

## 八代集シソーラスによる歌ことばの分析 —シソーラスの開発と可視化システム—

山元 啓史

オーストラリア国立大学

### 要 旨

本研究では、八代集 (ca. 905 年～1205 年) の和歌 (約 9500 首) を単位分割し、各単位に国立国語研究所の分類語彙表準拠の分類コードを添付したシソーラスを開発した。古語には現代語でいう標準表記というものがなく、さまざまに表記されるため、電子化テキストそのままの形では、語彙量を求めたり、単語を意味毎に分類したりするのは難しい。これらの処理を簡単にするためには、各単位語について表記を統一した上で、意味体系で整理したメタコードをつける必要がある。本稿では、上記の処理を施した八代集シソーラスの開発とその試用実験、さらにその応用として語彙構造可視化システムの試作について述べる。本シソーラスにより、従来、単位分割や行検索だけでは困難であった 1. 異形同語でも同じ語として集計、2. カテゴリによる集計、3. 存在しないカテゴリの指摘、などが可能になった。

## The Development of the *Hachidaishū* (ca. 905–1205) Thesaurus with the Classification Codes Based on Semantic Principles, and the Visualization System of Poetic Vocabulary Construction

Hilofumi Yamamoto

The Australian National University

### Abstract

Since classical Japanese words are written in various characters, it is difficult to analyze by computer. To solve this problem, it is necessary to define the semantic category of vocabulary and unify the orthography of each word. This study addresses 1) the development of the thesaurus for the *Hachidaishū* (ca. 905–1205), which comprises approximately 9,500 classical Japanese poems; and 2) the development of the visualization system of vocabulary construction as its application. Each poem is divided into tokens. A thesaurus code is attached to each token according to the *Bunrui Goi Hyo* (Japanese thesaurus based on the semantic principles) developed by the National Institute of Japanese Language. This thesaurus allows researchers to 1) analyse words as the same semantic elements although they are expressed in different orthographies, 2) calculate the number of words by each semantic category, and, 3) point out missing semantic categories in target texts, which are difficult by using only a morphological parser or string searches.

## 1 はじめに

古語で記された資料のうち、和歌は文学作品であると同時に言語や文化を研究する上でも貴重な資料である。なかでも八代集は、古今集(905年)から新古今集(1205年)までの300年間に成立した8つの勅撰集で、ほぼ9,500首の和歌が収められており、言語の変遷を通時的に分析するには非常に便利な資料である。従来にも八代集までを一区切りとして、歌ことば、歌の題材、それらの成立および展開など、さまざまな視点から数多くの研究が行われてきている(辻, 1998, p. 226)。

しかしながら、古語には現代語でいう標準表記というものはなく、さまざまに表記されるため、計算機で処理するには困難が伴う。古語の文献を電子化/データベース化するプロジェクトは数々あるが、文字列を分割し、そのまま計量分析するのは難しい<sup>1</sup>。

歌ことばの体系的な分析を本格的に試みるには、歌が示す意味や文化的色彩、歌題、掛詞、詞書、技巧、語感などさまざまな要素をデータベース化する必要がある。本研究では、その手始めとして、語の単位の認定と意味による分類コードを付したソーラスを開発することにした。

ソーラスは用語を体系的に管理したリストで、従来は類語検索に用いられていたが、次第に情報管理や検索の利便さから研究に多く用いられるようになってきた。中でも国立国語研究所の「分類語彙表」はさまざまな目的で利用されている(宮島・小沼, 1992)<sup>2</sup>。中野(1969, p.51)は分類語彙表を用いれば「語を体系化する方法と電子計算機に入力可能な数値を得ることができる。言語処理への意味の導入が重要な問題になっている現在、我々はより充実した意味情報を得なければならない」と述べ、分類語彙表による意味の分析の可能性を示唆した。田島(1995, p.9)は「意味をコード化し、数値化することができれば、恐らく語彙分析に効力を発揮するであろう」と述べ、自ら分類語彙表を利用した語彙構造の分析を行っている。同時に「語に意味コードをつける作業の困難さがあり、これを解決する必要がある」とコードづけ作業の困難さも指摘している。西端他(1989)も「各種分類コードの付加作業を行ったが、予想以上に作業に手間取った」とコードづけの作業の困難さを指摘している。本稿では、分類コードつき八代集ソーラスの開発、試用結果、そしてその応用としての語彙構造の可視化システムの試作について報告する。

