の置換 (substitution).

たとえば，文字列 `acdeba` と `abdac` の間の編集距離は 4 となる．図 1(a) にその様子を示す．図において，縦棒で結ばれた上下の文字対は，文字が同一であることに対応する．縦棒なしで向き合った文字対は置換操作に対応する．また，空白記号とそれに向き合った文字との対は，挿入操作もしくは削除操作に対応する．図から，文字列 `acdeba` に 4 回の編集操作を適用すれば文字列 `abdac` に変換できることがわかる．図 1(a) のような文字列間の対応づけをアラインメント (alignment) と呼ぶ．

上の編集距離の変種として，置換操作を禁じたものがある．この場合，文字の置換は，削除と挿入の 2 回の操作で実現される．図 1(b) にこの場合のアラインメントを示す．図 1(a) で上下に向き合って並んでいた文字 `c`, `b` が，(b) においては，各々，空白記号と向き合っていることに注意しよう．この場合，文字列の変換に必要な編集操作は 5 回であり，距離は 5

```
acdeba      ac deba     acdeb a
| |  |      |  |  |     |   | |
abd  ac     a bd  ac    a   bdac
```

(a) (b) (c)

図 1. アラインメント.

である．距離 5 を与えるアラインメントとしては，この他にも，図 1(c) がある．この場合には，縦棒で結ばれた文字のなす文字列は，aba であり，(a) や (b) の場合のそれが，ada と異なっていることにも注意しよう．文字列 aba と ada は，いずれも，文字列 acdeba と abdac の共通部分列 (common subsequence) であり，しかも長さが最大であることから，最長共通部分列 (longest common subsequence; 以下 LCS と略記する) になっていることがわかる．実際，LCS の長さと置換操作を禁じた場合の編集距離との間には，密接な関係があることが知られている．すなわち，二つの文字列間の距離は，文字列長の和から LCS 長の 2 倍を引いた値に等しい．この性質は，置換操作を許した場合には成立しない．ここでは，置換操作を禁じたが，逆に挿入・削除の操作を禁じて置換操作のみを許した場合，ハミング距離 (Hamming distance) が得られる．この場合，長さの同じ文字列間にのみ距離が定義される．

文書タイピングにおける綴り誤りの訂正や，ゲノム情報学における DNA の塩基配列やアミノ酸配列の比較 (e.g. Gusfield (1997)) などの応用においては，もう少し複雑な指標が用いられる．上で述べた編集距離の定義では，編集操作の回数のみを問題にした．この拡張として，各編集操作にコストを与え，適用した編集操作のコストの総和を最小にすることを考える．コストは，操作に関与する文字に依存して決まる．すなわち，文字 a を文字 b に置き換えるコスト $\delta(a,b)$，文字 a の削除に関わるコスト $\delta(a,\varepsilon)$，および文字 a の挿入に関わるコスト $\delta(\varepsilon,a)$ を定めておき，二つの文字列間の距離を，一方を他方へ変換する際の最小コスト和として定義する．この指標を，重み付き編集距離 (weighted edit distance) と呼ぶ．なお，編集操作に関わるコスト δ は，距離の公理を満たすように定める．

また，DNA 配列の比較等においては，編集操作として，連続した文字列の一括挿入や一括削除などの操作をも考慮する．たとえば，図 1(c) において，文字列 cde を一括して削除するような操作である．このような操作に関わるコストは，ギャップペナルティ (gap penalty) とよばれ，その値は，ギャップの長さの関数として与えられることが多い．通常，このギャップ関数として，一次関数 (affine function) や凸関数 (convex function) が用いられる．

4.2 統一的枠組み

現実の応用場面において，扱う問題に適した類似性指標を設計するためには，類似性自体を直感的に把握できることが重要である．4.1 節で示した編集距離は，一連の編集操作の適用によって文字列を変換する際の最小コストとして定義されており，定義自体は数学的に明確であるものの，類似構造の把握には，図 1 のようなアラインメントを表す図による助けが必要であった．

この節で提案する枠組みでは，二つの文字列間の共通構造を共通パターンとして捉え，類似度を共通パターンの最大スコアと定義する．そこで，類似性指標の違いは，

- 共通パターンの属するパターン集合 Π.
- Π の各パターンにスコアを与えるパターンスコア関数 Φ.

の二つということになる．たとえば，パターン集合として正規パターン (アルファベット Σ の文字と*から成る記号列で，*は Σ 上の任意の文字列と合致する) 全体の集合を用い，パターン中の文字の個数をそのパターンのスコアとすると，LCS の長さに基づく類似性指標が得られる．実際，文字列 acdeba と abdac は共通パターン a*d*a*を含むが，これは三つの文字を含んでいる．

Σ をアルファベットとし，Σ 上の文字列全体の集合を Σ^* で表す．空文字列を ε で表し，$\Sigma^+ = \Sigma^* - \{\varepsilon\}$ とする．**パターン**とは Σ 上の言語の‘表現’(description) をいい，各パターン π に対して，π の表す言語 $L(\pi)$ が一意に定まるものとする．パターン π が文字列 $w \in \Sigma^*$ に**合致する**とは，$w \in L(\pi)$ であるときをいう．パターン π が二つの文字列 $x, y \in \Sigma^*$ の**共通パターン**であるとは，π が両方に合致するとき，すなわち，$x, y \in L(\pi)$ であるときをいう．実数全体の集合を $\boldsymbol{R}$ で表す．以上で，文字列間の類似性指標 (非類似性指標) を定義するための準備が整った．

