分類語彙表による歌ことばシソーラスの開発
Thesaurus of Japanese Poetic Vocabulary Based on the Semantic Classifications Chart

山元啓史
Hilofumi Yamamoto

オーストラリア国立大学
The Australian National University
Faculty of Asian Studies ANU ACT 0200 Australia

あらまし：和歌の単語は、さまざまに表記されるため、単語の検索や一括集計などの処理が困難である。この処理を確実にするためには、シソーラスによって表記を統一する必要がある。本稿では和歌の単語を国立国語研究所の分類語彙表の体系コードにならって、歌ことば用のシソーラスコード辞書およびコード変換ツールを開発した。材料は八代集の和歌、約 9500 首を用いた。さまざまな表記を収集するために、できる限り多くの資料を参考にした。地名、人名は分類語彙表には一部のコードしか存在しないため、新たに作成した。その結果、同形異語、異形同語の特定と類語の分類ができた。今後もシソーラスを充実させる予定である。

Summary: It is difficult to search and count words appearing in classical poems because they are written in various ways. It is necessary to transliterate words into unified codes using a thesaurus in order to solve the above mentioned problem. This paper addresses a development of the thesaurus of classical Japanese poetic vocabulary based on the *Bunruigoihyō*. The *Hachidaishū* which consists of approximately 9,500 poems is used as a material for the development. In order to collect different spellings of words in poems various texts which have been published are analised. The thesaurus of proper names such as personal names and place names has been developed from scratch since they are not included in the *Bunruigoihyō*. The thesaurus of this project allows us to properly distinguish and classify words appearing in classical poems.

キーワード：和歌、歌ことば、単位分割、タグづけ、分類語彙表、八代集
Keywords: Japanese poetry, poetic vocabulary, tokenize, tagging, the *Bunruigoihyū*, the *Hachidaishū*

1 はじめに

多くの人々に使われる言語は通常方言や語族に系統があり、広い地域で使用され、多様な文化に影響を受けている[1]。ところが、日本語は系統として孤立し、どの語族にも属さない言語であり、使用人口第 6 位、1 億人以上もの人々によって話される言語である (宮島他, 1982, p.15–6) しかも、1000 年以上前の言語が均一な形で観察できる。古代日本語と現代日本語は「大筋」においては変わっていない。このような言語は世界の言語のうちでも珍しいといわれる (阪倉, 1977)。

古代語で記された資料のなかでも、和歌は文学作品であると同時に古代の言語や文化を知るための貴重な資料であり、とりわけ八代集は言語の変遷を通観する上ではなくてはならない、きわめて重要な資料で

[1] たとえば、英語は歴史ある言語の一つではあるが、およそ 400 年–1100 年ごろに使われたと推定される古い英語（古英語）は一形式にまとめられるものではなく、時期的に前後するいくつかの方言からなっていたり、名詞には性の区別があったりなど、現代の英語とは異なる点が多い。

ある。八代集は、古今集（905年）から新古今集（1205年）までの300年間に撰集された８つの勅撰集からなり、ほぼ9500首の和歌が収められている。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なされている (辻, 1998, p.226)。

ところが、八代集の史的展開の中では「千載集から新古今集にかけての時代において歌ことばの使用についての急激な変化が認められる (辻, 1998, p.227)」とも、「語彙の転換期は後拾遺集あたりから (浅見, 1986; 上野, 1976; 川村, 1991)」とも、「語彙論の立場から八代集の各歌集中の和歌の使用語句の性格を精査してみると『拾遺集』に転換期がある (西端, 1994)」ともいわれる。転換期に関する学説がさまざまであるのは一体何が問題となっているのか。

語彙の体系は一つの平面の上にかけるものではなく、意味、形、文体などいくつかの側面の総合として存在するゆえ、各側面ごとに見ていかなければならないといわれる (宮島, 1977, p.4)。特定の側面のみを切り取って体系づけると他の側面から見た場合、矛盾した結果となってしまう。意味は連続した世界であるため、その意味だけによって単語を認定しようとすると単語でさえも規定できなくなってしまう[2]。語の単位の認定方法、語の出現頻度、和歌の内容が示す文化的色彩、歌風や歌題の変化、技巧、意味内容、語感の違いなど、視点によっては転換期そのもののとらえ方は一意に決定できるものではないと考えられる。

このような語彙の多彩な側面を総合的に捉え、かつ他の研究との互換性を維持するためには、研究データの管理、具体的には各作品での用語の統一・管理が不可欠になる。シソーラスは用語を体系的に管理した語彙リストであり、語彙の計量研究によく用いられている。日本語のシソーラスである国立国語研究所が開発した「分類語彙表」(中野他, 1994)[3]はさまざまな目的で、数多くの研究で利用されている (宮島・小沼, 1992; 中野他, 1994)[4]。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、語彙は意味的存在であるとし「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、自らも分類語彙表を利用し、語彙の意味を分野別に検討し、総体論として語彙の構造を分析している。しかし「語に意味コードを付ける作業の困難さがあり、これを解決する必要がある」と述べ、コードづけ作業の困難さも指摘している。