## 2 データの作成方法

国文学研究資料館作成による二十一代集データベース<sup>3</sup>を主に利用し、その他さまざまな表記でも文字列の照合ができるよう、数多くの資料を参照し、表記の実例を集めた<sup>4</sup>。八代集の和歌テキストを品詞タグづけシステム kh(山元, 2007)で単位分割し、品詞情報を加え、その出力に t2c(山元, 2009 印刷中)(Token to Codes: 漢字見出し語、よみ、品詞の3つの情報を入力すると分類語彙表のコードを返すプログラム)を用いて、分類コードを加えた。kh, t2cの各処理で未登録語が見つかったら、辞書にその語を登録し、再び出力の確認を行った。この作業を未登録語がなくなるまで続けた。

t2cは、一般語(BG)、地名(CH)、人名(PN)の3種類の分類コード辞書を持っている。一般語(BG)には旧版分類語彙表(中野他, 1994)の索引データを利用した。旧版分類語彙表では個別の語に番号は付与されていないが、ここでは個別の語にも番号を付与した。任意2語の分類コードの遠近はその2語の意味の遠近とは関係のないものとした。分類コードは18桁で示し、その下4桁のうち、異形同語は下2桁に異なるコード、異語は上2桁に異なるコードを与えた。すなわち、先頭の16桁を有効桁数にすれば、異形同語も同じ語として処理できるようにした。

分類語彙表は現代語を前提に開発されているので、厳密には古代語のソーラスではない。平安末期と現代とではほぼ同義かつ同形態の動詞が存在する場合は問題が小さいが、動詞も転成品詞も現代に全く生きていない場合や同形態であっても語義が大きく変化している場合は、問題が大きく<sup>5</sup>、注意を要する。本研究

<sup>1</sup>たとえば、「家苞」は「いゑつと」「いゑづと」「いへつと」「いへづと」「家つと」「家苞」などのように表記されるが、これをもひとつの語としてコンピュータで処理するのは難しい。

<sup>2</sup>宮島・小沼(1992)発表時点で136の論文で利用されている。

<sup>3</sup>底本は国文学研究資料館蔵「正保版本二十一代集」。

<sup>4</sup>現段階では、国文学研究資料館二十一代集に加えて、岩波日本古典文学大系本、角川新編国歌大観収録の表記にも対応できるようにした。

<sup>5</sup>たとえば、(犬飼, 1988, p.38-39)は「あたふ」には「与える」の2.377を与え、動詞「くらがる」は現代に生きていないので、名詞「くらがり」との関連から「光」類の2.501を与える方法を示している。

では広辞苑と各種古語辞書を参照し、適切な番号が与えられるようにしたが、全く問題がないわけではない。辞書登録の結果、一般語 48,732、地名 1,408、人名 49 の計 50,189 レコードになった。以上で八代集に見られる用語は網羅されたが、同形異語<sup>6</sup>の問題があるため、t2c で処理しただけでは、複数のコードが出力される。厳格な意味の区別に対応させるには t2c の出力から文脈にふさわしいコードを選び出す（あるいは不要なコードを削除する）必要がある<sup>7</sup>。

表 1: 八代集シソーラス (古今集 2 番歌、紀貫之): 先頭 2 桁は歌集の番号、次 6 桁は歌番号、次 4 桁は語番号。A00 以外に A01, A02... が見られるのは、その語が多義で複数のコードを取ることを示す。以降順に、分類コード、漢字、よみ、代表形、ChaSen の辞書で用いられている品詞の分類番号。

