4A1-5

音素記号列からのテキストマイニングによる古典和歌の分析

Analysis of classic Waka with text mining from phoneme patterns

大柿高志 *1 Takashi Ookaki　吉仲亮 *2 Ryo Yoshinaka　山本章博 *3 Akihiro Yamamoto

*1京都大学工学部情報学科 Faculty of Engineering, Kyoto University
*2京都大学大学院情報学研究科 Graduate School of Informatics, Kyoto University

In recent years, many traditional waka (Japanese traditional poets) are stored in databases. This situation enables us to apply data-mining methods to the waka data. Before the situation appears, researchers on traditional waka in literature paid attention to the phrase of independent words, while in recent research, by applying mining methods to traditional waka we can extract various kinds of similarity among waka, including to find common Kana patterns .

In this research, we propose a method of mining from traditional waka databases with phoneme string, in order to take rhymes and sound symbolism into account in mining from waka. we assume a hypothesis that the more frequently phoneme pattern emerge in a waka, the more important it is. We define a refrain phoneme pattern as a pattern which appears in a waka more than two times.

1. はじめに

近年, 古典文学テキストのデータベース化が進んだことにより, マイニングの技術を文学研究, 特に古典和歌においても利用しようという動きが起こっている. これまでテキストマイニングで扱われるデータは, 自然言語処理の成果を利用するため, 形態素解析, 構文解析, 意味解析などで処理され, 音素記号列として解析することはほとんどなかった. しかし, 和歌は朗詠の文化という側面も持っている. 和歌データを音素の列とみなして分析する手法を提案することは, 新しい文学研究の方法として価値あることと考えられる. 例えば, 次の短歌

> 最上川/逆白波の/たつまでに/
> ふぶくゆふべと/なりにけるかも
> Mogamigawa/sakashiranamino/tatsumadeni/
> fubukuyuhubeto/narinikerukamo

について, 小説家であり, 劇作家・放送作家でもある井上ひさし [1] は, その母音に着目して次のように述べている.

> これの「音」の響きに注目しよう. とくに母音の並び方が大事だ. 五七五の上句には圧倒的に「ア」の母音が多い.（中略）「ア」は大きく, 広く, すべてを包み込む音だ. 最上河畔の吹雪という雄大な風景が「ア」の音でしっかりと捉えられている. だが, 下句の頭の音の並びはどうだ. 「ふぶくゆふべと」の七音のうち五音 (!) までもが「ウ」という母音をもっている. 「ウ」は籠った音, 吹雪の凄さに息もつけずに唸っている. そして最後の七音には「アイウエオ」の母音がすべて登場し, 歌が閉じられる. じつに見事な母音操作である.

上記の例は国文学研究において必ずしも正統的なものではない. しかし, このような言語音と意味（概念）との間の恣意的でない結びつきは音象徴と呼ばれ, 心理学の分野では言語音と図形イメージとの連想関係を示す「ブーバ/キキ効果」などがよく知られている. また, 日本語の形容語の場合においても, 「軽い」,「苦しい」など複数の語に対して, 音素の種類に応じて特定のイメージが喚起されやすいことは心理学実験により確かめられている [2]. 計算機の分野では, 三浦ら [3] が, 「強い」音象徴をもつ名前を判定する実験を行い, それが機械学習可能であることを確認した. また, 音象徴をデータマイニングに応用した研究としては五十嵐 [4] がオノマトペを利用した評判分析の研究を行っている.

本研究の目的は, テキストにおける文字情報を音素記号列とみなすことで, 韻および音象徴を考慮したマイニングを行うことを提案し, それを古典和歌の分析に適用することである. 和歌の音素パターンに何らかの有意性があるという仮説を立てた上で, 実際に頻出音素記号列を抽出し, その分析を行った.

2. 先行研究

2.1 計算機を用いた古典和歌の分析

2.1.1 語句レベルの分析

国文学の研究においては,「梅に鶯」,「紅葉に鹿」といった特定の組み合わせをもった語句（自立語）やその流行による分析が主であった. それに対し, 国文学研究における計算機の初期の利用法は, 索引やデータベースの作成やそれに対するキーワード検索であった. これは単純作業にかかる時間を短縮することで, それまで人手では事実上不可能なほど膨大だった検証を可能としたた点で, 大きな意義を持つものとされている [5].