西端他 (1989) はコンピュータが個人に普及しはじめた早い時期に自らプログラムを書き、和歌の語彙索引自動作成を試みた。科学研究費の助成を得て「和歌語彙データベースの開発」「平安朝和歌の語彙論的研究」などの研究を行い、平安時代から室町時代初期までの和歌集、物語作品所収の和歌、私家集を対象としたデータベースの開発を手がけた。後者の研究では「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」旨が報告されており、ここでもコードづけの作業が大仕事であることがわかる。またコンピュータを利用して作業を進めたために、語彙の諸問題が改めて認識されることも少なくない。たとえば、土屋 (1978, p.2) はコンピュータによる同語異語判別の作業に際して「カードを使った手集計の語彙調査では、この作業（同語異語判別の作業）は、カードを採る段階で、作業者の頭の中で大部分が行われてしまい、大仕事だとは意識されていない」と述べている。

本研究では、和歌の用語の統一・管理作業をできるだけ、計算機に任せ、なおかつ従来の研究成果が蓄積されるように、１）和歌のためのシソーラスコードと、２）それを和歌テキストに自動的に付けるためのツールを開発し、最終的にそれらを利用して、３）八代集のための歌ことばシソーラスを開発した。本稿では、その開発過程および問題、課題について報告する。

---

[2] 西尾 (1988, p.20) は、もし１個の事物の名称は１単語であるという意味優先の基準を立てようとすると、「藻塩草」の別称である「リュウグウノオトヒメノモトユイノキリハズシ」も１語になってしまうことを指摘している。

[3] 「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。本研究では1994年刊行のフロッピー版を用いた。

[4] 宮島・小沼 (1992) の調査によると、136の論文で分類語彙表が利用されていることが示されている。

2 方法

和歌のテキストは、国文学研究資料館作成による二十一代集データベースを利用した。底本は国文学研究資料館蔵「正保版本二十一代集」である[5]。八代集の成立、撰者、収録和歌数を表 1 に示す。和歌のそれぞれには新編国歌大観準拠の歌番号をつけた。

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

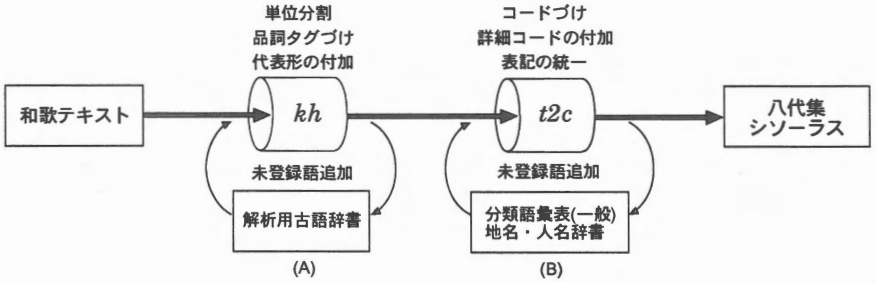

図 1: 八代集シソーラスの開発の流れ

八代集シソーラスの開発は図 1 に示す流れで行った。八代集の和歌テキストを 古文自動品詞タグ付けシステム kh (Kobun to Hinshi) (山元, 2007) で単位分割し、品詞情報を加える (A)[6]。次に、その出力を t2c (Token to Codes) で処理して、シソーラスコードを付加する (B)。kh は清濁のないテキストには対応していないため、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、清濁を補った上で処理した。分割の単位は国立国語研究所 β 単位（複合語は一次結合までを認める）とした。活用のない語は注釈書などにより判断できる限り、相当する漢字列が代表形として出力する。活用のない語は基本形（いわゆる終止形）およびその漢字列を代表形として出力する[7]。kh で処理をする際には、未登録の語や表記の異なる語を逐一登録しつつ、解析結果が正しいかどうか、全首について確認した。

次に t2c を使ってシソーラスコードを加える。t2c は単位切りした語を入力すると分類語彙表のコードを返すプログラムである。たとえば「立田」「竜田」「龍田」は互いに異表記の同義語（異形同語）であるが、コンピュータでは別の語として扱ってしまうので、t2c でシソーラスコードをつけた上で、代表形となる表記（この場合は「立田」）をつけて用語を統一する作業をする[8]。本シソーラスの開発では、t2c は