定義 1. Σ 上の文字列間の**類似性指標**とは，対 $\langle \Pi, \Phi \rangle$ をいう．ここで，Π は**パターン集合**とよばれ，各要素 $\pi \in \Pi$ について Σ 上の言語 $L(\pi)$ が対応する．また，Φ は Π から $\boldsymbol{R}$ への関数で，**パターンスコア関数**とよばれる．

定義 2. 類似性指標 $\langle \Pi, \Phi \rangle$ のもとでの文字列 $x, y \in \Sigma^*$ の類似度 $\mathrm{SIM}_{\langle \Pi, \Phi \rangle}(x, y)$ を，次式で定義する．

$$\mathrm{SIM}_{\langle \Pi, \Phi \rangle}(x, y) = \max\{\Phi(\pi) \,|\, \pi \in \Pi \text{ かつ } x, y \in L(\pi)\}\,.$$

非類似度の場合には，上式において，最大値でなく最小値をとればよい．

上では，パターンを Σ 上の言語を定義する表現としたが，パターンを文字とワイルドカードから成る列に制限しても，実にさまざまな類似性指標を扱うことができる．ここでは，ワイルドカードとして表 1 に示した 4 種類を導入する．

以下では，主な (非) 類似性指標を上記の枠組みに沿って記述する．なお，ここでのパターンスコア関数 $\Phi_i : \Pi_i \to \boldsymbol{R}$ は，いずれも，

$$\Phi_i(uv) = \Phi_i(u) + \Phi_i(v) \qquad (u, v \in \Sigma^*)$$

を満たす準同型写像である．

LCS の長さに基づく類似性指標．
パターン集合： $\Pi_1 = (\Sigma \cup \{*\})^*$.
パターンスコア関数： $\Phi_1(a) = 1$ $(a \in \Sigma)$ と $\Phi_1(*) = 0$ で定まる準同型写像 $\Phi_1 : \Pi_1 \to \boldsymbol{R}$.

ハミング距離．
パターン集合： $\Pi_2 = (\Sigma \cup \{\phi\})^*$.
パターンスコア関数： $\Phi_2(a) = 0$ $(a \in \Sigma)$ と $\Phi_2(\phi) = 1$ で定まる準同型写像 $\Phi_2 : \Pi_2 \to \boldsymbol{R}$.

表 1. ワイルドカード.

$*$	:	Σ 上の任意の文字列と合致するワイルドカード
ϕ	:	Σ の任意の文字と合致するワイルドカード
$[w]$	:	空文字列 ε と文字列 $w \in \Sigma^+$ の両方に合致するワイルドカード
$\phi(u_1 \vert \cdots \vert u_k)$	:	文字列 $u_1, \ldots, u_k \in \Sigma^+$ のすべてに合致するワイルドカード

重み付き編集距離.
パターン集合: $\Pi_3 = (\Sigma \cup \Delta_3)^*$. ここで, $\Delta_3 = \{[a] \,|\, a \in \Sigma\} \cup \{\phi(a|b) \,|\, a, b \in \Sigma \text{ かつ } a \neq b\}$.
パターンスコア関数: $\Phi_3([a]) = \delta(a, \varepsilon)$, $\Phi_3(\phi(a|b)) = \delta(a, b)$, $\Phi_3(a) = 0$ $(a, b \in \Sigma)$ によって定まる準同型写像 $\Phi_3 : \Pi_3 \to \boldsymbol{R}$.

ギャップペナルティ・重み付き編集距離.
パターン集合: $\Pi_4 = (\Sigma \cup \Delta_4)^*$. ここで, $\Delta_4 = \{[w] \,|\, w \in \Sigma^+\} \cup \{\phi(a|b) \,|\, a, b \in \Sigma \text{かつ } a \neq b\}$.
パターンスコア関数: $\Phi_4([w])$ を長さ $|w|$ のギャップに対するペナルティとし, $\Phi_4(\phi(a|b)) = \delta(a, b)$, $\Phi_4(a) = 0$ $(a, b \in \Sigma)$ とした準同型写像 $\Phi_4 : \Pi_4 \to \boldsymbol{R}$.

次の章では, 以上の枠組みに沿って和歌間の類似性指標を定め, その有効性を検証する.

5. 和歌に適した類似性指標

前章で示した枠組みのもとで類似性指標を設計する際, 以下の二つを決定しなければならない.

- どのような形式のパターンによって類似構造 (共通構造) を表すべきか.
- パターンとして表現された各々の類似構造に, どのようにスコアを割り当てるべきか.

これらの決定を問題領域の性質に合わせて行えば, 有効な指標が得られるものと考えられる. この章では, 古典和歌において本歌取りを中心とした類似歌を半自動抽出することを目的とし, そのための類似性指標について論じる. なお, ここで示す指標は, 山崎 他 (1998) で提案したものである.

5.1 句の順序の変化

本歌取りにおいては, 先行歌の表現が少なからず用いられることになる. したがって, 歌人は単なるイミテーションに堕してしまわないように注意を払う必要があった. 藤原定家 (1162–1241) は以下のように記している (久松 (1971)).

- 古歌を取りて新しき歌を詠ずる, 五句の内に三句に及ばば, 頗る過分, 珍し気なし. 二句の上に三字, 四字これを許す. 『詠歌大概』
- 五七五の七五の字をさながら置きて, 七々の字を同じく続けつれば, 新しき歌に聞きなされぬ所ぞ侍る. 『近代秀歌』

第 1 の項目から, 文字列の対応は, 和歌を句に分割して, 句ごとに対応をとればよいと考えられる. また, その際, 句の順序の変化を想定しなければならないことが, 第 2 の項目からわかる. 実際, 次に示す本歌取りの例では,『古今集』147 番歌の初句, 第二句, 第四句に対して,『新古今集』216 番歌の初句, 第四句, 第二句が, それぞれ対応している.

