八代集用語のモデリングシステム

山元啓史

東京工業大学大学院社会理工学研究科

要旨

本研究では、八代集（約905年〜1205年）の和歌（約9500首）をデータベースとした歌ことばのモデリングシステムを開発した。従来のシステムは単語1つをキーとして検索、作図するものであったが、実際に研究で必要となる場面は、複数のネットワークモデルを比較し、語彙環境として互いにどのような布置にあるのかを調査したい場面であった。そこで、同時に2つのネットワーク図を生成し、両者を比較することによって、相対的特徴（relative salience）を分析するシステムを開発した。その結果、1語によるモデルでは、捉えにくかった相対的特徴は、2つのネットワークを統合したモデルを通して明示的に把握できるようになった。

The Modelling System of Japanese Poetic Vocabulary for the *Hachidaishū* (ca. 905–1205)

Hilofumi Yamamoto

Graduate School of Decision Science and Technology, Tokyo Institute of Technology

Abstract

This paper addresses the development of the visualization system of vocabulary construction for the *Hachidaishū* (ca. 905–1205), which comprises approximately 9,500 classical Japanese poems. The existing system has a thesaurus of poetic vocabulary as a database with which one keyword per model could be searched. Although sufficient, the capability of searching for two keywords at the same time rather than only one is deemed more effective. Therefore, an improved feature of the new system in this project is designed to handle two keywords at a time, and to be able to generate a combined network model by two keywords. As the result of the development of the new system, the differences between two keywords can be clearly observed, and the specific features of two models can be defined more easily as a relative salience.

1 はじめに

「八代集の用語モデリングシステム」は歌ことばの特徴をネットワーク (グラフ理論) で可視化するものである。和歌一首に出現する任意の2語の組合せをひとつのパターン (共出現パターン) として、すべてのパターンを生成し、その頻度と各パターンの情報量に基づく計算により、語と語の依存関係を描く。これらのパターンは数理表現であるので、語と語の依存関係（結びつく語の種類、量）が観察できるだけでなく、指定した条件にしたがい、パターンの分割、差分、統合などの操作が可能である。この

ツールを用い、八代集（約905年〜1205年）のおよそ300年間の語彙の変遷を歌集別に分析している。

本システムは従来1つの単語をキーとして検索、作図するものであった。しかし、実際に必要となる場面は、任意の2語がどのような布置であるのかを相対的に見る場面であった。具体的には、山元(2006)では一般名詞の「鶯」と「時鳥」について、山元(2005)では歌枕の「立田」と「吉野」について、それぞれのモデルから、差分をとることにより、相対的な違いを考察した。

本来、語の意味は単独の語で確定するのではなく、周辺の語群（文脈）によって相対的に識別され、その語の意味や機能がはっきりしてくる。そこで、任意2語の相対的特徴（relative salience）が簡易に出力できる機能を実装した。本稿ではその機能の実装について報告する。

2 仕様

2.1 内部データ

内部データとして、八代集の全和歌（約9500首）を収録したシソーラスを実装する。これは、和歌用形態素解析器(山元, 2007, 自作)を用いて、単語分割し、国立国語研究所の分類語彙表準拠の管理番号を付記したものである(山元, 2009a)。和歌の表記は現代語の表記とは異なり、さまざまに表記される[1]。その違いをシソーラス・コードを利用して調整する。

2.2 データの生成

ネットワーク1つにつき、1つの共出現テーブルを生成する。共出現テーブルとは、1首に含まれる語を2語ずつとりだして作られた単語対の一覧表である。この対を共出現パターンと呼び、その一覧表を共出現テーブルと呼ぶ。共出現パターンは、対の各々の単語をノードとして、また、それらを結ぶ線をエッジとして、グラフが表現できる。共出現パターンの数は検索キーで抽出された歌の数によって異なるが、数百〜数千になる。これらすべてのパターンをネットワーク図として描画すると真っ黒な塊になってしまうので、重みづけ計算をパターン毎に行い、重要なパターンから閾値までを順に出力する。

