モデリングによる歌ことばの変遷と分析
—八代集・歌ことばシソーラスの開発—

山元 啓史

オーストラリア国立大学

要 旨

八代集の語彙、古今集 (905 年頃) から新古今集 (1205 年) まで約 9500 首、300 年間の歌ことばの変遷を、共出現パターンによるモデルを通して分析した。各歌を品詞分解し、その語を用いて分類語彙表に準拠したシソーラスを開発した。同じ歌に共出現する任意の 2 語を 1 パターン (共出現パターン) としてすべてのパターンを生成した。共出現パターンには、どのような歌に出てくるような特徴のないものも含まれるため、共出現ウエイトを定義し、ウエイトの高いパターンから順に、グラフによって可視化モデルを作図した。300 年間の歌ことばの変遷を観察するために、対象語について歌集毎にモデルを作成し、それらのモデルを相互に比較し、分析した。

An Analysis of Historical Changes of Japanese Poetic Vocabulary using Computer Modelling
—Poetic Thesaurus for the *Hachidaishū* (ca. 905–1205)—

Hilofumi Yamamoto

The Australian National University

Abstract

This paper addresses the methods of computer modelling of Japanese poetic vocabulary and the analysis of the historical changes of poetic vocabulary in the *Hachidaishū* anthologies. Using the poems of the *Hachidaishū* (the Eight Anthologies compiled by the order of Emperors, ca. 905–1205), I generate co-occurrence patterns consisting of any two words in each poem, and calculate the co-occurrence weight of each pattern. I analyse the historical differences between the models of earlier anthologies and those of later anthologies through the visualization of those models.

1 はじめに

八代集は万葉集に続く和歌の基礎資料としてだけでなく、古代語研究においても重要な資料である。従来より、八代集を材料とした計量分析は多数あり [1, 2, 3]、八代集の索引データベースも開発されている [4, 5]。しかしながら、多くの研究では大系本等の索引を利用しているため、単語の長さや表記が一定でなく、コンピュータで単語を集計したり、歌集間で語彙を比較したりするのは困難であった。また、古典研究で用いられるデータは底本所有権、校訂翻刻作業の著作権などの理由により、一般公開、研究データの再利用、研究成果の追試、再検討なども容易ではなかった。しかし、最近では国文学研究資料館の二十一代集データベースおよびその検索システム (http://ocelot.nijl.ac.jp/dlib/21dai/)、ヴァージニア大学日本語テキストイニシアティブ (http://etext.virginia.edu/japanese/)、関西大学図書館の「八代集の世界」では印影本の画像公開 (http://www.lib.kansai-u.ac.jp/etenji/hachidaisyu/)、高千穂大学の渋谷栄一氏翻刻による電子テキストの配布 (http://www.takachiho.ac.jp/~eshibuya/) など、電子化された資料がインターネットで公開されるようになり、データの共有や追試が実施しやすくなってきている。

山元 [6, 7] は歌ことば (いわゆる歌語) と歌枕 (和歌に見られる地名) のモデル化を検討し、語と文脈の関係、モデル中の語彙と実際の和歌の関係など、体系と詳細がモデリングによって、一瞥できることを示した。しかし、その材料は古今集に限られており、他の歌集を用いた広範囲な歌ことばのモデリングはまだ検討されていない。たとえば「桜と吉野 (地名)」の関係である。一般的に「桜といえば吉野」とよくいわれるが「桜と吉野」の関係は古今集の時代にはまだ成立しておらず、それが成立するのは西行の時代 (新古今集) を待たなければならない [8, p.434][9, p.125] [1]。確かに山元 [6] のモデルでも「雪の吉野」は見られたが、「桜の吉野」は明瞭には見られなかった。しかし、モデリングの妥当性・信頼性を得るには、さらに新古今集のデータでもモデルを作り、その中に「桜の吉野」が見られるかどうか、確認する必要がある。

さて、八代集の語彙の転換期は後拾遺集 (あるいは拾遺集) あたりからといわれている [1, 3, 10, 11]。八代集は古今集 (約 905 年) から新古今集 (1205 年) までの 300 年間に成立した 8 つの勅撰和歌集である。各勅撰集は 300 年の間に等間隔で成立しているわけではなく、その間隔はさまざまであり、中でも拾遺集と後拾遺集の間は最も大きく、約 80 年の隔たりがある (図 1)。