例 1. (本歌取り)
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 番 (よみ人しらず)

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 番 (藤原公経)

5–7–5–7–7 の短歌形式の場合，五つの句から成るため，$5! = 120$ 通りの対応付けを考えなければならない．また，上の例では，3 対とも文字列として全く同一であったが，たとえば「ながなくさとの」に対する「ながなくさとは」のように，若干数の文字が異なることも少なくない．そこで，120 通りの対応付けの各々について，対応付けられた句の間での類似度の総和を求め，これを最大にするような対応付けを考える．その最大値を和歌と和歌の類似度と定義する．すると，次には，句ごとの類似度をどのように定義するかが問題となる．

5.2 句ごとの類似度を与える指標

著者らは，山崎 他 (1998) において，まず，LCS 長に基づく指標を用いて実験を行い，その結果から，共通パターンにおける文字の連続性を考慮すべきであるとの着想に達し，これに基づく指標を提案した．以下の例をみてみよう．

例 2. **(本歌取り)**
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 番 (源宗干)
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 番 (慈円)

「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，それぞれ一致している．しかし，これらはほとんど無意味である．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，意味がある．このような文字の偶然の一致は，形態素解析を行わない限り避けられない問題であるが，文字が連続していれば，ある程度偶然の一致の可能性が低くなると考えられる．このような観点から，文字の連続性を重視したスコア付けを行うことにした．

その指標は以下のようなものである．パターン中の連続文字列の長さに注目する．たとえば，$\pi = *a*bc*d*$ においては，左から，$1, 2, 1$ である．正整数全体の集合から正実数全体の集合への写像 f を仮定し，パターンスコア関数の値を $\Phi(\pi) = f(1) + f(2) + f(1)$ のようにすることを考えよう．ここで特に $f(\ell) = \ell$ $(\ell > 0)$ とすれば，$\Phi(\pi)$ は π 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

$$f(n+m) > f(n) + f(m) \tag{5.1}$$

でなければならない．この条件を満たす f は無数に存在するが，ここでは，$f(\ell)$ を ℓ の 1 次関数に限定し，

$$f(\ell) = \ell - s \qquad (0 < s < 1) \tag{5.2}$$

とおいた．このパラメータ s は $s = 0.9$ としたが，その決定には，本歌取りに関する少量の正例・負例を用いた．すなわち，慈円の『拾玉集』の 3,472 番から 3,571 番までの 100 首とその先行歌である『古今集』歌の 100 対を正例とし，それ以外の組合せである 9,900 対を

負例とした．特定の一人の歌人の歌である点など，訓練例としては必ずしも適当とはいえず，問題は残る．

山崎 他 (1998) の指標．
パターン集合： $\Pi = (\Sigma \cup \{*\})^*$.
パターンスコア関数： 次で定まる写像 $\Phi : \Pi \to \mathcal{R}$

$$\Phi(u_1 * \cdots * u_k) = \sum_{i=1}^{k} f(|u_i|) \qquad (u_1, \ldots, u_k \in \Sigma^*, k \geq 1)\,.$$

ここで，f は (5.2) 式で定めるものとし，便宜上，$f(0) = 0$ とする．

5.3 歌集間の類似歌抽出

『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，類似度の値を計算し，その上位のものについて人手で調査した．本歌取りか否かの判断は，ある程度主観的なものであるため，客観的データを出すために,『新古今集』の代表的な注釈書 2 冊 (久保田 (1979)，田中・赤瀬 (1992)) の脚注を調べ，指摘の有無をチェックした．その結果を表 2 に示した．括弧内の数は，上述の注釈書に本歌・類想歌・参考歌など，表現上，影響関係のある歌の指摘がみられた対の数を示す．

表からわかるように，類似度の値が 11 以上となる 73 対のうち 50 対が注釈書に指摘されていた．残りの 23 対は，一部の例外を除いて，本歌取りではないと考えられるものであった．また，指標の値が 13 以上である 15 対のうちでは，13 対について指摘があった．本歌取りでなかった 2 対を以下に示す．

表 2. 類似度の度数分布 (山崎 他 (1998) の指標; 指標 A)．括弧内の数は，田中・赤瀬 (1992)，久保田 (1979) の注釈書に本歌・類想歌・参考歌など，表現上，影響関係のある歌の指摘がみられた対の数を示す．

類似度	度数	累積度数
16– 17	2(1)	2(1)
15– 16	1(1)	3(2)
14– 15	4(4)	7(6)
13– 14	8(6)	15(12)
12– 13	26(18)	41(30)
11– 12	32(20)	73(50)
10– 11	77( )	150( )
9– 10	137( )	287( )
8– 9	332( )	619( )
7– 8	1066( )	1685( )
6– 7	3160( )	4845( )
5– 6	10089( )	14934( )
4– 5	35407( )	50341( )
3– 4	134145( )	184486( )
2– 3	433573( )	618059( )
1– 2	873904( )	1491963( )
0– 1	717547( )	2209510( )

例 3.　(本歌取りでない対)

すまのあまの/しほやく煙/風をいたみ/おもはぬ方に/たなびきにけり
『古今集』708 番 (よみ人しらず)