すべての単語の *idf* (Robertson, 2004; Rocchio, 1971)をあらかじめ計算しておく[2]。次に共出現パターンを生成し、先程の *idf* 値とパターンの出現頻度を使って、各パターンの重みを計算する。共出現パターンは単なる2語組のリストではあるが、共出現パターンで描画されたグラフには、もとの文（和歌）にある文脈が再現されることがわかっている(山元, 2005)。

本研究では共出現パターンに、重みづけ計算を導入するために以下のような計算式を用いた。テキスト群(d)において任意の1語(t)が特徴的であるかを評価する式 *tf-idf* (1) (Manning and Schütze, 1999)を拡張し、任意の2語のパターン(t_1, t_2)がどの程度特徴的であるのかを評価する式(2)を用い、パターンの重み(cw)を計算する。

$$w(t,d) = (1+\log\ tf(t,d)) \cdot idf(t) \quad (1)$$

$$cw(t_1,t_2,d) = (1+\log\ ctf(t_1,t_2,d)) \cdot cidf(t_1,t_2) \quad (2)$$

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

ただし、(2)の前半はt_1とt_2の2語が共出現した時のテキストの数。(2)の後半$cidf(t_1,t_2)$は、(1)の$idf(t)$を拡張し、2語の *idf* 値の幾何平均(3)としたものである。

異なる2つのネットワークを比較するためには、ネットワーク図に用いるノード・エッジの数をある基準によって、一意に選ばれなければならない。そこで、cw値を相互に比較できるように、一旦正規化を行い、標準得点1σ以上の共出現パターンを出力することにした。

2.3 出力設計

出力は、検索語の分布表、共出現テーブル(html, text)、ネットワーク図(svg, png)、ノード当該の歌の一覧、の4種類である。

検索語の分布表では、語を検索した場合、八代集の全体に渡ってどのように分布しているかを一覧にする。検索語自体（たとえば「ほととぎす」）、検索語を一部含む語（山ほととぎす、山時鳥）、検索語と表記の異なる語（ほとゝきす、ほとときす、ほととぎす、時鳥）なども総合して一瞥できるようにする。ネットワーク図を描画する際には、すべて語形を作図の対象にしたり、特定の語形（たとえば「山時鳥」）のみを対象にしたりできるようにする。それらを個別に取り出して、作図できるようにする。

[1] データベース開発では、注釈書に見られる伝本の表記をできる限り拾い集めたが、十分であるかどうかは、不明である。

[2] *idf* 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、$idf(t) = \log N/df(t)$で定義される。ただし、Nはすべての資料の数、$df(t)$は、語tの出現する資料の数である。情報量の一種である。

また、古今集から新古今集まで、歌集毎あるいはすべてを出力できるようにする。

共出現テーブルは、単語をキーとして八代集データベースを検索し、その結果得られた和歌を対象に、共出現パターンを生成し、一覧表にしたものである。本システムでは2つの用語（A, B）の表を生成する。この表に基づいてネットワーク図を作成する。

ネットワーク図は Graphviz/neato (`http://www.graphviz.org/`) で描く。デフォルトでブラウザに出力されるのは、svg である。svg に対応していないブラウザは png で出力される。ネットワーク図は単語を示すノード（大きさは単語の相対的出現頻度を表す）と単語と単語の関係を示すエッジ（傍らの数字は頻度共出現の頻度）からなる。2つの共出現テーブル（A, B）にあるデータを用いて $A \cup B$ を描く。$A \cap B$ に相当するノードは、グレーで塗りつぶし、2つネットワークで共有されていることを示す。

ネットワーク図を観察する際、あるノードがどのような和歌からきているのかを知りたくなる。その際は、ノードをマウスでクリックすると直接当該の和歌が出力できるとよい。png で出力された図の場合には、クリッカブルマップによって、svg で出力された図の場合には、svg の内部のノード毎に サーバ側で処理できるように、URL と当該の和歌の歌番号（新編国歌大観準拠）を埋め込み、処理を行う。