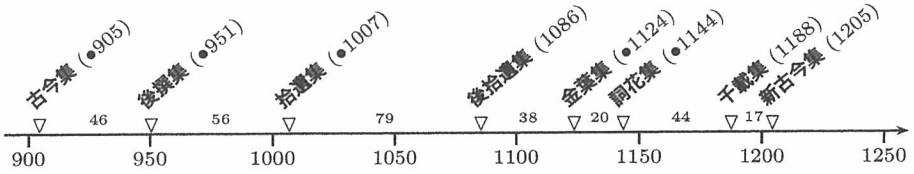

図 1: 八代集の各勅撰集の成立年とその間隔。● はおよそ。▽ 間の小数字は成立間の年数。

上野 [10, p.6–8] は「『古今』『後撰』両集の時代は、〈褻の歌〉の時代であり、『後撰集』の〈晴の歌〉への蔑視ははなはだしかった[2]。しかし『拾遺集』になると、〈晴の歌〉は大量に採択され、(中略) 和歌史に変化がおこったのだ」と述べている。川村 [11] は後拾遺集以後の転換は、すでに拾遺集に見られることを指摘している。たとえば、「〈霞＝春の到来〉という図式は、これを勅撰集史上に窺えば、この『拾遺集』巻頭において確立し、以後霞は勅撰集の立春歌群の欠くべからざる素材として、重く用いられるに至る [11, p.375]」と述べ、またそれまで春歌の表現であった「解氷のテーマは『後拾遺集』以降においては、独立したテーマとして立春・初春歌群を彩ることになる [11, p.377]」と述べている。歌ことばの転換期が拾遺集あるいは後拾遺集あたりからとするならば「桜の吉野」は新古今になって急に歌題として詠まれたのではなく、その兆候が転換期より徐々に見られ、そして新古今に確立したことも十分予想できる。古今集・新古今集の二集を調べるだけでなく、八代集を通して均一な条件で調査すれば、推移の過程も見られるのではないだろうか。

[1] 古来、吉野は山岳信仰の霊地で万葉集には天武天皇が山入りをした歌があり、古今集では隠遁の地としての印象が強く、桜の詠んだ歌は僅かに確認されはするが、雪を読み込んだ歌の方が目立つといわれる [9, p.125]。小林 [9, p.125] は「『五代集歌枕』(藤原範兼、1107–1165、平安時代後期成立) は、34 首にわたり吉野山の作例を挙げるが、そのうち桜花を詠んだものは 7 首に過ぎない。西行以後、吉野山は桜の名所として定着したと言えよう」と述べている。

[2] 褻 (け) は「日常生活のなかでの贈答や述懐」のこと、一方、晴 (はれ) は「賀宴や歌会・歌合での歌や屏風料歌」のこと。

2 方法

2.1 材料

テキストは、国文学研究資料館作成による二十一代集データベース[3]を利用した。八代集の成立、撰者、収録和歌数を表 1 に示す。八代集収録の和歌の内、長歌をのぞく 9484 首を用いた。長歌はテキストの長さが著しく異なるものがあるため除外した。岩波新日本古典文学大系本をはじめ、八代集関連出版書籍を参考にし、清濁を補い、古文自動品詞タグ付けシステム kh [12] で単位分割した。分割の単位は国立国語研究所 β 単位に、歌の番号は新編国歌大観にしたがった。「立田／竜田／龍田」など異表記の同義語はコンピュータではを別語として扱われてしまうので、すべての語は t2c でシソーラスコードに変換し、表記を統一した[4]。本研究開発のシソーラスは、一般語、地名、人名の 3 種類である。一般語のシソーラスには旧版分類語彙表 [13] を利用した。分類語彙表は現代語用に開発されたものであるが、古代語の語彙分類にもよく用いられている。分類語彙表の中には古代語としても利用できる語もあった。広辞苑を参照し、現代語の体系にそのまま古語を当てはめてよいと考えられる時には、そのエントリを利用したが、表記を変更しなければならないものもかなりあった。また、相当する語がない場合や意味が著しく異なる場合には広辞苑・古語辞典の意味から、分類語彙表の体系にふさわしい番号を当てた。固有名詞 (地名・人名) は分類語彙表にはないので、別途作成した。一般語 48732 レコード、地名 1408 レコード、人名 49 レコード、計 50189 レコードになった。現時点では八代集に見られる語とその表記はカバーされている。「吉野川」「吉野山」のように地名が山や川と連接している場合には「吉野」だけでもモデルが計算できるように地名 (吉野) と対象 (山/川) を分割した[5]。