しかのあまの/しほやく煙/かぜをいたみ/立ちはのぼらで/山にたなびく
『古今集』1592 番 (よみ人しらず)

例 4.　(本歌取りでない対)

春霞/たなびく山の/さくら花/見れどもあかぬ/君にもあるかな
『古今集』684 番 (紀友則)

紫の/雲にもあらで/春霞/たなびく山の/かひはなにぞも
『新古今集』1448 番 (円融院)

いずれの対も，文字列の共通性からは非常に類似している．前者は，両歌ともに読人しらず歌である．新古今歌は万葉歌を再録したものであり，したがって，古歌の類型表現のバリエーションとも見られよう．一方，後者は，単なる類想歌であると見られる．高い類似度を得たのは,「はるがすみたなびくやまの」という表現を共通にもつためだが，これは和歌によくみられる慣用的表現である．そこで，表現の生起確率を考慮することにより，このような対の類似度の値を小さくする方法が考えられよう．この点に関しては，6.2 節で論じる．

類似度が下がるにつれ，注釈書に指摘がないものが多くなる．その中には，以下に示すように，本歌取りと考えてしかるべきものも含まれていた．

例 5.　(本歌取りとすべき対)

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 番 (坂上是則)

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 番 (二条院讃岐)

この対では,「ながらへて ··· まに年ぞへにける」という表現が一致しており，しかも，歌枕「ながらのはし」(古今)「松山」(新古今) が，それぞれ,「ながらへて」(古今)「まつと ···」を導くという共通性をもっている．田中・赤瀬 (1992) 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 番 (光孝天皇)

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．なお，この対の類似度は 11.5，全体の順位は 55 位であった．

ここで示した指標は，上位 73 対のうち 50 対が本歌取りと指摘されているという点で非常に良い性質をもっているといえる．しかし，指摘のない対の中にも，上で示した例のように，本来指摘されてしかるべき歌が見落とされている場合があるはずである．和歌文学研究者の興味をひくのは，まさにそのような，従来指摘のない歌であり，したがって，指摘のあった割合だけをもとに指標の評価を行うべきではない．

6. 多様な類似性への対応

前章では，共通する文字列に着目した類似性指標を提案し，その指標が，これまで看過されていた本歌取りを高い類似度をもつものとして拾うことができるなどの点で，非常に有効であることを示した．しかし，和歌の類似性は多様であるから,「共通する文字列に着目した類似性」に絞ったとはいっても，有効な指標が一意的に定まるとは考えにくい．むしろ，研究者の視点やその時々の興味に応じて研究者自身が指標を変更し，新たに類似度が高くなった対を調べる，というシナリオで研究を進めていくことが有効だと思われる．この章では，第 4 章で導入した枠組みのもとで別の視点から新たな指標を提示し，その評価を行う．

6.1 順序自由パターンに基づく指標

6.1.1 山崎 他 (1998) の指標の問題点

前章で示した山崎 他 (1998) の類似性指標は，以下のようなものであった．

- 和歌を 5–7–5–7–7 の句に分割し，句ごとに求めた類似度の和を全体の類似度とする．
- 句と句の対応付けとしては，$5! = 120$ 通りのすべてを考慮して，その中で最も類似度の総和が大きくなるものを選ぶ．
- 句と句の類似度は，共通の正規パターンにおいて文字の連続性を考慮して与える．

しかし，この指標では，次のような場合をうまく扱うことができない．

例 6. **(本歌取り)**

君まさで/煙たえにし/しほがまの/浦さびしくも/見え渡るかな
『古今集』852 番 (紀貫之)

ふるゆきに/たくもの煙/かきたえて/さびしくもあるか/しほがまのうら
『新古今集』674 番 (藤原兼実)

この本歌取りの例では，先行歌の第二句に現れていた「煙」「たえ」の語句が，新古今歌では第二句と第三句とに分かれて現れている．また逆に，先行歌では第三句と第四句にまたがっていた「しほがまのうら」が新古今歌ではまとまって結句となっている．前章で示した類似性指標は，句と句の対応のみを扱うために，このような場合をうまく扱うことができない．実際，この対の類似度の値は 10.6 であり全体の順位は 92 位にとどまる．以下に示す和歌対についても同様のことがいえる．

例 7. **(本歌取り)**

かはづなく/ゐでの山吹/ちりにけり/花のさかりに/あはましものを
『古今集』125 番 (よみ人しらず)

あしびきの/山吹の花/ちりにけり/ゐでのかはづは/いまやなくらむ
『新古今集』1162 番 (藤原興風)

前章の指標による類似度は 9.7，順位は 161 位であった．

6.1.2 新しい指標の定義

そこで，句に分割せずに，31 文字全体を比較して，共通して現れる部分文字列に着目する．この際，共通部分文字列の生起順序は問わない．

定義 3. 非空文字列の多重集合 $\{u_1, \ldots, u_k\}$ $(u_1, \ldots, u_k \in \Sigma^+, k > 0)$ を**順序自由パターン** (*order-free pattern*) と呼び，$\pi(u_1, \ldots, u_k)$ のように表す．順序自由パターン $\pi(u_1, \ldots, u_k)$ の表す言語を，

$$L(\pi(u_1, \ldots, u_k)) = \bigcup_{\sigma \in \mathcal{S}_k} L(*u_{\sigma(1)}* \cdots *u_{\sigma(k)}*)$$