3 操作例

3.1 検索キーの入力

検索キーは、グラフモデルの中心となる語である。漢字、平仮名、あるいは語コード[3]のいずれでも可能である (図 1)。

3.2 条件の指定

2つのキーを入力し、OK ボタンを押すと、図 2 のように一覧表が出力される。一覧表には、該当する文字列とその八代集の各歌集における頻度と全体の頻度が示される。グラフモデルの作図に必要なパラメタの指定を行う。

[3]旧版分類語彙表国立国語研究所 (1994) をもとに歌ことばを体系的にシソーラスコードでデータベース化してある山元 (2009a)。たとえば「梅」は「BG-01-5520-20-04」と定義されている。このコードによって、語の分割単位、表記、分類カテゴリを選びだす。

1. 分析対象の用語の選択
   どの語のモデルを出力するか、一覧表左端のラジオボタンで選ぶ。

2. 対象とする歌集の選択
   古今集から拾遺集までを見たければ、From は古今集、To は拾遺集を選ぶ。古今集だけの場合には、From も To も古今集を選ぶ。

3. 集計のレベルの指定
   単語の表記を区別せずに集計する時には、level 16 を選ぶ。区別するときには、level 18 を選ぶ。16, 18 という数字は、シソーラスコードの桁数を意味する。

4. 共出現ウエイト計算方法の選択
   共出現パターンの重みづけ計算方法を指定する。現在はデフォルトは 7。12 は現在実験中の計算方法。

5. 語の単位 (Unit) のサイズの指定
   たとえば「吉野山」の場合は「吉野」と「山」の2語に分割して集計するのか、「吉野山」としてそのまま集計するか、を指定する。Unit 0 はそのまま、Unit 1 は地名に限り分割、それ以外はそのまま、Unit 2 はすべて最も短い単位に分割する。ネットワーク図を描く時には、できるだけ短い単位の方が、意味の似通ったノードを複数出力しないで済む。

6. Log-likelihood (Dunning, 1993)
   Dunning (1993) の2語の関連性検定を行う ($p < .05$, あるいは $p < .01$ レベル) か、行わないか、を指定する。

7. Z-score
   パターン重み付け得点 (cw) を標準得点に変換し、0 σ 以上、1 σ 以上、2 σ 以上の各場合で、パターンを出力する。

8. Co-occurrence Weight
   cw がある値以上になった場合のみを出力の対象とする。

9. CW Table Out
   共出現テーブルを出力する。ネットワーク図は一目でわかる便利さはあるが、図は結果が見せられているだけであり、どういう変数や要素により、その図が作られているのかが、見えにくい。計算の過程やすべてのノードに関わる基

[Make new] [Help] [Info] [Home]