2.1.2 かな文字レベルの分析

キーワードによる検索は, 文学者があらかじめ着目する語句を勘を頼りに定めておかねばならないという点で問題があった. 近藤 [6] は N-gram 統計を利用し, キーワードをあらかじめ指定せずに研究上重要と思われる文字列を抽出した. また, 竹田ら [7] は助詞・助動詞などによる言い回しを重視し, かな文字列に対する類似性指標を定義することで類似歌の抽出を行った. 類似歌を分析した結果, それまで見つかっていなかった新たな本歌取りが発見された.

竹田によれば, 形態素解析の代わりに和歌をかな文字列とみなすことの利点としては, 前述した近藤が

連絡先: 大柿高志，ookaki@ii.ist.i.kyoto-u.ac.jp

- タグ付け作業を必要としないこと
- 一つの言葉に 2 つの意味を持つ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-gram 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

などを指摘していると述べている.

このように, 計算機を用いた和歌研究においては, 単語情報などを捨てて完全に文字列の連鎖として扱うことで, 解釈における曖昧性を排除するという立ち場での研究が進んでいる.

2.1.3 音素レベルの分析

六井ら [8] は構築した和歌情報可視化システムにおいて, 和歌における韻を考慮した. 各句の最後の文字の母音を属性として用いて, 和歌の種類を識別する実験を行い, 作者が韻を作為的にこめることを示した.

2.2 本研究の位置付け

本研究は, 和歌のテキスト全体を音素列とみなし, また, 音象徴の影響を考慮しているなどの点で六井らの研究とは大きく異なる. 和歌や文章を音素記号列として解析し, データマイニングを行う方法は, 竹田らの行う文字列解析と同様の利点を保ちつつ, 言葉の音をも考慮したより豊かな情報を抽出できる可能性がある.

3. 音素記号列からのマイニング

3.1 音素記号と音素記号列

かな文字は五十音表にしたがい, 母音+子音の組みで表される. 五十音表を構成する音素を記号化して表したものを音素記号と呼ぶものとし, また, 音素記号の列を音素記号列とする. 半母音については子音と区別せず, 子音記号に含まれるものとする. 撥音「ん」は/n/とする. 促音「っ」および長音符「ー」については, 和歌には現れないため考慮しない. 旧仮名遣いである「ゐ」および「ゑ」はそれぞれ/i/, /e/で表すものとした.

3.2 同音反復パターン

多数の和歌にひとつずつ現れるパターンと, 和歌一首内において複数回現れるパターンでは, 出現頻度は同じでも, 後者の方が和歌一首ごとの特徴を見る上では重要であるという考えに基づき, 次のような同音反復パターンを定義する.

同音反復パターンの定義音素記号列で表された文（和歌一首）において頻度 2 以上であるすべての頻出部分文字列を, その和歌における同音反復パターンと呼ぶものとする.

同音反復パターンのうち, 頻出度が高いものほど音象徴を強調し, また, そのパターンの長さが長いほど特異的なパターンであると考えられる. ここで, 頻出度最大の同音反復パターンを最大頻度同音反復パターン, 長さ最大の同音反復パターンを最大長同音反復パターンと定義する. 音素記号列からのマイニングを行う上で, このような同音反復パターンを求めることは, 和歌分析における重要な手掛かりとなり得ると考えられる.

4. 万葉仮名, 上代特殊仮名遣いに関する扱い

８世紀の奈良時代における母音体系は「上代特殊仮名遣い」などと呼ばれ, 平安時代初期に生まれた平仮名, 片仮名とは大きく異なっていたと考えられている. そのため, 本研究におけるデータセットからは, ８世紀以前の和歌は除外した. ただし, 9 世紀以降であっても, 厳密には発音は時代ごとに変化している. 本論文においては, 9 世紀以降における発音の変化は万葉仮名などと比べて現代仮名遣いに近いとみなし, 単純に五十音表にしたがって音素記号への変換を行うこものとした. 厳密な発音については考慮していない.