表 1: 八代集の詳細: *印はおよその成立年。国文学研究資料館正保版本「二十一代集」による。

	歌集名	勅/院宣	成立	撰者	首
1	古今集	醍醐天皇	*905	紀友則, 紀貫之, 凡河内躬恒, 壬生忠岑	1111
2	後撰集	村上天皇	*951	清原元輔, 紀時文, 源順, 大中臣能宣, 坂上望城	1425
3	拾遺集	花山院	*1007	花山院	1351
4	後拾遺集	白河法皇	1086	藤原通俊	1218
5	金葉集	白河院	*1124	源俊頼	712
6	詞花集	崇徳院	*1144	藤原顕輔	415
7	千載集	後白河院	1188	藤原俊成	1288
8	新古今集	後鳥羽上皇	1205	源通具, 藤原有家, 藤原定家, 藤原家隆, 藤原雅経, 寂蓮	1978

2.2 共出現パターン

ここで作成するモデルは「歌ことばの語用特性」を示すモデルである[6]。複数の和歌をひとつのモデルに統合するためには、和歌から文脈を損うことなく最小の単位で、かつ代表的な要素を抽出する必要がある。

一般的に頻度主体の語彙の計量分析を行うと、頻度の高い語は分析に含まれるが、頻度の低い語 (いわゆる低頻度語) は分析から洩れてしまう。水谷 [14, p.169] は「高使用率語が、(それが術語であっても) 案外にキーワードとしての情報量に乏しいという事実に、気づかれたに違いない」と述べ、必ずしも高頻度語がキーワード性の高い語ではないことを指摘している。石井 [15, p.25] は「低頻度語とは、一般的に用いられにくい語がたまたま用いられたために低頻度になったものではなく、用いられるべくして用いられ、しかも、低頻度であるという語であ」り、それは「語彙論的な性格にではなく、それが用いられた文章の性格に規定されている」と述べている [15, p.26]。語の必要性や文章の性格を重視する上では、高頻度語ではなく、

[3]底本は国文学研究資料館蔵「正保版本二十一代集」である (`http://ocelot.nijl.ac.jp/dlib/21dai/README-21dai.html`)。

[4]t2c (Token to Code) は単位切りした語を入力すると国立国語研究所開発旧版分類語彙表準拠のシソーラス体系コードを返すプログラム。

[5]このように個別な処理を行うと語の単位が統一されない欠点がある。本研究では、あらかじめできるだけ長い単位、短い単位でシソーラスコードを登録しておき、用語の性質に応じてプログラムのオプションスイッチで組み合せを選べるようにし、可能なかぎり、分析単位が統一されるようにした。

[6]ここでは語の意味や概念を扱っていないので、それらを直接的に示すモデルではない。和歌を言語の使用と見なし、その文脈や語の相互関係を拾いあげることによって、間接的に意味や概念を分析するものである。

むしろ低頻度語に注目して分析を進めるべきである。そこで、すべての単語について *idf* (inverse document frequency) [16, 17] を計算し、語の重要度 (キーワード性) でモデルを作ることにした。*idf* (1) はある特定のテキストにしか出現しない語か、どんなテキストにも出現する語なのかを示す値である。

$$idf(t) = \log \frac{|N|}{tf(t)} \tag{1}$$

ただし、$|N|$ はすべての資料の数 (すべての和歌の数)。$tf(t) = 0$ の時は、$idf(t) = 0$ とする。$tf(t)$ は、語 t が出現する資料の数 (和歌の数) である。ある語の *idf* 値が十分に高ければ、その語はキーワードとして利用できる。たとえば、全 9484 首中、ラ変動詞「あり」は 1201 首に出現する。(1) により、9484/1201=7.89.. となり、log 7.89 はおよそ 2.07 となる。一方、名詞「鶯」は 101 首に出現するが、これは 9484/101=93.90.. で、log 93.90 はおよそ 4.54 となり、「鶯」は「あり」よりもキーワード性が高いことがわかる。和歌は 1 首が 31 文字に限られるため、助詞・助動詞の出現に偏りが見られ、高い *idf* 値がつくものが見られた[7]。そこでこれらは計算から除外した。

一般的な文章の語彙調査を行なうとその頻度分布は、L 字型分布になるといわれている [18, p.25][19, p.6]。短い作品の場合、歌謡曲のような繰り返しの多いテキストは頻度 2 の方が 1 よりも多く、必ずしも L 字型分布にはならないが [19, p.7][8]、和歌の場合は短くても、また助詞・助動詞を除外しても、L 字型分布になった[9]。L 字型分布は高頻度語の種類が少なく、逆に低頻度語の種類が多い分布であるが、そのL字型分布について *idf* を計算すると、ちょうどL字をひっくり返して、ややゆるやかにした J 字型の分布になる。*idf* によって重要度の低い大多数の語を排除し、重要度の高い語を分析に用いることができる。