と定める．ここで，$\mathcal{S}_k$ は集合 $\{1, \ldots, k\}$ 上の置換全体の集合を表す．

次のような類似性指標を考えよう．パターン集合 Π を，順序自由パターン全体の集合とし，各順序自由パターン $\pi(u_1, \ldots, u_k)$ に対するスコアを，ある関数 f を用いて，$\sum_{i=1}^{k} f(|u_i|)$ と定める．

共通部分文字列である各 u_i について連続性を考慮せず，$f(n) = n$ $(n \geq 1)$ とするならば，すべて 1 文字単位で考えてよいから，各文字について二つの文字列中での生起頻度のうち小さい方の値を求め，それらを合計すれば，求める類似度が得られる．すなわち，文字 c の文字列 x における生起回数を $n_c(x)$ で表すとき，文字列 x, y の類似度は，$\sum_c \min\{n_c(x), n_c(y)\}$ で与えられる．

しかし，一般の f の場合には，文字 c ごとに

$$\frac{\max\{n_c(x), n_c(y)\}!}{|n_c(x) - n_c(y)|!}$$

通りの可能性があり，これを各文字ごとに掛け合わせたすべての場合を検討する必要がある．したがって，最悪時の計算時間は，入力長 n に対し $O(n!)$ となる．しかしながら，短歌形式の場合，入力長 n は 31 文字程度で増加しないため，実際の計算時間が耐えられるものであるなら問題はない．

ここでは，関数 f として

$$f(n) = \begin{cases} 0, & n = 1 \text{ のとき} \\ n, & n > 1 \text{ のとき} \end{cases}$$

を用いた．すなわち，1 文字だけの文字の一致は偶然の一致の可能性が高いと考え，2 文字以上連続したものだけを考慮した．以上をまとめると，次のようになる．

順序自由パターンに基づく指標．
パターン集合： $\Pi = \{\pi \mid \Sigma \text{ 上の順序自由パターン }\}$.
パターンスコア関数： 次で定まる写像 $\Phi : \Pi \to \boldsymbol{R}$

$$\Phi(\pi(u_1, \ldots, u_k)) = \sum_{i=1}^{k} f(|u_i|) \qquad (u_1, \ldots, u_k \in \Sigma^+).$$

ここで，$f(1) = 0$, $f(n) = n$ $(n > 1)$ とする．

6.1.3 結果

順序自由パターンに基づく指標を用いて,『古今集』と『新古今集』の間のすべての対の類似度を算出した．最も類似度の高い対は以下の 2 対であり，類似度の値はいずれも 21 であった.

例 8. (**本歌取り**)

さむしろに/衣かたしき/こよひもや/我をまつらむ/うぢのはしひめ
『古今集』689 番 (よみ人しらず)
はしひめの/かたしき衣/さむしろに/待つよむなしき/宇治の曙
『新古今集』636 番 (後鳥羽院)

例 9. (**本歌取り**)

花のちる/ことやわびしき/春霞/たつたの山の/うぐひすのこゑ
『古今集』108 番 (藤原後蔭)
霞たつ/春の山辺に/さくら花/あかずちるとや/鶯のなく
『新古今集』109 番 (よみ人しらず)

『古今集』689 番歌の結句の「うぢ」と「はしひめ」が,『新古今集』636 番歌では結句と初句とに分かれているが，これらに対応して類似度が計算されていることがわかる．もう一方の対についても同様である．なお，前章で示した指標では，これらの対の類似度は，それぞれ，11.5，10.5 であり，その順位は，55 位，121 位であった．さらに，例 6 と例 7 に示した和歌対の類似度は，いずれも 19 となり，これは全体の 5 位に上昇した.

なお，ここでは指標の「変更」を行ったが,「改善」をねらったものではないことに注意されたい．すなわち，前章の指標とくらべ，ここで示した指標が優れているわけではない．いうまでもなく，視点が異なれば有効な指標も異なる．ここで得られた指標は，特に，句を超えて共通する語句に着目した場合において有効である.

6.2 パターンの生起頻度に基づく指標

6.2.1 共通表現の頻度

次の対に注目されたい.

例 10. (**本歌取りでない対**)

春霞/たなびく山の/さくら花/見れどもあかぬ/君にもあるかな
『古今集』684 番 (紀友則)
紫の/雲にもあらで/春霞/たなびく山の/かひはなにぞも
『新古今集』1448 番 (円融院)

この対は,『古今集』『新古今集』間で類似度を算出した際，類似度による順位が，前章で示した指標では 12 位，6.1 節で示した順序自由パターンに基づく指標では 16 位であった．この二首は,「春霞たなびく山の」が共通するために，類似度が高くなったものである．しかし，このような表現は，和歌にはありふれたものである．一般に，共通表現の頻度が比較的高い場合には，特定の二首についてその類似性を指摘することの価値は，まず認められない．逆に，二首間の共通表現がそれ以外の和歌でほとんど見られないときには，そこに表現上の直接的な影響関係のある可能性が高い.