5. 計算機実験

実験に用いたデータセットはすべて国際日本文化研究センター [9] の和歌データベース (190,420 首) および俳諧データベースから, かな文字で清音表記された和歌を抽出し, 母音または子音の音素記号列に変換したものを用いた. また, そうして得られたデータのうち, 必要に応じてその一部を実験に利用した. ただし, 前述したように「上代特殊仮名遣い」が使われている 8 世紀 (700 年代) 以前の歌集は除外した. また, 清音かな表記でない漢字のみのデータも除外するものとした. ただし, 和歌の一部または全体が欠損したデータは各データセットに含まれている場合がある. データセットに対して行った頻出文字列マイニングは, 坪井 [10] の提案したアルゴリズムを利用した. 頻出部分文字列マイニングを行った結果は, データの傾向を分析するため, 頻出度順でランキングを行った. 結果に対するソートは, Microsoft Office Exel 2007 を使用した. なお, 頻出度, 抽出された頻出文字列とともに, 一首あたりにパターンの出現する割合を求めた.

5.1 作品集成立年代ごとの和歌および俳句の比較

5.1.1 実験

データセットデータベースから抽出した和歌を作品集成立年代ごとに切り分け, 母音および子音の音素記号列に変換した. 一律に 100 年ごとに分割し, 作品集成立年代が不明な歌集はデータセットから除外した.

実際のデータセットは以下のようになった. 800 年代 (811 首), 900 年代 (14,800 首), 1000 年代 (7,325 首), 1100 年代 (24,543 首), 1200 年代 (60,967 首), 1300 年代 (63,302 首), 1400 年代 (16,191 首), 1500 年代 (2,587 首)

俳句のデータセットは, 同じく国際日本文化研究センターにある俳諧データベースのうち,「冬の日 6 巻」,「春の日 4 巻」,「春の日」,「阿羅野員外 10 巻」,「阿羅野」,「ひさご 5 巻」,「猿蓑 4 巻」,「猿蓑」,「炭俵 8 巻」,「炭俵」,「続猿蓑 5 巻」,「続猿蓑」をひとまとめのデータセット (3,486 首) とした.

実験方法各年代ごとの和歌データセットおよび俳句のデータセットに頻出部分文字列マイニングを行った結果を頻出度の高いものから順に 50 位までランキングしたが, 必要に応じて文字列の長さごとでも順位を比較した. 結果からわかる音素記号列のパターンの性質・傾向について考察した. さらに, 特徴的だと考えられるパターンが実際に現れる和歌をランダムに取り出し, どのような語句や言い回しにパターンが現れているかを確認し, 考察を行った.

5.1.2 結果・考察

頻出パターンのランキングを行った結果の一部を表 1, 表 2 としてまとめる.

各年代ごとの比較母音, 子音ともに多少の順位の変動があるものの, 現れる頻出音素パターンは各年代でほぼ同じであり, 頻出傾向に目立った違いは見られなかった.

一方, 個別のパターンを見ていくと, 例えば母音のパターンについて, どの年代においても/o/の連続したパターンおよび/a/の連続したパターンが多い. 例えば, 表 2 において, 頻出度の 1 位と 2 位は, どの年代においても常に 1 位/ooo/, 2 位/aaa/となっている.

表 1: 長さ 1 の頻出音素パターンのランキング

順位	800s	900s	1000s	1100s	1200s	1300s	1400s	1500s	俳句
1	a	a	a	a	a	a	a	a	a
2	o	o	o	o	o	o	o	o	i
3	i	i	i	i	i	i	i	i	o
4	u	u	u	u	u	u	u	u	u
5	e	e	e	e	e	e	e	e	e

表 2: 長さ 3 の頻出音素パターンのランキング (一部)

順位	800s	900s	1000s	1100s	1200s	1300s	1400s	1500s	俳句
1	ooo	ooo	ooo	ooo	ooo	ooo	ooo	ooo	aaa
2	aaa	aaa	aaa	aaa	aaa	aaa	aaa	aaa	iaa
3	ioo	aai	aai	aai	oaa	oaa	oaa	oaa	aai
4	oaa	oaa	ioo	oaa	aai	aai	aio	aio	ooo
5	aio	ioo	oaa	iaa	aio	aio	aai	aai	aua
...	...	...	...	...	...	...	...	...	...

このようなパターンの頻出度に関する特徴は, 母音, 子音を問わず各年代を通してすべて共通に見られたことから, 和歌における音素の頻出パターンに時代を越えて普遍的な性質があると考えられる.

ただし, このようなパターンの特性は, 和歌の性質によるものだけとは限らず, 日本語一般において共通する性質である可能性がある. そこで, 和歌と俳句を比較することでこれを考察する.