Computer Modelling of Japanese Poetic Vocabulary

Key 1: 梅

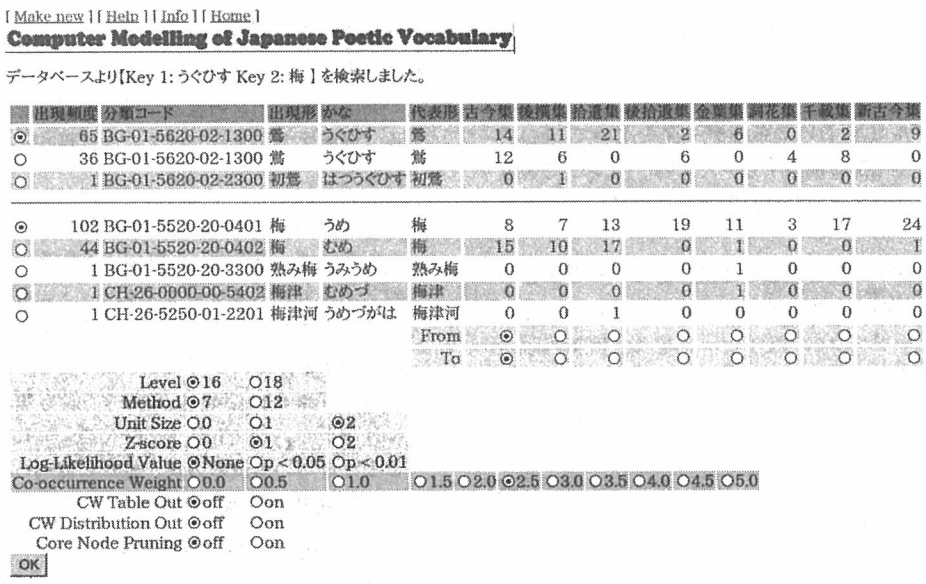
Key 2: 桜

OK

--

GPL Copyleft Hilofumi Yamamoto (yamagen[at]ryu.titech.ac.jp)

図 1: キー入力画面: 「梅」「桜」を入力して、OK ボタンを押す。



[Make new] [Help] [Info] [Home]

Computer Modelling of Japanese Poetic Vocabulary

データベースより[Key 1: うぐひす Key 2: 梅]を検索しました。

	出現頻度	分類コード	出現形	かな	代表形	古今集	後撰集	拾遺集	後拾遺集	金葉集	詞花集	千載集	新古今集
◉	65	BG-01-5620-02-1300	鶯	うぐひす	鶯	14	11	21	2	6	0	2	9
○	36	BG-01-5620-02-1300	鶯	うぐひす	鶯	12	6	0	6	0	4	8	0
○	1	BG-01-5620-02-2300	初鶯	はつうぐひす	初鶯	0	1	0	0	0	0	0	0
◉	102	BG-01-5520-20-0401	梅	うめ	梅	8	7	13	19	11	3	17	24
○	44	BG-01-5520-20-0402	梅	むめ	梅	15	10	17	0	1	0	0	1
○	1	BG-01-5520-20-3300	熟み梅	うみうめ	熟み梅	0	0	0	0	1	0	0	0
○	1	CH-26-0000-00-5402	梅津	むめづ	梅津	0	0	0	0	1	0	0	0
○	1	CH-26-5250-01-2201	梅津河	うめづがは	梅津河	0	0	1	0	0	0	0	0
					From	◉	○	○	○	○	○	○	○
					To	◉	○	○	○	○	○	○	○

Level	◉16	○18							
Method	◉7	○12							
Unit Size	○0	○1	◉2						
Z-score	○0	◉1	○2						
Log-Likelihood Value	◉None	Op < 0.05	Op < 0.01						
Co-occurrence Weight	○0.0	○0.5	○1.0	○1.5	○2.0	◉2.5	○3.0	○3.5	○4.0	○4.5	○5.0
CW Table Out	◉off	Oon							
CW Distribution Out	◉off	Oon							
Core Node Pruning	◉off	Oon							

OK

図 2: グラフモデル作図のための各種設定画面: キーにしたがい、データベースより該当の語句が八代集各歌集における頻度とともに出力される。

礎データがわかるように共出現テーブルを提供する。共出現テーブルの例を図 3 に示す。

10. CW Distribution Out

共出現テーブル内の *cw* 値あるいは *Z* 値の分布曲線を描画する。(現在、実装中)

11. Core Node Pruning

検索した用語 (キー) の *cw* が極めて大きいとき、グラフは真っ黒な塊になる[4]。これは、ほとんどのノードがキーのノードと関係を持つため、数多くのノード・エッジで隠れてしまって黒く塗りつぶされてしまうためである。これを「スポークエフェクト」と呼んでいる。この時は閾値を操作するのではなく、キーそのものを削除すると見えやすくする。これをプルーニング (刈り込み) という。(従来の 1 語によるネットワークシステムでは実装済みだが、本システムの場合は検討中)

以上により、ネットワーク図は svg あるいは png 形式で出力される。さらに、その図の任意のノードをクリックするとノードに関連する和歌が出力される。svg は XML によってマークアップされたテキストファイルである。利用者がブラウザで表示され