次に各和歌から共出現パターンを生成し、2 語の *idf* 値とパターンの頻度を用いて共出現ウエイト *cw* を計算する。共出現パターンは単純な 2 語の組み合せではあるが、それによるモデルには元の文にある文脈が含まれることがわかっている [6]。しかしながら、すべてのパターンを描くと図は真っ黒な塊になってしまうので、すべての共出現パターンのうちから重要なパターンのみを選び出す必要がある。そこで、*tfidf* (2) [20] を拡張し、任意の 2 語のパターンが特徴的であるかどうかを評価する関数を定義する。

$$w(t,d) = (1 + \log\ tf(t,d))\ idf(t) \tag{2}$$

$$cidf(t_1,t_2) = \sqrt{idf(t_1)\ idf(t_2)} \tag{3}$$

tfidf (2) は、語 t について *idf* で重要度を計算し、抽出されたテキスト d に t が含まれる頻度 *tf* (term frequency) を掛けて、t がキーワードとして有効かどうかを評価する関数である[10]。これを前半 *tf* と後半 *idf* に分け、まず、後半に相当する関数 *cidf*(3) を定義し、八代集すべての和歌を用いて計算した。*cidf* (co-occurrent document frequency) は 2 語の *idf* の幾何平均である。これを単語重要度とする。「鶯、時鳥、梅、桜、立田、吉野」の 6 語でそれぞれ検索抽出した和歌群から共出現パターンを生成し、*cidf* の分布を調べた。いずれもやや右裾広がりの正規分布に近い分布であった (図 2 左)。

つぎに、*tfidf* (2) の前半 *tf* に相当する関数 *ctf*、および *tfidf* 全体に相当する関数、共出現ウエイト *cw* (4) を定義した。

$$cw(t_1,t_2,K) = (1 + \log\ ctf(t_1,t_2,K))\ (1 + \log \frac{|N|}{ctf(t_1,t_2,N)})\ cidf(t_1,t_2) \tag{4}$$

N はすべてのテキスト群、ここではすべての八代集の和歌である。K はある条件によって抽出されたテキスト群、たとえば「吉野」で検索抽出された和歌群である。

$\log\ |N|/ctf(t_1,t_2,N)$ は *idf* の式 (1) と同じ考え方 (inverse) である。全テキスト (すべての和歌) N について調査し、2 語 (t_1,t_2) が共出現するテキスト (和歌) の数を分母に全テキスト数 $|N|$ を分子にした比を

---

[7]たとえば、係助詞の「なむ」は和歌ではあまり用いられないが、ぜんぜん用いられないわけではなく、古今集 425 番歌で用いられており、その時の *idf* 値は 9.16 という高い値が付いた。

[8]古今集 3 番「春霞／たてるやいつこ／みよしのゝ／吉野の山に／雪はふりつゝ」のように「吉野」が 1 首に 2 度詠まれる場合もあるが、およそ和歌の語は 1 首に 1 回の使用であるので、語の頻度 *tf* は歌の頻度 *df* と考えてもよい。

[9]語の使用分布がL字型分布になると言われているが、なぜ延べ語数の大部分が少数の高頻度語によってカバーされるのか、なぜ異なり語の大部分が頻度 1 の語をはじめとする低頻度語によって占められるのかはあまり明らかにされていない。類する分布型としては英単語の *Zipf* の法則が有名で、他にも 80:20 の法則、ロングテールの法則などと言われる。

[10]Google、Yahoo などの検索システムで、WEB ページの文章からキーワードを抽出するのに用いられている。

対数変換したものである。これは2語 (t_1, t_2) が共出現することは和歌の世界 N においてよくあることなのか稀なことなのかを示す。もし、ある2語の組み合せがどの和歌にも出現する場合には、その2語の組み合せには特徴がないと考える。これを組み合せ重要度とする。

$ctf(t_1, t_2, K)$ は抽出されたテキスト群 K における2語の共出現頻度 ctf (co-occurrent term frequency) であり、(2) の tf を2語に拡張したものである。共出現ウエイト cw は、2語 (t_1, t_2) について和歌全体 N から得られる単語重要度と組み合せ重要度の積に、抽出和歌群 K での共出現頻度を掛けて計算する。