|                |     |                    |    |     |    |    |
|----------------|-----|--------------------|----|-----|----|----|
| 01:000002:0001 | A00 | BG-01-4240-01-0100 | 袖  | そで  | 袖  | 02 |
| 01:000002:0002 | A00 | BG-02-5130-01-2100 | 潰つ | ひつ  | 潰つ | 47 |
| 01:000002:0003 | A00 | BG-08-0064-16-0100 | て  | て   | て  | 64 |
| 01:000002:0004 | A00 | BG-02-1515-08-0105 | 掬ぶ | むすぶ | 掬ぶ | 47 |
| 01:000002:0005 | A00 | BG-09-0010-04-0200 | き  | き   | き  | 74 |
| 01:000002:0006 | A00 | BG-01-5130-03-0201 | 水  | みづ  | 水  | 02 |
| 01:000002:0007 | A00 | BG-08-0061-07-0100 | の  | の   | の  | 61 |
| 01:000002:0008 | A00 | BG-02-5160-01-0101 | 凍る | こほる | 凍る | 47 |
| 01:000002:0009 | A00 | BG-09-0010-03-0300 | り  | り   | り  | 74 |
| 01:000002:0010 | A00 | BG-08-0061-10-0100 | を  | を   | を  | 61 |
| 01:000002:0011 | A00 | BG-01-1624-02-0100 | 春  | はる  | 春  | 02 |
| 01:000002:0012 | A00 | BG-02-1513-01-0100 | 立つ | たつ  | 立つ | 47 |
| 01:000002:0012 | A10 | BG-02-1521-06-0200 | 立つ | たつ  | 立つ | 47 |
| 01:000002:0012 | A20 | BG-02-3330-11-0200 | 立つ | たつ  | 立つ | 47 |
| 01:000002:0012 | A30 | BG-02-3391-02-1100 | 立つ | たつ  | 立つ | 47 |
| 01:000002:0013 | A00 | BG-01-1641-02-1100 | 今日 | けふ  | 今日 | 02 |
| 01:000002:0014 | A00 | BG-08-0061-07-0100 | の  | の   | の  | 61 |
| 01:000002:0015 | A00 | BG-01-5151-01-0100 | 風  | かぜ  | 風  | 02 |
| 01:000002:0016 | A00 | BG-08-0065-14-0100 | や  | や   | や  | 65 |
| 01:000002:0017 | A00 | BG-02-1550-05-0200 | 解く | とく  | 解く | 47 |
| 01:000002:0017 | A10 | BG-02-3060-09-0400 | 解く | とく  | 解く | 47 |
| 01:000002:0018 | A00 | BG-09-0010-02-0100 | らむ | らむ  | らむ | 74 |

表 1 はシソーラスの例である。「立つ」と「解く」は多義なので、複数のコードが出力されている。「解く」の例でいうと、BG-02-1550-05 (合い・組み・解け:「紐を解く」)、BG-02-3060-09 (思考・認識・知解:「問題を解く」) の 2 つが出力されている。どれがふさわしいかは、文脈を見て一つ一つ判断して選ばなければならないが、簡単に選べない場合も少なからずある。表記の統一を行うだけなら、最初に見られる「立つ (BG-02-1513-01)」や「解く (BG-02-1550-05)」を選び、それらを分析の対象とすればよい。

### 3 検索例

開発したシソーラスの利用により、1) 異形同語を同じ語として集計、2) カテゴリで集計、3) 存在しないカテゴリの指摘、などが可能となる。特に 3) は体系を持ったリストと照合しない限り指摘できない。

まず、異形同語の集計であるが、「立田」を例に説明すれば、「立田」の分類コード CH-29-0000-00-1800 の上 16 桁で検索すると「立田」「竜田」「龍田」の 3 種類が抽出され、それぞれの頻度、立田 (54)、竜田 (5)、龍田 (4) とそれら合計 (63) が得られた。

つぎに概念やカテゴリ集計であるが、BG-01-5520 は植物を示すコードで、これで検索すると、「松」をはじめ、203 種類の植物名とその頻度が得られた。

最後に欠落している概念の検索例を示す。紀貫之は古今集仮名序に「やまと歌は人の心を種としてよづの言の葉とぞなれりける世の中にある人、事、業しげきものなれば、心に思ふことを見るもの聞くものにつけて、言ひいだせるなり」と、人間のさまざまな心模様、喜怒哀楽が和歌には表現されているのだと述べている。ところが、筆者が知る限り (あるいは開発作業を通してデータを見る限り)、「あの時喰ったあれは本当にうまかった」「あれをもう一度食べたい」など、「食」の歌はほとんど見られない。久保田 (2003, p. 7) は「概して王朝文学では飲食という行為は描写の対象として軽視されている。和歌文学に至っては、食