[4]この時、出力するパターンの数を *cw* の値を上げることで制限することもできるが、それは恣意的な操作である。

KEY CT BG-01-5620-02-13 鶯 26 101 4.53 0.00 101 K:1-1 L:0.00 U:2 M:7 Z:0.00

No.	cw	cfq	bgcode (t_1)	idf	fq	tfq	bgcode (t_2)	idf	fq	tfq	t_1—t_2	l2	Z
1	4.98	5	BG-01-4250-02-01	7.08	5	13	BG-02-3840-01-03	7.08	2	8	笠—縫ふ	106.16	6.51
2	4.04	3	BG-01-4250-02-01	7.08	5	13	BG-02-1570-07-02	6.16	1	20	笠—擦る	47.54	4.67
3	3.98	5	BG-01-4250-02-01	7.08	5	13	BG-01-5620-02-13	4.53	26	101	笠—鶯	53.03	4.56
4	3.82	5	BG-01-4250-02-01	7.08	5	13	BG-01-5520-20-04	4.16	6	146	笠—梅	72.96	4.23
5	3.80	3	BG-01-4250-02-01	7.08	5	13	BG-03-5020-03-06	5.44	1	41	笠—青い	47.54	4.19
6	3.79	3	BG-01-4250-02-01	7.08	5	13	BG-01-5520-26-01	5.42	1	42	笠—柳	47.54	4.17
7	3.54	3	BG-01-4200-05-01	4.74	1	82	BG-01-4250-02-01	7.08	5	13	糸—笠	47.54	3.70
8	3.42	16	BG-01-5620-02-13	4.53	26	101	BG-02-3030-03-01	3.09	16	426	鶯—鳴く	534.21	3.45
9	3.35	2	BG-01-4250-02-01	7.08	5	13	BG-02-1130-08-01	5.47	1	40	笠—押頭す	36.84	3.32
10	3.20	2	BG-01-4250-02-01	7.08	5	13	BG-02-5810-06-07	4.98	1	66	笠—老ゆ	36.84	3.02
11	3.20	6	BG-01-5520-20-04	4.16	6	146	BG-01-5620-02-13	4.53	26	101	梅—鶯	171.56	3.02
12	3.18	5	BG-01-4250-02-01	7.08	5	13	BG-02-3120-01-01	2.89	3	523	笠—言ふ	87.43	2.99
13	3.05	2	BG-01-5620-02-13	4.53	26	101	BG-02-3840-01-03	7.08	2	8	鶯—縫ふ	21.15	2.73
14	3.00	2	BG-01-4250-02-01	7.08	5	13	BG-02-1210-02-01	4.39	1	117	笠—隠る	34.07	2.64
15	2.92	2	BG-01-5520-20-04	4.16	6	146	BG-02-3840-01-03	7.08	2	8	梅—縫ふ	28.74	2.48
16	2.91	2	BG-01-4250-02-01	7.08	5	13	BG-02-1570-01-01	4.11	3	159	笠—折る	27.34	2.45
17	2.84	1	BG-01-1760-08-02	6.52	1	14	BG-01-4700-12-07	7.21	1	7	末—国	14.51	2.32
18	2.82	10	BG-01-5620-02-13	4.53	26	101	BG-02-1527-01-01	2.56	10	714	鶯—来	149.81	2.29
19	2.82	10	BG-01-1624-02-01	2.56	10	709	BG-01-5620-02-13	4.53	26	101	春—鶯	429.09	2.29
20	2.82	3	BG-01-5530-08-05	4.70	3	87	BG-01-5620-02-13	4.53	26	101	枝—鶯	95.82	2.29
21	2.80	16	BG-01-5530-12-01	2.08	16	1148	BG-01-5620-02-13	4.53	26	101	花—鶯	383.23	2.25
22	2.73	1	BG-02-1570-07-02	6.16	1	20	BG-02-3840-01-03	7.08	2	8	擦る—縫ふ	14.51	2.11
23	2.73	1	BG-01-3123-04-12	7.08	1	8	BG-02-1562-03-13	6.16	1	20	便る—頼ふ	21.79	2.11
24	2.73	1	BG-03-1940-01-03	6.76	1	11	BG-03-1940-02-01	6.45	1	15	疾し—遅し	19.01	2.11
25	2.70	5	BG-01-4250-02-01	7.08	5	13	BG-01-5530-12-01	2.08	16	1148	笠—花	64.13	2.04