[5] http://ocelot.nijl.ac.jp/dlib/21dai/README-21dai.html
[6] kh の詳細については山元 (2007) を参照いただきたい。
[7] 仮名表記は実際にどんな漢字が当てられるのが適当かわからない。たとえば、コンピュータでコードづけする際には【明く・開く・空く】（あるいは【飽く・厭く・倦く】）のいずれであるか、よくわからない。
[8] 異形同語の問題は日本語の検索においてしばしば問題となる。日本語の情報検索がいかに困難であるかという例として「きんのたまごをうむにわとり」をあげられる (Halpern, 2002)。おそらく「金の卵を産む鶏」が標準的な表記であろうが、「たまご」（卵、玉子、たまご、タマゴ）、「にわとり」（鶏、にわとり、ニワトリ、庭鳥）、「うむ」（産む、生む）などがあるので、それらを組み合わせると 24 通りの表記にもなる。

kh からの出力をそのまま入力として受けとり、処理するので、できるだけ多くの情報（【漢字：よみ：品詞】の３つのデータ）を利用し、より適切な情報を辞書より探して出力するようにしている。

t2c が利用する辞書データは、一般語（BG）、地名（CH）、人名（PN）の３種類である。一般語の辞書（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。分類語彙表のコード体系は、語彙をまず品詞で「1. 体（名詞）」「2. 用（動詞）」「3. 相（形容詞・副詞）」「4. その他（接続詞・感動詞など）」の４つに分け、その下位をそれぞれを意味で「1. 抽象的関係」「2. 人間活動の主体」「3. 人間活動—精神および行為」「4. 生産物および用具」「5. 自然および自然現象」の５部門に分けている。さらに下位項目に細分した上で、具体的に実際の語を管理している。ただし、「2. 用」「3. 相」には「2. 人間活動の主体」と「4. 生産物および用具」はなく、「4. その他」には「1. 抽象的関係」「3. 人間活動—精神および行為」のみがある (犬飼, 1988; 田島, 1999; 山田, 2002)[9]。さらに、文法的な性質を主な役割とする語（文法質）(田島, 1999) も取り出して分類ができるように、上記の品詞区分１から４に加えて、田島 (1999) の提案する新設コードの５から 18 を追加した[10]。本研究では、地名（CH）と人名（PN）を同時に利用するため、一般語のデータを BG（分類語彙表の略）とし、シソーラスコードの先頭に識別子としてつけた。また、旧版分類語彙表では小数点によるコード記法であったが、それを改めてすべて同じ桁数で揃えるようにした。

表 2: 一般語のコード体系：BG-01-5520 は「植物」。BG-01-5520-17 は「柑橘類」。１レコードは１行で、フィールドの区切りは':' である。各フィールドは左より、シソーラスコード、品詞番号１、品詞番号２、品詞番号３、出現形、よみ、代表形、下位シソーラスコードとなっている。下位シソーラスコードは語をさらに分析した時に該当するコードで、'+' が区切りである。

```
BG-01-5520-17-0100:02:00:00:きんかん:きんかん:金柑
BG-01-5520-17-0101:02:00:00:金柑:きんかん:金柑
BG-01-5520-17-0200:02:00:00:だいだい:だいだい:橙
BG-01-5520-17-0201:02:00:00:橙:だいだい:橙
BG-01-5520-17-0300:02:00:00:ゆず:ゆず:柚
BG-01-5520-17-0301:02:00:00:柚:ゆず:柚
BG-01-5520-17-0400:02:00:00:たちばな:たちばな:橘
BG-01-5520-17-0401:02:00:00:橘:たちばな:橘
BG-01-5520-17-0500:02:00:00:やぶこうじ:やぶこうじ:薮柑子
BG-01-5520-17-0501:02:00:00:薮柑子:やぶこうじ:薮柑子
BG-01-5520-17-0600:02:00:00:みかん:みかん:蜜柑
BG-01-5520-17-0601:02:00:00:蜜柑:みかん:蜜柑
BG-01-5520-17-0700:02:00:00:なつみかん:なつみかん:夏蜜柑:BG-01-1624-03-0101+BG-01-5520-17-0600
BG-01-5520-17-0701:02:00:00:夏蜜柑:なつみかん:夏蜜柑:BG-01-1624-03-0101+BG-01-5520-17-0600
BG-01-5520-17-0800:02:00:00:ザボン:ざぼん:ザボン
BG-01-5520-17-0801:02:00:00:朱欒:ざぼん:ザボン
BG-01-5520-17-0900:02:00:00:ネーブル:ねえぶる:ネーブル
BG-01-5520-17-1000:02:00:00:オレンジ:おれんじ:オレンジ
BG-01-5520-17-1100:02:00:00:レモン:れもん:レモン
BG-01-5520-17-1101:02:00:00:檸檬:れもん:レモン
BG-01-5520-17-1200:02:00:00:からたち:からたち:枸橘
BG-01-5520-17-1201:02:00:00:枸橘:からたち:枸橘
BG-01-5520-17-1202:02:00:00:枳殻:からたち:枸橘
BG-01-5520-17-1300:02:00:00:さんしょう:さんしょう:山椒
BG-01-5520-17-1301:02:00:00:山椒:さんしょう:山椒
BG-01-5520-17-1400:02:00:00:はなたちばな:はなたちばな:花橘:BG-01-5530-12-0100+BG-01-5520-17-0400
BG-01-5520-17-1401:02:00:00:花たちばな:はなたちばな:花橘:BG-01-5530-12-0100+BG-01-5520-17-0400
BG-01-5520-17-1402:02:00:00:花橘:はなたちばな:花橘:BG-01-5530-12-0100+BG-01-5520-17-0400
BG-01-5520-17-1500:02:00:00:やまたちばな:やまたちばな:山橘:BG-01-5240-05-0100+BG-01-5520-17-0400
BG-01-5520-17-1501:02:00:00:山たちばな:やまたちばな:山橘:BG-01-5240-05-0100+BG-01-5520-17-0400
BG-01-5520-17-1502:02:00:00:山橘:やまたちばな:山橘:BG-01-5240-05-0100+BG-01-5520-17-0400
```