6.2.2 新しい指標の定義

本稿では，類似性指標を，パターン集合とパターンスコア関数の対として捉える．前章で示した指標と 6.1 節で示した指標は，いずれも，パターンのスコアを統語的に与えるものであった．これに対し，以下では，パターンのスコアを，パターンの生起頻度に依存した方法で与える．

S を Σ^+ の有限部分集合とし，S に属する文字列の類似度のみを考えることにする．パターン π の S に関する稀少度 (rarity) を以下のように定義する．

定義 4. パターン π の S における生起確率を $\Pr(\pi; S)$ とする．このとき，

$$\log_2(1/\Pr(\pi; S))$$

を，パターン π の S に関する**稀少度** (*rarity*) とよぶ．

パターンの稀少度をそのままパターンのスコアとする類似度指標を考えよう．すなわち，類似度の高い文字列対とは稀少度の高い共通パターンをもつ対である，ということになる．稀少度の定義はパターンの形式に依存しないため，類似性指標のパターン集合として任意のものを用いることができる．

稀少度の考え方に基づいて，和歌の類似性を扱うための新しい指標を，以下のように定める．

稀少度に基づく指標．
パターン集合： $\mathit{\Pi} = \{\pi \mid \pi \text{ は } \Sigma \text{ 上の順序自由パターン }\}$.
パターンスコア関数： 次で定まる写像 $\mathit{\Phi} : \mathit{\Pi} \to \boldsymbol{R}$

$$\mathit{\Phi}(\pi) = \log(1/\Pr(\pi; S)) \qquad (\pi \in \mathit{\Pi}).$$

ここでは単純に，以下のようにした．パターン $\pi(u_1, \ldots, u_n)$ の生起確率について，次のように仮定する．

$$\Pr(\pi(u_1, \ldots, u_n); S) = \prod_{i=1}^{n} \Pr(u_i; S).$$

ここで，$\Pr(u; S)$ は文字列 u の S における生起確率を表す．生起確率 $\Pr(u; S)$ は，

$$\Pr(u; S) = \frac{|S \cap L(*u*)|}{|S|}$$

として求める．このとき，上の仮定より，

$$\mathit{\Phi}(\pi(u_1, \ldots, u_n)) = -\sum_{i=1}^{n} \log(\Pr(u_i; S))$$

を得る．ただし，長さ 1 の文字列 u に対しては特別に $\log_2(\Pr(u; S)) = 0$ とした．

6.2.3 結果

このように定めた指標を用いて,『古今集』と『新古今集』の間の比較を行った．集合 S としては，二十一代集を用いた．その結果, はじめにあげた『古今集』684 番歌と『新古今集』1448 番歌の類似度は相対的に下がり，全体の 93 位になった．頻度を考慮しない指標では 16 位であったことを考えると，目的とした効果は得られたようである．

第 5 章で示した指標を A, 6.1 節で示した指標を B, 上で示した指標を C とする．指標 B, C についての類似度の度数分布を，それぞれ，表 3，表 4 に示す．

表 2 と同様，括弧内の数は，田中・赤瀬 (1992)，久保田 (1979) の注釈書において，本歌・類想歌・参考歌など，表現上，影響関係のある歌の指摘がみられた対の数を示す．指摘のあった割合に関していえば，指標 B と C は，指標 A に比べやや劣っているようにみえる．しかし，前章で述べたように，この割合だけをもとに指標の優越性を云々すべきではない．

指標 C では，頻度を考慮した結果，指標 A, B では下位であった対が浮上した．例を三つあげてみる．

表 3. 類似度の度数分布 (順序自由パターンに基づく指標; 指標 B)．括弧内の数は，田中・赤瀬 (1992)，久保田 (1979) の注釈書に本歌・類想歌・参考歌など，表現上，影響関係のある歌の指摘がみられた対の数を示す.

類似度	度数	累積度数
21	2(2)	2(2)
20	2(1)	4(3)
19	4(4)	8(7)
18	7(5)	15(12)
17	11(9)	26(21)
16	22(11)	48(32)
15	54(20)	102(52)
14	110( )	212( )
13	247( )	459( )
12	608( )	1067( )
11	1506( )	2573( )
10	3534( )	6107( )
9	7847( )	13954( )
8	20744( )	34698( )
7	30312( )	65010( )
6	104053( )	169063( )
5	70713( )	239776( )
4	370354( )	610130( )
3	75323( )	685453( )
2	792725( )	1478178( )
1	0( )	1478178( )
0	731332( )	2209510( )

表 4. 類似度の度数分布 (頻度を考慮した指標; 指標 C)．括弧内の数は，田中・赤瀬 (1992)，久保田 (1979) の注釈書に本歌・類想歌・参考歌など，表現上，影響関係のある歌の指摘がみられた対の数を示す.

類似度	度数	累積度数
46– 48	1(1)	1(1)
44– 46	1(1)	2(2)
42– 44	2(2)	4(4)
40– 42	4(2)	8(6)
38– 40	4(3)	12(9)
36– 38	8(5)	20(14)
34– 36	17(11)	37(25)
32– 34	31(18)	68(43)
30– 32	55(18)	123(61)
28– 30	117( )	240( )
26– 28	256( )	496( )
24– 26	547( )	1043( )
22– 24	1285( )	2328( )
20– 22	2850( )	5178( )
18– 20	6285( )	11463( )
16– 18	13624( )	25087( )
14– 16	27532( )	52619( )
12– 14	54444( )	107063( )
10– 12	97295( )	204358( )
8– 10	171823( )	376181( )
6– 8	253714( )	629895( )
4– 6	368142( )	998037( )
2– 4	480141( )	1478178( )
0– 2	731332( )	2209510( )

例 11.　(類型表現)

きみが世は/限もあらじ/ながはまの/まさごのかずは/よみつくすとも
『古今集』1085 番 (よみ人しらず)

君が代の/としのかずをば/白妙の/浜のまさごと/たれかしきけむ
『新古今集』710 番 (紀貫之)