図 3: 共出現テーブルの出力例: 検索キー「鶯」の場合の共出現テーブル（紙面の関係から上から 25 番までの一部）。1 番上にテーブル出力の時の条件が示される。キーのシソーラスコード「BG-01-5620-02-13」、キー「鶯」、このテーブルでの出現頻度「26」、当該歌集全体での出現頻度「101」、*idf* 値「4.53」、*Z* 値によって取り出された時の *cw* 値「2.50」 *idf* 値の元となった *N* の値「101」対象とする歌集の範囲「K:1-1（古今から古今まで）」Loglikelihood 値「L:0.00」、ユニット種別「U:2」、計算法「M:7」標準得点の閾値「Z:0.00」テーブルは、左より、共出現パターンの通し番号（No.）、共出現ウエイト（cw）、共出現パターンの頻度（cfq）、用語 1 のシソーラスコード（bgcode (t1)）、用語 1 の idf 値、抽出された用語 1 の頻度（fq）、全体の用語 1 の頻度（tfq）、用語 2 のシソーラスコード（bgcode (t2)）、用語 2 の idf 値、抽出された用語 2 の頻度（fq）、全体の用語 2 の頻度（tfq）、共出現パターン（t1–t2）、実験中の計算メソッド（13）、共出現パターンの標準得点（Z）

た svg ファイルをそのまま自分のコンピュータにダウンロードすれば、サイズを変更しても劣化のない画像が得られる。ファイルには各パターンのリスト、重要度の値など、実際に作図に用いられたデータが記載されており、それらデータを利用者が各自の論文に引用することができる。また、隠れたノード・エッジも svg には記録されているので、svg を編集できるソフト（たとえば、inkskape）を用いれば、色を加えたり、ノードの位置を修正したり、不要な要素を取り除いたりすることができる。

3.3 ネットワーク図の出力例

「梅」と「桜」の２語をキーとするモデル（図４上）と「梅」と「鶯」の２語をキーとするモデル（図４下）の２つの図をを出力した。出力の条件ともに、重要度 2.5 以上、Unit 2、Loglikelihood 指定なしで、古今集のテキストが対象である。ノードの大きさは用いられた全単語の頻度を相対的に示している。エッジに添えられた小さな数字はパターンの頻度を示している。共有するノードはグレーで塗りつぶされる。

図４（上）を見ると「梅」「桜」が共有するノードは「花」と「折る」だけである[5]。「梅」も「桜」もともに「花」であり、それらの「花」はともに「折って」賞 (め) でる以外には、共通する観念はあまり見られないようだ。違いとしては「香る梅」に対して、「散る桜」が図から見て取れる。これにより２語の相対的特徴が把握できた。

一方、図４下の「梅」「鶯」の図を見ると共有するノードの数は多く、２語の関係は 密接であることがわかる。上の図と見比べると「桜」と「梅」のノードは共有していないことから、同じ歌で同時に詠まれることがないのがわかる。下の図では、「鶯」「梅」ともにグレーで示され、同じ歌で詠まれることがわかる。ただし、「香」のノードが共有されていないからといって、「鶯」と「香」を詠んだ歌がないわけではない。たとえば、古今集 13 番歌紀友則の「花のかを／風のたよりに／たくへてそ／鶯さそふ／しるへにはやる」という歌がある。あくまでもある数理的基準で一意に選び出された操作によって、このようなことが起こっているのである。ゆえに、実際の和歌を常に見ながら、よりよい出力方法について検討していく必要がある。