<sup>6</sup>語の形は同じで、文脈的に意味の異なる語。たとえば、花の「うのはな」と豆腐 (おから) の「うのはな」。

<sup>7</sup>分類コードは 1 語に 1 つ付くわけではなく、複数に付くことがある (中野, 1969, p.44)。

い物・飲み物それ自体が意識的に排除されている」と述べている。そこで、シソーラスを使って飲食物が本当に見られないのか、確かめてみる<sup>8</sup>。しかしながら、“ないものを探す課題”を満足させるキーワードはない。思いつくままに食物名を検索したとしても、古語の食物名を漏らさず検索するのは難しい。

食料は、分類コードでは BG-01-4300 から始まるが、BG-01-43 で検索すると、BG-01-4300 (品目名以外、おかず、常食、飼料、餌…)、BG-01-4360 (薬剤、薬品)、BG-01-4370 (化粧品) なども出力されてしまうので、4300、4360、4370 を除いた番号で検索する<sup>9</sup>。その結果、12 首から「塩、蓼水、飯、餅、磯干鯛」の 5 品目が得られた (表 2)。

表 2: シソーラスコード「食料」(BG-01-43) を八代集シソーラスより検索した結果: 先頭の数字は行番号。次の 2 桁は歌集を示す。古今集 (01)、後撰集 (02)、拾遺集 (03)、後拾遺集 (04)、金葉集 (05)、詞花集 (06)、千載集 (07)、新古今集 (08)。後撰 (=02)1095 番には 2 品が出現する。

|    |    |        |      |     |                    |     |       |     |
|----|----|--------|------|-----|--------------------|-----|-------|-----|
| 1  | 01 | 000708 | 0005 | A00 | BG-01-4330-03-0100 | 塩   | しほ    | 塩   |
| 2  | 01 | 000758 | 0005 | A00 | BG-01-4330-03-0100 | 塩   | しほ    | 塩   |
| 3  | 01 | 000894 | 0009 | A00 | BG-01-4330-03-0100 | 塩   | しほ    | 塩   |
| 4  | 02 | 001095 | 0001 | A00 | BG-01-4330-03-0100 | 塩   | しほ    | 塩   |
| 5  | 02 | 001095 | 0014 | A00 | BG-01-4310-08-0700 | 蓼水  | ただみ   | 蓼水  |
| 6  | 03 | 000423 | 0005 | A00 | BG-01-4330-03-0100 | 塩   | しほ    | 塩   |
| 7  | 03 | 001350 | 0006 | A00 | BG-01-4310-02-0201 | 飯   | いゝ    | 飯   |
| 8  | 04 | 001203 | 0005 | A00 | BG-01-4310-06-0102 | 餅   | もちひ   | 餅   |
| 9  | 05 | 000501 | 0007 | A00 | BG-01-4321-01-0600 | 磯干鯛 | いそひたひ | 磯干鯛 |
| 10 | 08 | 001115 | 0004 | A00 | BG-01-4330-03-0100 | 塩   | しほ    | 塩   |
| 11 | 08 | 001590 | 0007 | A00 | BG-01-4330-03-0100 | 塩   | しほ    | 塩   |
| 12 | 08 | 001592 | 0005 | A00 | BG-01-4330-03-0100 | 塩   | しほ    | 塩   |
| 13 | 08 | 001701 | 0007 | A00 | BG-01-4330-03-0100 | 塩   | しほ    | 塩   |

表 3: 八代集に見られる食物を含む歌 (抜粋): 行頭の番号は、表 2 の行頭番号を示す。

1. すまのあまの／塩やく煙／風をいたみ／思はぬかたに／たなひきにけり【古今 708】
5. しほといへは／なくてもからき／世中に／いかにあへたる／たゝみ成らん【後撰 1095】
7. しなてるや／かた岡山に／いゝにうへて／ふせるたひ人／あはれおやなし【拾遺 1350】
8. みかの夜の／もちいはくはし／わつらはし／きけはよとのに／はゝこつむ也【後拾遺 1203】
9. あふことは／かたねふりなる／いそひたい／ひねりふすとも／かひやなからん【金葉 501】

それぞれの食品を実際の歌 (表 3) で吟味すると、「塩」は塩焼煙の「塩」なので、ごちそうではない。「ただみ」は蓼の葉のしぼり汁で「からい／つらい」の意味を添えた上で、作者名「みぶのただみ (壬生忠見)」を掛けた語である (片桐, 1990, p.325)。「飯」は「飯に飢へて」とあり、空腹を示す。「餅」は「三日の夜の餅は喰はじ」とあり「新婚三日目に食べる祝いの餅は食うまい」の意味 (久保田・平田, 1994, p.391)。「磯干鯛」の歌 (金葉集 501 番) は (川村他, 1989, p.142) によると「歌意不詳」とある。食物名はまったくないわけではない<sup>10</sup>が、9500 首中の 12 首 (0.12%)、さらに厳しく「蓼水」「餅」「磯干鯛」しかを認めないなら、わずか 3 首 (0.03%) であり、久保田 (2003) の指摘を計量的に裏づけることができた。