指標 A, B, C のもとでの上の対の順位は，それぞれ，916 位，213 位，36 位であった．指標 C において生起確率を考慮したことにより，順位が大きく上昇していることがわかる．最大スコアを与えた共通の順序自由パターンは，

$$\pi(\text{きみがよ}, \text{はまの}, \text{まさご}, \text{かず})$$

であった．上の 2 首は，いずれも，限りない「きみがよ」の数を浜辺の砂粒の数に託したもので，賀歌の表現類型の一つということができるが，これらの語句の組合せは珍しい．これは，各句索引を用いたのでは検索しにくい対である．

例 12.　(異伝)

萩が花/ちるらむをのの/つゆじもに/ぬれてをゆかむ/さ夜はふくとも
『古今集』224 番 (よみ人しらず)

あき萩の/さき散る野辺の/夕つゆに/ぬれつつきませ/よはふけぬとも
『新古今集』333 番 (柿本人麿)

この対の指標 A, B, C のもとでの順位は，それぞれ，620 位，213 位，38 位であった．最大スコアを与えた共通の順序自由パターンは，

$$\pi(\text{はぎ}, \text{ちる}, \text{つゆ}, \text{ぬれ}, \text{よはふ}, \text{とも})$$

であった.『新古今集』歌は，万葉歌を再録したものである．また,『古今集』歌はよみ人しらず歌で，詠歌年代は古今集時代よりも古いと思われる．したがって，どちらも古歌であって詠歌年代の前後関係は判断しにくいが，単なる類想歌という以上の共通性があり，異伝 (一方の歌から他方の歌が派生した) の可能性もある.『古今集』のほとんどの注釈書は，この万葉歌を指摘している．ぜひ，順位を上げたい対である．

例 13.　(同一技巧)

けさはしも/おきけむ方も/しらざりつ/思ひいづるぞ/きえてかなしき
『古今集』643 番 (大江千里)

あさ露の/おきつる空も/おもほえず/きえかへりつる/こころまどひに
『新古今集』1172 番 (源周子)

この対の指標 A, B, C のもとでの順位は，それぞれ，52997 位，1068 位，67 位であった．指標 A と比べると，順位が劇的に上昇していることがわかる．共通パターンは，

$$\pi(\text{おき}, \text{りつ}, \text{おも}, \text{きえ}, \text{つる})$$

であった．古今歌は「しも(霜)」，新古今歌は「露」であるが，どちらも「置く」ものである．これが「起く」との掛詞になっており,「しも」「露」の縁語「消ゆ」(死ぬ)が用いられ

るという表現技巧が共通している．両歌は，いずれも後朝 (きぬぎぬ) の歌であり，この特定の場面における共通の表現技巧を見出したと言える．

このように，和歌の類似性を，順序自由パターンの形式で捉え，さらに頻度を考慮することで，既知の常套表現をできる限り排除して，より緊密な類似性をもつ歌の対を得ることができた．

7. むすび

従来，古典和歌の表現に関する研究は，ある一つの表現に着目し，それと同じ用例を博捜した上で，その結果を検討するという方法をとってきた．その際に,『新編国歌大観』の各句索引で検索しようと,『新編国歌大観』CD-ROM 版付属の検索ソフトウェアを用いようと，根本的な手法の差はないと言ってよい．だが，その方法では,『新編国歌大観』に収載された 45 万首の歌を対象に，表現上の特色を経験と勘とで見つけ出し，古典和歌の表現における継承と変容の歴史を把握するのに，これから先，どれだけの時間と労力がかかるであろうか．

そこで本稿では，これまでの検索とは全く異なる新たな方法を提案した．すなわち，一首の和歌を単なる文字列として捉え，二つの歌集間で，共通する文字列を多く含む和歌の対を計算機プログラムにより抽出するという手法である．これにより，表現上の類似歌を発見することができる．これまでの研究方法は，この，和歌研究の端緒となる段階にたどり着くまで，手作業で個々の用例にあたるしかなかったが，それを，計算機プログラムに肩代わりさせ，網羅的に提示させようというのである．そして，古典和歌における多様な類似性を鑑み，異なる類似性指標を設定してみた．ここでは，一首の歌を五句に分割した上で，句ごとに共通文字列をとる方法と，句を分割せずに，共通文字列をとる方法を試み，さらに後者については，その表現の生起頻度を考慮した．その結果，従来の方法では見つけ出しにくい類似歌の対が，少なからず指摘されたことで，これらの類似性指標の有効性を示すことができた．

しかし，本稿で示した指標では，次のような類似性をうまく扱うことができない．たとえば,「たてるやいづこ」「たてるはみやこ」という句を口ずさんでみると，たいへんよく似た感じがする．それは，同じ 7 拍の句の中に，共通文字列「たてる」と「こ」が，句中の同じ場所，すなわち，句頭から 3 拍目までと，句末の 1 拍に位置しているからであろう．そしてさらに言えば，共通パターン「たてる∗こ」のワイルドカード∗に合致する「やいづ」と「はみや」についても,「や」と「は」，および「い」と「み」は，それぞれ母音が共通している．ところが，句間の類似性指標として山崎 他 (1998) で示した指標を用いたのでは，このような対の類似度は低くなってしまう．こういった類似を捉えるためには，単に文字の一致・不一致を見るのではなく，母音または子音の一致までも考慮した，新しい類似性指標を定義する必要がある．