「香る梅」の実例がどのような歌であるのかを見るために、図４のネットワークモデルで「香」をクリックする。すべてのノードにもとの和歌の管理番号が埋め込んであるので、図５のように、そのノードに該当する和歌が出力される。この和歌のテキストには国文学研究資料館開発の「二十一代集」(中村他, 1999) を用いた[6]。

4 おわりに

従来のシステムに加えて任意２語の検索と共有ノードの作図を可能にするシステムを開発した。語の相対的特徴を１つの図によって示すことができた。今後はさまざまな２語の違いを相対的特徴の点から明らかにしていくことができよう。

今回のシステムのさらなる改良も計画している。相対的特徴は「２語」いうように語を軸とした比較で捉えるだけでなく、任意の２つの視点というような、より抽象的な視点で、処理できるシステムに改良していきたい。

たとえば、同じ語であっても詠み手によって使い方が異なることがある。「桜」は大半が「うつせみの世にも似たるか桜花咲くと見し間にかつ散りにけり」（古今集 73 番歌）というように散り惜しむ歌として詠まれている (片桐, 1983)。ただし、同じ「桜」であっても、「西行桜」はどのような点で、それまでの「桜」と異なるのであろうか。室町時代に世阿弥作の能楽作品にまで発展した西行の「桜」にはそれまでの「桜」とは異なる要素があるのではないかと期待する。この場合、作家を軸とした２点、つまり、西行の「桜」と他の「桜」を比較することにより、その相対的特徴が観察できよう。しかし、現在のシステムには作家をキーに抽出する機能はないので、これはできない。また、実際には「桜」を詠んではいても、テキストに出現する語は「花」である場合も多い。その場合は単純にシソーラスのコードを用いて上位のコードを一括して検索すればよい？というわけではない。「桜」以外の花が含まれてしまうからだ。今後の課題である。

現在のシステムは八代集のみを扱っているが、将来的には、二十一代集の歌すべてについて同様の処理ができるようにもしていきたい。これは現在の八代集対応和歌形態素解析辞書を拡張し、二十一代集