以上のように語を意味の単位で検索・集計できることを示したが、集計から作表までいくつかのツールを組み合わせたり、中間的なファイルを作成したり、プログラムを書いたりしなくてはならないので、ある程度お決まりのお作法が考えられるなら、お気楽にしておきたい。次ではそれら加工・集計などの処理に可視化の仕組みを加えたシステムについて述べる。

## 4 シソーラスを利用した語彙構造可視化システム

本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、すべての共出現パターンを生成し、

<sup>8</sup>(田島, 2000, p. 54) は「その分野の語がないことを示すのも重要である」と述べている。

<sup>9</sup>grep なら、% grep "BG-01-43[1-5]" hachidaishu.db で一括検索できる。

<sup>10</sup>物名 (ものな) にも食料が詠まれている事実があるが、物名は歌の内容を意味するものではなく、多くは単語や句をまたぐため、語を単位とするシソーラスには現段階では登録されていない。

その頻度と各パターンの情報量に基づく計算により、モデルを描いている。グラフは数理モデルであるので、語と語の依存関係(結びつく語の種類、多さ)が観察できるだけでなく、条件によって、パターンを分割、差分、合成などの操作が可能である。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、二十一代集すべてについて同様の処理ができるようにし、インターネット公開したいと考えている。

本システムをインターネットで公開する主な目的は、開発したシソーラスの利用方法を具体的に示すことではあるが、他に(1)このような方法がコンピュータ技術を持つ研究者だけでなく、広くさまざまな研究者によって、利用できるようにすること、(2)多くの研究者が利用することにより、意見の交換(モデルの妥当性、従来学説との比較など)を活発にすること、(3)文学だけでなく、言語、文化など広く人文科学の領域においてコンピュータによる研究方法の普及と理解を促進させることなども意図している。

グラフ生成は次のような手順によって行われる。まず、調査したい和歌を検索する。たとえば「梅」について調べたいときには、「梅」をもつ歌を選び出す。「梅」をもつ歌を対象に、1首に含まれるすべての単語を総当たり戦で2語1組のパターンを作る。1首が終わったらつぎの歌も総当たり戦でパターンを生成し、すべての歌について同じ処理を施す。このようにして作った1パターンはちょうどグラフでいう2ノード(2語)1エッジ(語と語の関係)になる。これを共出現パターンと呼ぶ。共出現パターンの数はキーワード(たとえば、「梅」)が出現する歌の数によるが、ほぼ数百~数千になる。しかし、すべてのパターンをグラフで描くと真っ黒な塊になってしまう。そこで各パターンについて重要度を計算し、その重要度の高いパターンから、順に出力していく<sup>11</sup>。

具体的な利用方法であるが、まず、図1に示すようにキーワードを入力する。キーワードは、語彙構造分析のターゲットであり、グラフモデルの中心となる語である。キーワードは、漢字、平仮名、あるいは分類コードのいずれでもよい。

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

## Computer Modelling of Japanese Poetic Vocabulary

Key:

--  
GPL Copyleft Hilofumi Yamamoto (yamagen[at]ucsd.edu): This file can be freely distributed.

図1: 検索キー入力画面: 検索キー「吉野」を入力して、OK ボタンを押す。

キーワードを入力して、OK ボタンを押すと、図2の一覧表が出力される。一覧表には、キーワードに該当する語、八代集の各歌集における頻度、全体の頻度が示される。

つぎに、ネットワークモデルを作図するために必要なパラメタの指定を行う。

1. どの語のモデルを出力するかを決める。  
一覧表左端のラジオボタンで選ぶ。
2. どの歌集からどの歌集までのデータを利用するかを決める。  
もし、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. 集計のレベルを指定する (Level)。  
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、語コードの桁数を意味する。
4. 重要度の計算方法を選ぶ (Method)。  
計算方法は試作段階であるため、現在は 7,12~16 の方法が用意されている。計算方法の詳細については省略する。今後も試行錯誤が必要のため、オプションを用意するだけでなく、計算式を直接入力できるようにすることも検討中である。
5. 語の単位のサイズを指定する (Unit)。