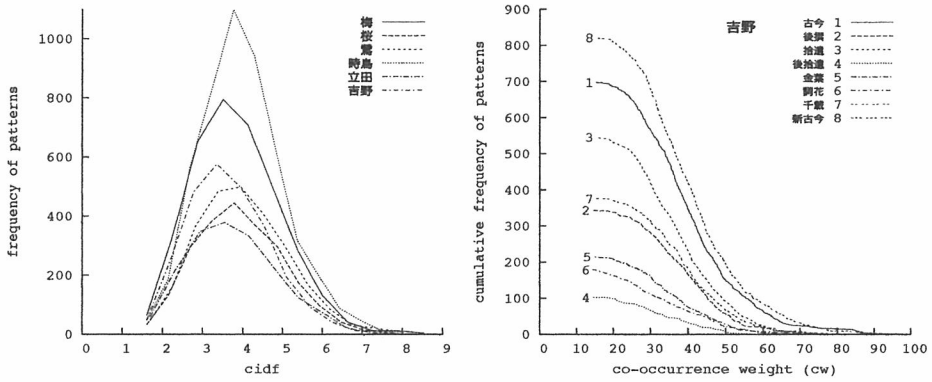

図 2: $cidf$ の分布 (左) と、cw 毎の累積パターン数 (右)：$cidf$ は「梅、桜、鶯、時鳥、立田、吉野」の6語で抽出されたテキストについて、cw 毎の累積パターン数は各歌集の「吉野」について計算した。

このようにして得られた cw は正規分布に近く、どのモデルも近い値で収束する (図2左) が、キーや抽出されたテキストの量によって多少変動があるため、このままでは相互にモデルを比較することはできない。そこで cw を z 値に変換し、正規化する。正規化した後の分布型が正規分布かどうかを確かめるために、適合度テスト (Kolmogorov-Smirnov Test) を実施したところ、すべてのモデルについて正規分布と考えてもよいことがわかった。

確率変数 x が正規分布にしたがう時、平均からのずれが $\pm 1\sigma$ 以下の範囲に x が含まれる確率は 68.26 % であり、右の 1σ の点から右裾だけを抽出すると、およそ 16 % のパターンを一律に抽出することができる。また、1σ は変曲点 (下方向きから横方向に変わる点) であり、これは不要なパターンの値から重要なパターンの値に変わる点と考えることができる。変曲点を利用すれば、調査者が任意に cw を決める必要がなくなるため、客観的なモデルの比較が可能になる。

以上の手続きで得られた共出現パターンを用いて、古今集から新古今集までのそれぞれの「吉野」のモデルを作成し、Graphviz (`http://www.graphviz.org/`) によって可視化した。

3 結果と考察

まず、モデルの可視化について述べる。八代集の各集について「吉野」のモデルを可視化した。図3左は古今集のモデルである。「山、雪、降る」が中心に見られ、「桜」は左端に見られる。山元 [6] のモデルでは「桜」は明瞭には見られなかった。前回 [6] と異なる点は、前回が古今集のデータだけで cw を計算したのに対して、今回は八代集全体の和歌を利用して cw を計算した点である。前回のモデルでは「吉野」とその他の語をむすぶ辺がほとんどで、それらは頻度の高いパターンから描かれたものであった。今回のモデルでは頻度の低いパターンも見られ、「妹背山 (古今集 828 番)」「縦しや世の中 (古今集 794, 828 番)」などのサブネット (さらに枝わかれするフラクタル状の群) が見られた。「妹背」は吉野川を流れる山の名称、「縦しや」は吉野をかけた掛詞であり、今回のモデルでは幹だけでなく、枝葉もよく見え、そこに描かれたエピソードも推測できた。

それでは、抽出テキスト数が少ない場合はどうなるのだろうか。図3右は後拾遺集のモデルである。後

拾遺集には３首しかないが、その３首が三菱の形状で描かれている。*cw* によれば、抽出テキストの数が多い場合には幹も枝葉も、抽出テキストの数が少ない場合には枝葉のモデルが得られることがわかる。