[5]現実の歌で共有する単語は数多くあるが、ある基準から見て、その重要度にしたがい、結果的に選ばれたノードの中では「２語だけ」という意味である。

[6]正保版本を底本としつつ、研究に重要な異本の情報が詳細に加えられている。使用については国文学研究資料館の許諾を得ている。

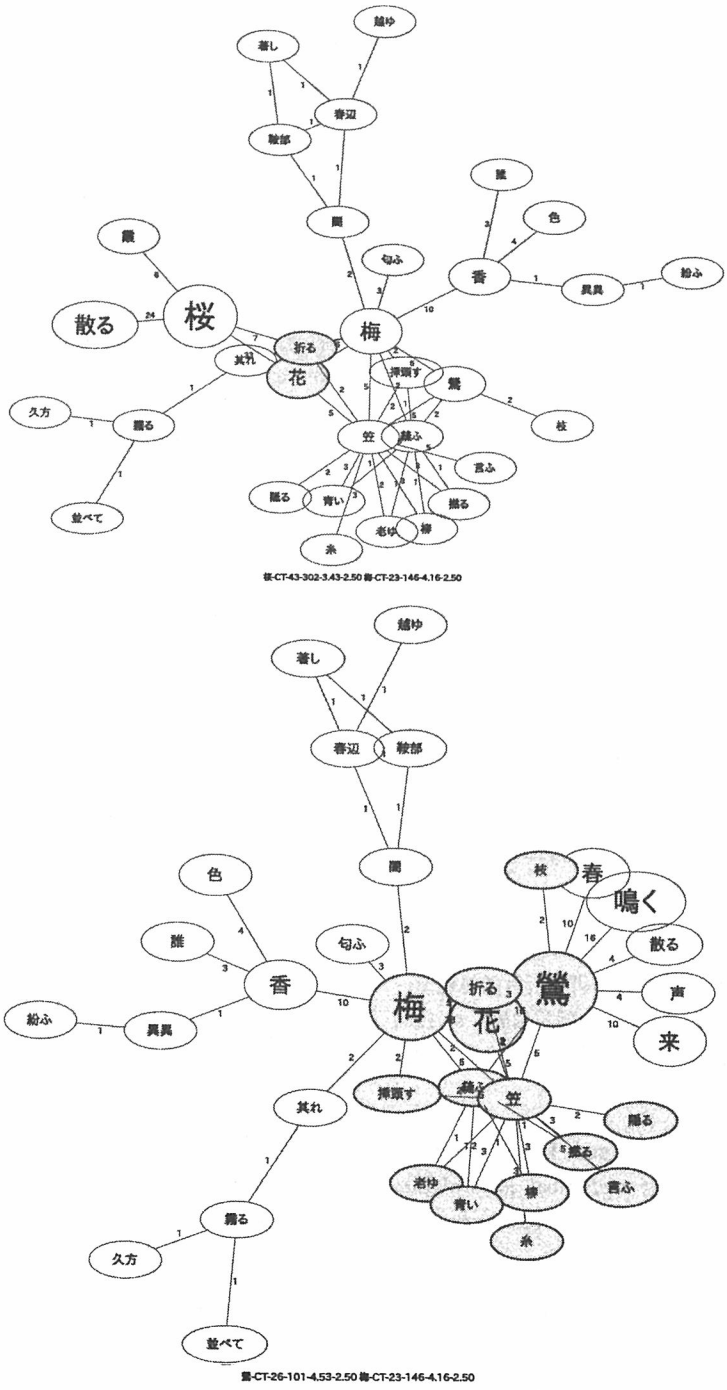


図 4: グラフモデルの出力「梅／桜」（上）と「梅／鶯」（下）：「梅」と「桜」で共有するノードはあまり多くなく、和歌での２語の関係は同じ花であっても、異質なものであることがわかる。一方「梅」と「鶯」（下）を見ると共有するノードの数は多く、２語の関係は密接であることがわかる。

図 5: 和歌一覧の出力: 図 4 の「香」ノードをクリックすると当該の和歌が出力される。10 の和歌が出力されている。

対応にする研究が進められている (山元, 2009b)。近いうちに実現したいと考えている。

参考文献

Dunning, Ted (1993) "Accurate Methods for the Statistics of Surprise and Coincidence", *Computational Linguistics*, Vol. 19, No. 1, pp. 61–74.

片桐洋一 (1983) 『歌枕歌ことば辞典』, 第 35 巻, 角川小辞典, 角川書店, 東京.

国立国語研究所 (編) (1994) 『分類語彙表／フロッピー版』, 第 5 巻, 国立国語研究所言語処理データ集, 大日本図書, 東京. 『分類語彙表』は 1964 年に国立国語研究所資料集 6 林大担当として刊行された。.

Manning, Christopher D. and Hinrich Schütze (1999) *Foundation of statistical natural language processing*, Cambridge, Massachusetts: The MIT press.

中村康夫・立川美彦・杉田まゆ子 (1999) 『国文学研究資料館データベース古典コレクション 『二十一代集』(正保版本) CD-ROM』, 岩波書店, 東京.

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

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

山元啓史 (2005) 「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, 81–8 頁.

——— (2006) 「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』, 第 2006 巻, 第 17 号, 21–28 頁.

——— (2007) 「和歌のための品詞タグづけシステム」, 『日本語の研究』, 第 3 巻, 第 3 号, 33–39 頁.

——— (2009a) 「分類コードつき八代集用語のシソーラス」, 『日本語の研究』, 第 5 巻, 第 1 号, 46–52 頁.

——— (2009b) 「和歌解析用 MeCab 辞書の開発—八代集解析済みコーパスによる学習—」, 『第 15 回公開シンポジウム人文科学とデータベース発表論文集』, 31–36 頁.