和歌と俳句の比較和歌と俳句では, 年代ごとに分割した和歌の比較実験と同様に, 似たような頻出パターンが現れている. ただし, 年代ごとに和歌を比較した場合よりも, 和歌と俳句の比較の方が, 抽出される頻出パターンは異なったものが多い. 例えば、表 2 を見ると, 和歌のデータにおいては, どの年代においても/ooo/のパターンは/aaa/のパターンよりも頻出だが, 俳句は/ooo/のパターンの頻出する割合が和歌よりも格段に低い. このように, 年代ごとの比較と比べてやや違いが見られた.

俳句との間に頻出パターンの差異が現れる原因は, 和歌と俳句の音に関する性質によるものと考えられる. 俳句の詠まれる年代が 1600 年代に偏っていることから, 単なる時代による違いとも考えられるが, それでは他の年代ごとの比較と比べて, パターンの違いが大きい理由を十分に説明できない.

パターンの含まれる和歌ある程度の長さを持ち, 頻出度が特に高いパターンについて, 実際にどのような語句や和歌に現れているかを確認した. 具体例として, 長さ 4 以上の母音のパターンのうち, 頻出度が最大である/oooo/のパターンについて調べてみた. 以下はその一例である.

おおはらや/をしほのやまも/けふこそは/
かみよのことも/おもひいつらめ
Ooharaya/wosihonoyamamo/kehukosoha/
kamiyonokotomo/omohiiturame

また, 「こころ」や「おもほ（おもふ)」, また逆説の接続助詞「とも」などが, /o/の連続したパターンを含む語句として各時代共通に繰り返し確認された. このほかに多く現れる語句としては, 「ほとときす」, 「(の) とこよ」,「このよ」などがある. これらの語句が「とも」,「の」,「(と, を) こそ」などの/o/を含む付属語とともに現れることで/o/の連続パターンが形成されている. 抽出された語句の特徴として, 抽象的イメージを喚起する語句が多いが, これは同じ音象徴/o/を連続していることが, 語句の抽象性を高めていると考えることもできる.

このような音素パターンと和歌の関係は, 研究者が和歌における音を考慮した詳細な分析を行うための基礎づけとなりうると考えている.

一方, 繰り返し現れるような語句が存在せず, 和歌ごとのさまざまな言い回しに含まれるようなパターンも散見された. これは, /oooo/のような同じ音素/o/の連続の場合, 単純に/o/を複数個含む語句の組み合わせによってパターンが形成されているのに対し, /aaiu/のようなパターンは組み合わせの数がより複雑になり, 珍しい言い回しのなかにしか現われないためだと考えられる.

5.2 歌の種類による比較

5.2.1 実験

データセット古今集は, 各巻ごとにテーマが分かれており, 巻一:春上, 巻二:春下, · · · · · , 巻二十:大歌所歌, 異本所載歌となっている. このうち季節の歌と恋の歌に着目し, それぞれのデータセットを作成した. 古今集の巻一:春上から巻六:冬までを季節の歌 (342 首) とし, 巻十一:恋一から巻十五:恋五までを恋の歌 (360 首) とし, それぞれデータセットとして用いることとした.

実験方法古今集の季節の歌, 恋の歌を母音および子音の音素記号列に変換し, その各々に対して頻出部分文字列マイニングを行い, 頻出度の順でランキングした. また, その結果のうち特徴的だと思われるパターンに対して, 実際にどのような和歌に現れているかを確認し, 分析を行った.

5.2.2 結果・考察

どちらにも共通して頻出するパターンが多く見られた一方, 頻出パターンの出現頻度の差も見られた. 例えば表 3 より, すべての長さのパターンのランキングにおいて, 季節の歌で 26 位だった/ooo/パターン (一首あたり 0.69) は, 恋の歌においては 14 位 (一首あたり 1.19) であり, その出現頻度の比はおよそ 1.7 倍である. また, 子音についても歌の種類により頻出パターンの違いが見られた.