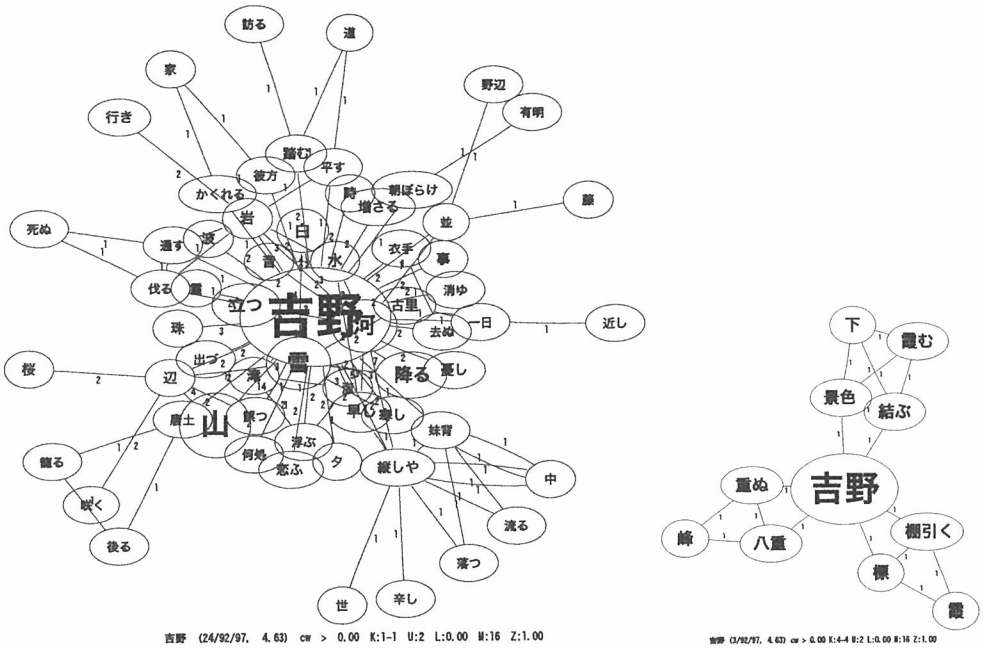

図 3: 古今集 (左) と後拾遺集 (右) の「吉野」: エッジの数字は共出現の頻度。ノードの大きさは各モデルにおけるその語の相対的頻度を示す。どちらも同じ z 値 1 以上を出力。

図 4 右は新古今集のモデルである。このモデルでは「雪、山」だけでなく「春、花、桜、峰、散る」などの「桜の吉野」に関する語が見られる。また、サブネットがいくつか見られる。特に「枝折 (しほり)」は西行の歌「吉野山／こそのしほりの／道かへて／また見ぬかたの／花を尋ねん (新古今集 86 番)」であることがわかる。「雪」から「花 (桜)」への変遷がうかがえる。

では、どの時代あたりから「花 (桜) の吉野」が見えはじめるのだろうか。可視化モデルでは後撰集から後拾遺集までは顕著に見られなかった。図 4 左は金葉集のモデルである。金葉集でも「吉野」そのものがあまり詠まれていないのだが、少ないながらも図形からは峰にかかる吉野山の桜花の関係が明らかに見える。千載集には「桜」も「雪」も見られないが「桜」の代りに「花と吉野」の関係が見られた。新古今集では「雪」が多少復活しているように見えるが、この時代には「桜」という語ではなく「花」という語が「桜」として詠まれており「桜と吉野」の関係が強くなっていると考えられる。ある関係が常套的に用いられ、その関係が確立すると「花といえば桜」というように上位語の「花」で下位語の「桜」を表現するようになる。このような提喩法[11]もうまく計算で処理したいと思うかもしれないが、すべての上位下位関係がシソーラスで解決できるほど語彙の構造は単純ではない [14, pp.180–2]。

つぎに「雪」と「桜」に限り、「吉野」の歌群から共出現パターンを抽出し、一覧表にした (表 2)。「雪」と「桜」の共出現関係を比較してみると古今集から拾遺集までは「雪」を含む共出現パターンが多く、いずれも上位を占めている。金葉集になるとこの立場は逆転し、「桜」を含む共出現パターンが上位を占めるようになる。ただし「吉野」の歌そのものが少ないことや「雪→桜」の推移を対比して分析できるような「雪」あるいは他の語が見られないため、この時代に「桜の吉野」が確立していたかどうかは断言できない。新古今ではまだ「雪」は見られはするものの「吉野山／桜か枝に／雪ちりて／花をそけなる／年にもあるかな (新古今集 79 番：西行)」や「みよし野は／山もかすみて／白雪の／ふりにし里に／春はきにけり (新古