もし，長さの等しい句の間だけで類似性を考えるのであれば，対応する文字が一意的に定まるため，そのような指標は比較的容易に定義できる．しかし，実際には,「ひとしれぬ」と「ひとにしられぬ」など，5 拍の句と 7 拍の句が類似している場合が少なからず見られる．第 5 章で示した指標では，これら二つの句の共通パターン「ひと∗し∗れぬ」において共通文字列が三つに分離しているため，高い類似度は与えられない．この類の例としては，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」などがある．このような対を類似したものとして扱うためには，同じ仮名が句中で何番目に配置されているか，といった規則性を見出し，それに基づいて類似度

を定義する必要があろう．

本稿では，類似歌を計算機プログラムによって抽出する手法を述べてきたが，得られた結果のどの和歌の対に着目し，和歌史的にどう意味づけていくかは，言うまでもなく，和歌研究者に委ねられている．計算機は，研究の糸口を示すに過ぎない．だが，古典和歌 45 万首には，その数の多さに加え，成立年代未詳の歌集や，ほとんど研究の手が及んでいない歌集が含まれている．それらを研究対象として，でき得る限り多くの角度から表現を分析し，新たな知見を得て，それらを和歌史的に位置づけるためには，本稿のような計算機を用いた手法の導入が，もはや必要不可欠であろう．『新編国歌大観』CD-ROM 版の古典和歌テキストデータは，それを可能にした．この新手法による古典和歌の表現分析は，いま，まさに機が熟したと言えよう．

参 考 文 献

Brāzma, A., Ukkonen, E. and Vilo, J. (1996). Discovering unbounded unions of regular pattern languages from positive examples, *Proceedings of the 7th International Symposium on Algorithms and Computation* (*ISSAC'96*), 95–104, Springer, Berlin.

福田智子 (2000a). 「人の親の心は闇にあらねども」──藤原兼輔の歌再考 ── (投稿中).

福田智子 (2000b). 為忠集再考, 和歌文学会第 46 回大会研究発表資料.

Gusfield, D. (1997). *Algorithms on Strings, Trees, and Sequences*: *Computer Science and Computational Biology*, Cambridge University Press, New York.

久松潜一 (1971). 『歌論集 一』, 中世の文学, 三弥井書店, 東京.

門田隆史, 石野 明, 竹田正幸, 松尾文碩 (2000). 主旋律の類似性について, 情報処理学会「人文科学とコンピュータ」「音楽情報科学」合同研究会研究報告, **2000**(49), 15–24.

久保田淳 (1979). 『新古今和歌集』, 新潮日本古典集成, 新潮社, 東京.

近藤みゆき (1999). 平安時代和歌資料における特殊語彙抽出についての計量的研究と利用ツールの公開, 文部省科学研究費特定領域研究 (A)「人文科学とコンピュータ」1998 年度研究成果報告書, 68–77.

近藤みゆき (2000). n グラム統計処理を用いた文字列分析による日本古典文学の研究──『古今和歌集』の「ことば」の型と性差──, 千葉大学人文研究, **29**, 187–238.

Laird, P. D. (1988). *Learning from Good and Bad Data*, Kluwer, Dordrecht.

村上征勝, 今西祐一郎 (1999). 源氏物語の助動詞の計量分析, 情報処理学会論文誌, **40**(3), 774–782.

Rissanen, J. (1978). Modeling by the shortest data description, *Automatica*, **14**, 465–471.

Shinohara, T. (1982). Polynomial-time inference of pattern languages and its applications, *Proceedings of the 7th IBM Symposium on Mathematical Foundations of Computer Science*, 191–209.

竹田正幸, 福田智子, 南里一郎, 山崎真由美 (1999). 和歌データベースにおける特徴パターンの発見, 情報処理学会論文誌, **40**(3), 783–795.

Tamari, K., Yamasaki, M., Kida, T., Takeda, M., Fukuda, T. and Nanri, I. (1999). Discovering poetic allusion in anthologies of classical Japanese poems, *Proceedings of the 2nd International Conference on Discovery Science* (*DS'99*), 128–138, Springer, Berlin.

田中 裕, 赤瀬信吾 (1992). 『新古今和歌集』, 新日本古典文学大系, 岩波書店, 東京.

山崎真由美, 竹田正幸, 福田智子, 南里一郎 (1998). 和歌データベースからの類似歌の自動抽出, 情報処理学会「人文科学とコンピュータ」研究会研究報告, **98**(97), 57–64.

Yamasaki, M., Takeda, M., Fukuda, T. and Nanri, I. (2000). Discovering characteristic patterns from collections of classical Japanese poems, *New Generation Computing*, **18**(1), 61–73 (Preliminary version in: Proc. DS'98 (LNAI 1532)).

Discovering Similar Poems from Anthologies of Classical Japanese Poems

Masayuki Takeda

(Department of Informatics, Kyushu University)

Tomoko Fukuda

(Fukuoka Jo Gakuin College)

Ichirō Nanri

(Junshin Women's Junior College)

Mayumi Yamasaki and Koichi Tamari

(Department of Informatics, Kyushu University)

WAKA is a form of traditional Japanese poetry with a 1300 year history. In this paper we attempt to semi-automatically discover similar poems in anthologies of WAKA poems. The key to success is how to define the similarity measure on poems. We introduce a unifying framework that captures the essence of string similarity measures. This framework makes it easy to design new measures appropriate for discovering similar poems. We proposed three types of similarity measures. Using them, we report successful results in finding similar poems between Kokinshū and Shinkokinshū, which are known as the best two of the twenty-one imperial anthologies. Most interestingly, we have found several instances of poetic allusion which have never been pointed out in the long history of the WAKA research.

Key words: Classical Japanese poems, analysis of expressions, similarity measures, similar poems, machine discovery.