<sup>11</sup>山元(2006)では一般名詞について、山元(2005)では歌枕(和歌に詠まれる地名)について、報告しているのでご参照いただきたい。

たとえば、「吉野山」の場合は「吉野」「山」の2語に分割して集計するか、「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。

6. 標準得点による抽出 (Z-score)。

パターンの分布は正規分布に近似することがわかっている。このオプションでは、標準得点による抽出を行わない、1 標準偏差以上を抽出、2 標準偏差以上を抽出、が選べる。

7. Log-likelihood(Dunning, 1993) による検定。

Log-likelihood による2語の関連性検定を行わない、行う ( $p < .05$ , あるいは  $p < .01$  レベル)、を指定する。指定した場合はレベルに応じて検定で棄却されたエッジは出力されない。

8. 重要度の値による描画パターンの抽出 (Co-occurrence Weight)。

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

**Computer Modelling of Japanese Poetic Vocabulary**

データベースより「吉野」を検索しました。

| 出現頻度 | 辞書コード                 | 出現語 | 節名    | 代詞語  | 古今集 | 後撰集 | 拾遺集 | 後拾遺集 | 後撰集 | 新古今集 | 新古今集 |    |
|------|-----------------------|-----|-------|------|-----|-----|-----|------|-----|------|------|----|
| ◎    | 62 CH-29-0000-00-2600 | 吉野  | よしの   | 吉野   | 15  | 6   | 10  | 1    | 5   | 3    | 7    | 15 |
| ○    | 32 CH-29-0000-00-2603 | み吉野 | みよしの  | 吉野   | 8   | 3   | 6   | 2    | 0   | 3    | 2    | 8  |
| ○    | 20 CH-29-5240-05-0803 | 吉野山 | よしのやま | 吉野山  | 0   | 2   | 5   | 1    | 3   | 0    | 3    | 6  |
| ○    | 7 CH-29-5250-01-0701  | 吉野河 | よしのがは | 吉野河  | 5   | 1   | 0   | 0    | 0   | 0    | 0    | 1  |
| ○    | 5 CH-29-0000-00-2601  | 御吉野 | みよしの  | 吉野   | 2   | 2   | 0   | 0    | 0   | 0    | 0    | 1  |
| ○    | 2 CH-29-5250-01-0700  | 吉野川 | よしのがは | 吉野川  | 0   | 0   | 0   | 0    | 0   | 0    | 2    | 0  |
|      |                       |     |       | From | ◎   | ○   | ○   | ○    | ○   | ○    | ○    | ○  |
|      |                       |     |       | To   | ◎   | ○   | ○   | ○    | ○   | ○    | ○    | ○  |

  

|                      |       |           |           |      |      |      |      |      |      |
|----------------------|-------|-----------|-----------|------|------|------|------|------|------|
| Level                | ◎16   | ○18       |           |      |      |      |      |      |      |
| Method               | ◎7    | ○12       | ○13       | ○14  | ○15  | ○16  |      |      |      |
| Unit Size            | ◎0    | ○1        | ○2        |      |      |      |      |      |      |
| Z-score              | ◎0    | ◎1        | ◎2        |      |      |      |      |      |      |
| Log-Likelihood Value | ◎None | ◎p < 0.05 | ◎p < 0.01 |      |      |      |      |      |      |
| Co-occurrence Weight | ◎1.0  | ◎1.5      | ◎2.0      | ◎2.5 | ◎3.0 | ◎3.5 | ◎4.0 | ◎4.5 | ◎5.0 |

ok

図 2: ネットワーク出力のための各種設定画面: 要求されたキーワードに基づいて、データベースより該当の語句がリストアップされる。八代集各歌集における頻度も出力される。

上記の条件を指定し、OK ボタンを押すと、図 3 に示すようなネットワークが出力される。図 3 の条件は、キーワードが「吉野」、金葉集から千載集のテキストを対象とし、集計レベルが 16 (表記による違いを無視する)、重要度計算方法が 16、unit size が 2 (「吉野山」であれば「吉野」と「山」に分割する)、Z 値が 1 以上、Loglikelihood 指定なし、重要度 2.5 以上のパターンをグラフとして描画したものである。描かれた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。