[11] 提喩法 (synecdoche) 上位概念で下位概念を表したり、逆に下位概念で上位概念に置き換えたりする比喩の一。

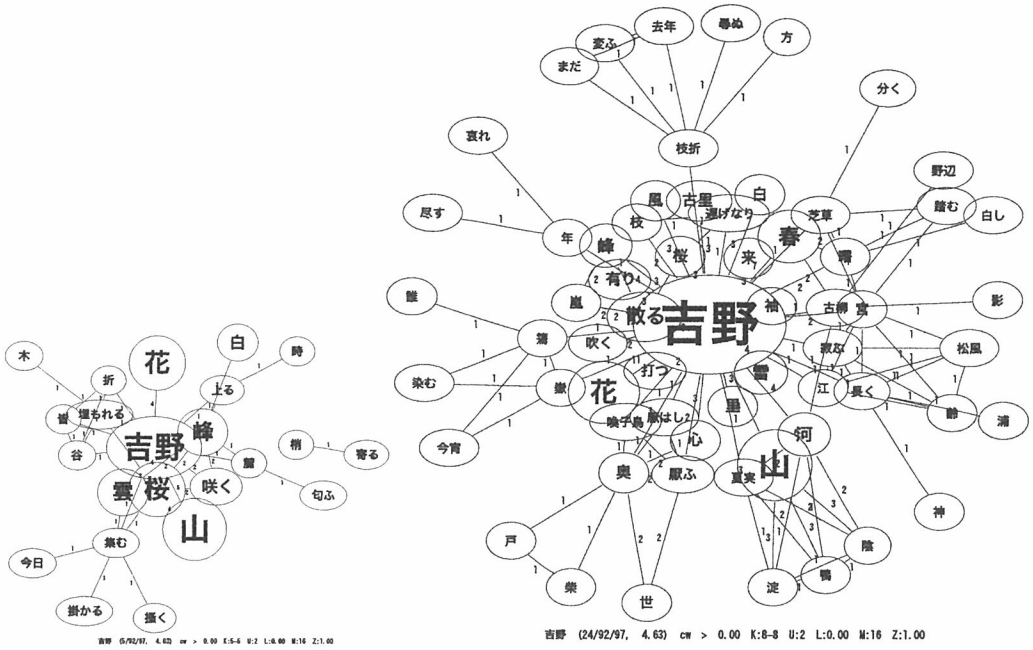

図 4: 金葉集 (左) と新古今集 (右) の「吉野」: エッジの数字は共出現の頻度。ノードの大きさは各モデルにおけるその語の相対的頻度を示す。どちらも同じ z 値 1 以上を出力。

今集 1 番：良経)」のように桜と一緒に詠まれたものや「春望」をテーマにしたものになり、「春の吉野」へ変遷している様子がうかがえる。以上、重要度の観点から共出現パターンを通して分析すると古今集での「吉野の雪」は、金葉集で変化を見せ、新古今集では「花」との関係に変化していったことがわかる[12]。

4 おわりに

本研究では「吉野」を題材に歌ことばのシソーラスの開発とモデル化を通して、古今集 (905 年) から新古今集 (1205 年) まで 300 年の歌ことばの変遷について分析した。

和歌は古代より伝わる日本の文芸であるが、そのコンパクトに納められたメッセージは文芸研究としてだけでなく、意味の圧縮、比喩の生成など言語の意味の諸相が歴史的に観察できる貴重な資料である。一つ一つの実例を示す研究方法では歌そのものを取り扱うため、意味や文脈に基づいた分析は可能であるが、複数例からなる特性、一般性、傾向、成分、親疎関係、依存構造など、科学的分析で求められる、いわゆる「体系」が示しにくい。一方、語彙分析で用いられる頻度や使用率を集計した語彙リストや一覧表は語彙の成分を全体的に把握することはできるが、その文脈を確認するには、再び実例にもどらなくてはならない「もどかしさ」がある。本研究の共出現パターンによるモデリングでは、上記問題点を解決し、体系的な語の布置を示しつつ、分析の対象とする単語だけでなく、具体的な用法ならびに実際の和歌に存在する文脈・周辺語彙も含めて検討することができた。

参考文献

[1] 浅見徹: 八代集における季節, 国語語彙史の研究 (国語語彙史研究会 (編)), Vol. 7, 和泉書院, pp. 111–31 (1986).

[12] 詠み手の活躍した時代に和歌を分類した上で、分析し直すとまた興味深い結果が見られると思われる。

表 2: 各集の「吉野」のモデルから抽出した雪あるいは桜の共出現パターン：八代集 (長歌のぞく 9484 首) 中「吉野」は $idf = 4.63$、出現和歌数=92、頻度=97。歌集名右の括弧数字は、その歌集での出現頻度。