特に特徴的な出現頻度の差が見られた/ooo/のパターンについて, 季節の歌, 恋の歌それぞれでパターンを含む和歌を調べ

表 3: 季節の歌と恋の歌の比較 (古今集)

順位	パターン	季節の歌のパターンの出現頻度 (A)	恋の歌のパターンの出現頻度 (B)	比 (B/A)
1	ooo	0.690	1.191	1.727
2	oio	0.406	0.602	1.483
3	uoo	0.304	0.447	1.470
4	ooa	0.491	0.716	1.459
5	oo	0.211	0.301	1.422

てみると,「こころ」,「おもふ」(ものおもふ) といった心情に関わる語句が多く, 恋の歌ほどそうした抽象的な想いを歌に詠むのではないかと思われる.

5.3 最大長同音反復パターンによる韻の抽出

データセット以下の和歌を対象データとした.

たちわかれ/いなばのやまの/みねにおふる/
まつとしきかは/いまかえりこむ

実験方法音素記号列に変換済みの和歌一首に対して頻出文字列マイニングを頻度閾値 ξ=2 として適用し, 同音反復パターンを求める. 求めた同音反復パターンのうち, 最大長同音反復パターンを求め, 分析を行った.

5.3.1 結果・考察

求めた同音反復パターンのうち, 最大長同音反復パターンとして/aiaaei/が得られた. しかし, このとき同音反復パターンが現れる和歌のかな文字列を見ると,「たちわかれ/い」,「は/いまかえり」であり, 句切れをまたいでしまっている. このように, 必ずしも韻を踏んでいるとはみなせないパターンが抽出されることがわかった. 和歌の韻を考慮したマイニングを行うためには, 同音反復パターンの現れる位置, 句切れの頭か接尾に接しているか, といったことを考慮する必要があると考えられる.

6. まとめと今後の課題

時代ごとおよび和歌と俳句の頻出部分文字列抽出の結果だけでなく, 季節の歌と恋の歌の比較においても母音, 子音ともに頻出パターンにかなりの類似が見られた. これは, 和歌の音素パターンに, その時代や和歌の種類などを超えた一定の傾向が存在することを示唆している. 時代や和歌の種類によって用いられる語句が変化していくことを考えれば, 注目すべき結果だと言える.

将来的には, 普遍的な傾向をもつ音素パターンのなかから, 時代ごとに特徴的なパターンを発見することを目標としている. 題材の流行や意味を踏まえて選択された語句が, 音韻においてもその時代・作者などの特徴を反映したものであることが示せれば, 文学研究に新しい視点をもたらす, 大きな意義をもつと考えられる. 今後, 国文学研究者の方からさらなる意見をもらうことが重要だと考えている.

参考文献

[1] 井上ひさし:井上ひさしコレクション 人間の巻:みごとな音の構築 (2005)

[2] 丹野眞智俊:日本語音韻における音象徴の存在, 神戸親和女子大学児童教育学研究 (2003)

[3] 三浦智, 村田真樹, 保田祥, 宮部真衣, 荒牧英治:音象徴の機械学習による再現:最強のポケモンの生成, 言語処理学会 第 18 回年次大会 発表論文集 (2012)

[4] 五十嵐沢馬, 笹野遼平, 高村大也, 奥村学:オノマトペの音象徴を利用した評判分析, The Association for Natural Language Processing(2012)

[5] 竹田正幸, 福田智子:古典和歌からの知識発見-モビルスーツを着た国文学者-, 43 巻 9 号 情報処理 (2002)

[6] 近藤みゆき:『古今和歌集男性特有表現一覧 (改訂版)』 : N-gram 分析による古典研究のこれまでとこれから, 實踐國文學 (2011)

[7] 竹田正幸, 福田智子, 南里一郎, 山崎真由美・玉利公一:和歌データからの類似歌発見, 統計数理, 第 48 巻第 2 号 289.310(2000)

[8] 六井淳, 辻野晃一, 山本桃子:ラフ集合論に基づくオンライン情報統合化～和歌情報可視化システムの提案～, The 21st Annual Conference of the Japanese Society for Artificial Intelligence(2007)

[9] 国際日本文化研究センター | データベースの案内 http://www.nichibun.ac.jp/graphicversion/dbase/database.html

[10] 坪井祐太:頻出部分文字列のマイニング (抽出, マイニング)(言語理解とコミュニケーション), 電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション (2003)

[11] 工藤浩, 小林賢次, 真田信次, 鈴木泰, 田中穂積, 土岐哲, 仁田義雄, 畠弘巳, 林史典, 村木新次郎, 山梨正明:日本語要説, ひつじ書房 (1993)