グラフはブラウザの対応状況によって 2 種類の画像フォーマットで出力される。将来的にすべてのブラウザで、SVG(Scalable Vector Graphics) が対応されるものと思われる。SVG は XML(Extensible Markup Language) 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いつまでも劣化のない画像が得られる。SVG で描かれたファイルには、各パターンのリスト、重要度の値、描画に用いられた条件など、実際に作図に用いられたデータが含まれており、それらを利用者が自分の論文に引用することができる。

実際にモデルを見ているとその各ノードはどの歌が起因しているのかを確かめたい。そこで、各ノードをマウスでクリックしてみる。各ノードには歌集番号と歌番号の ID が埋め込まれている。SVG では、その ID を利用した外部リンクで、そのノードが出現する和歌にリンクされている。一方、PNG ではクリックカプルマップ方式で、リンクされている。たとえば、図 3 のネットワークモデルで「花」をクリックすると、図 4 のように、そのノードに該当する和歌が出力される。この出力には国文学研究資料館開発の「二十一代集」電子テキストが用いられている<sup>12</sup>。

<sup>12</sup>インターネット公開については既に国文学研究資料館知的財産委員会より許諾を得ている。





## Computer Modelling of Japanese Poetic Vocabulary

1. 50047 顕季 桜花／咲ぬる時は／よし野山／立ものほらぬ／峯のしら雲
2. 50052 忠隆 吉野山／みねになみよる／白雲と／見ゆるは花の／梢なりけり
3. 50065 御匣殿 さくら花／雲かゝるまで／かきつめて／よしの山と／けふは見る哉
4. 50523 定信 みな人は／吉野山／の／桜花／おりしらぬ身や／谷の埋木
5. 60022 匡房 白雲と／みゆるにしろし／みよしの山と／吉野の山の／花さかり哉
6. 70065 顕昭 吉野川／みかさばさしも／まさらしを／青根をこすや／花の白浪
7. 70070 為業 吉野山／花のさかりに／なりにけり／たゆる時なき／峯の白雲
8. 70076 俊成 御吉野の／花のさかりを／けふみれば／こしのしらねに／春風ぞ吹
9. 70080 季通 芳野山／はなはなかはに／散にけり／たえくたえ／残る／峯のしら雲
10. 70093 俊恵 御よしの山と／山下風や／拂らむ／梢にかゝる／花のしら雲
11. 71073 定家 いつくにて／風をも世をも／恨まし／吉野山奥も／花はちりけり

図 4: ノード「花」に相当する和歌の出力: 和歌のテキストは貴機関開発の「二十一代集」を利用。

片桐洋一 (1990) 『後撰和歌集』, 新日本古典文学大系, 岩波書店, 東京.

川村晃生・柏木由夫・工藤重矩 (1989) 『金葉和歌集、詞花和歌集』, 新日本古典文学大系, 岩波書店, 東京.

久保田淳・平田喜信 (1994) 『後拾遺和歌集』, 新日本古典文学大系, 岩波書店, 東京.

久保田淳 (2003) 「文学における食」, 『国文学』, 第特集「食の文化誌」巻, 第7号, 6-7頁.

宮島達夫・小沼悦 (1992) 「言語研究におけるシソーラスの利用」, 『国立国語研究所報告』, 第104巻, 第13号, 1-30頁.

中野洋・林大・石井久雄・山崎誠・石井正彦・加藤安彦・宮島達夫・つる岡昭夫 (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.

中野洋 (1969) 「新聞語彙調査の類別語彙表について」, 『電子計算機による国語研究II』, 第34巻, 国立国語研究所報告, 秀英出版, 東京, 38-54頁.

西端幸雄・藤田久・成田徹 (1989) 『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.

田島毓堂 (1995) 「語彙と単語」, 『日本語学』, 第14巻, 4-11頁.

—— (2000) 「語彙研究法としての「意味分野別構造分析法」のために—阪倉論文の顕彰と批判—」, 『国語と国文学』, 第918巻, 43-56頁, 東京大学国語国文学会.

辻勝美 (1998) 「歌語の研究史—現状と展望—」, 小町谷照彦・三角洋一 (編) 『歌ことばの歴史』, 笠間書院, 東京, 217-238頁.

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.

—— (2009 印刷中) 「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.