	t_1–t_2	cw	z	ctf	$idf(t_1)$	$idf(t_2)$
古今集 (24)	雪–吉野	86.06	3.33	10	3.18	4.63
	雪–降る	65.15	1.76	5	3.18	3.26
	桜–辺	64.32	1.70	2	3.43	4.69
	雪–寒し	63.36	1.62	2	3.18	4.92
	雪–辺	61.87	1.51	2	3.18	4.69
	雪–白	60.36	1.40	4	3.18	3.18
	雪–古里	55.34	1.02	2	3.18	4.37
後撰集 (11)	雪–吉野	54.69	1.33	3	3.18	4.63
	雪–降る	52.40	1.12	3	3.18	3.26
	雪–崩る	51.40	1.03	1	3.18	8.06
	桜–吉野	51.28	1.02	2	3.43	4.63
拾遺集 (15)	雪–吉野	80.25	3.74	8	3.18	4.63
	雪–消ゆ	55.90	1.54	2	3.18	3.83
	雪–山	54.92	1.46	8	3.18	2.08
	雪–峰	54.35	1.40	2	3.18	3.95
	雪–宿	52.42	1.23	2	3.18	3.37
	雪–古道	50.48	1.05	1	3.18	7.77
後拾遺集 (3)	N/A					
金葉集 (5)	桜–吉野	72.27	3.34	4	3.43	4.63
	桜–峰	52.17	1.44	2	3.43	3.95
	桜–咲く	51.68	1.40	2	3.43	3.71
	桜–雲	51.00	1.33	2	3.43	3.43
	桜–山	49.48	1.19	4	3.43	2.08
	桜–集む	48.33	1.08	1	3.43	6.59
	桜–埋もれる	47.56	1.01	1	3.43	6.38
詞花集 (6)	N/A					
千載集 (9)	N/A					
新古今集 (24)	桜–吉野	63.56	1.64	3	3.43	4.63
	桜–散る	62.38	1.55	3	3.43	3.14
	雪–吉野	62.18	1.53	4	3.18	4.63
	桜–遅げなり	56.96	1.14	1	3.43	9.16

[2] 木村雅則: 平安勅撰和歌集の語彙—マクロとミクロの観点から—, 国語語彙史の研究 (国語語彙史研究会 (編)), Vol. 14, 和泉書院, pp. 25–47 (1994).

[3] 西端幸雄: 語彙史の立場から見た『拾遺和歌集』—使用語句の性格を統計的に見る—, 国語語彙史の研究 (国語語彙史研究会 (編)), Vol. 14, 和泉書院, pp. 318–303 (1994).

[4] 西端幸雄, 藤田久, 成田徹: パーソナルコンピュータ語彙索引自動作成の試み, 和泉書院, 大阪 (1989).

[5] 久保田淳 (編): 八代集総索引, 岩波書店 (1995).

[6] 山元啓史: 古今集データベースによる歌語の視覚化, 人文科学とデータベース、第 11 回シンポジウム, 人文科学とデータベース協議会, 大阪, pp. 81–8 (2005).

[7] 山元啓史: 歌ことばの可視化とコノテーションの抽出—グラフによる共出現パターンの作り方—, じんもんこん 2006, 人文科学とコンピュータシンポジウム, No. 17, pp. 21–28 (2006).

[8] 片桐洋一: 歌枕歌ことば辞典, 角川小辞典 35, 角川書店, 東京 (1983).

[9] 小林一彦: 吉野山：歌語・歌枕事典, 国文学解釈と教材の研究, Vol. 34, No. 13, p. 125 (1989).

[10] 上野理: 後拾遺集前後, 笠間書店 (1976).

[11] 川村晃生: 摂関期和歌史の研究, 三弥井書店 (1991).

[12] 山元啓史: 和歌のための品詞タグづけシステム, 日本語の研究, Vol. 3, No. 3, pp. 33–39 (2007).

[13] 中野洋: 分類語彙表／フロッピー版, 大日本図書, 東京 (1994).

[14] 水谷静夫: 語彙, 朝倉日本語新講座 2, 朝倉書店, 1 edition (1983).

[15] 石井正彦: 使用頻度“1”の語と文章－高校『物理』教科書を例に－, 国立国語研究所研究報告集, Vol. 17, 秀英出版, 東京, pp. 23–55 (1996).

[16] Robertson, S.: Understanding inverse document frequency: on theoretical arguments for IDF, *Journal of Documentation*, Vol. 60, pp. 503–520 (2004).

[17] Rocchio, J. J.: The SMART Retrieval System: Experiments in Automatic Document Processing, *Relevance feedback in information retrieval* (Salton, T. G.(ed.)), Prentice-Hall, Englewood Cliff, NJ, 1 edition, pp. 313–323 (1971).

[18] 中野洋: いわゆるＬ字型分布からはずれる語彙量の分布について, 計量国語学, No. 76, pp. 25–31 (1976).

[19] 水谷静夫: 短い作品の語彙の量的構造 昭和初期流行歌の調査から 1 , 計量国語学, No. 72, pp. 1–12 (1975).

[20] Manning, C. D. and Schütze, H.: *Foundation of statistical natural language processing*, The MIT press, Cambridge, Massachusetts (1999).