[9] 自然言語処理でよく用いられる日本語語彙大系 CD-ROM 版 NTT コミュニケーション科学基礎研究所 (1999) は「30 万語の収録語は 3000 種の意味分類を用いて定義されており、最大規模の日本語シソーラスとなっています」と謳っているが、意味体系の記述は分類語彙表を踏襲している。ただし、日本語語彙大系では固有名詞が充実している。

[10] 田島 (1999, p.120–2) による新設コードは、5. 接頭辞、6. 接中辞,7. 接尾辞、8. 助詞、9. 助動詞、10. 補助動詞・補助形容詞、11. 関係詞、12. 語尾、13. 前置詞・介詞、14. 意味不明、15. 固有名詞 16. 記号であるが、15. 固有名詞は別に CH を作成したので、使用していない。

表 2 に一般語のデータ例を示す。分類語彙表には個々の語を示す番号は与えられていない。本研究では個別の語にも番号を割り振った。ただし、番号が近いことは意味が近いこととは関係ないものとした。個別の語の番号は 4 桁で表示し、異形同語の場合は下 2 桁を、別の語の場合には上 2 桁を変更することにした。現代語のテキストを処理することも考慮し、現代語の形態素解析システム ChaSen に用いられている品詞番号を追加した。シソーラスコードは 18 桁で表示される。はじめの 16 桁を有効桁数にすれば、異形同語と判定されたものは同語として処理できる[11]。

分類語彙表は現代語を前提に開発されているので、厳密には古代語のシソーラスではないが、平安末期と現代とに、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 (犬飼, 1988, p.38(271))[12]。ところが、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、問題が大きい (犬飼, 1988, p.38(271))。その場合には広辞苑と古語辞書を参照し、できるだけふさわしい番号が割り振られるように努めた。分類語彙表には、地名は一部だけで歌枕はほとんどなく、人名もないので、これらは新規に作成した。表 3 に地名と人名のデータ例を示す。

表 3: 地名（CH）吉野川と人名（PN）のデータ例：地名・人名いずれも区切りは ':'。左より、シソーラスコード（CH は地名項目、29 は県コード、5250 は一般語 (BG) の河川のコード、品詞番号 1（11 は地名、08 は人名）、品詞番号 2、品詞番号 3、漢字表記、よみ、代表形、下位シソーラスコード。

```
CH-29-5250-01-0700:11:00:00:吉野川:よしのがは:吉野川:CH-29-0000-00-2600+BG-01-5250-01-0102
CH-29-5250-01-0701:11:00:00:吉野河:よしのがは:吉野河:CH-29-0000-00-2600+BG-01-5250-01-0103
CH-29-5250-01-0702:11:00:00:芳野河:よしのがは:芳野河:CH-29-0000-00-2604+BG-01-5250-01-0103
CH-29-5250-01-0703:11:00:00:吉野川:よしのがわ:吉野川:CH-29-0000-00-2600+BG-01-5250-01-0100
CH-29-5250-01-0704:11:00:00:吉野河:よしのがわ:吉野河:CH-29-0000-00-2600+BG-01-5250-01-0101
CH-29-5250-01-0705:11:00:00:芳野河:よしのがわ:芳野河:CH-29-0000-00-2604+BG-01-5250-01-0101
CH-29-5250-01-0706:11:00:00:吉野の河:よしののかわ:吉野の河:CH-29-0000-00+BG-01-5250-01-0101
PR-01-00UT-01-0100:08:00:00:宇多:うた:宇多
PR-01-00UT-01-0200:08:00:00:宇多天皇:うたてんのう:宇多天皇
PR-01-KT00-01-0100:08:00:00:友則:とものり:友則
PR-01-SH00-01-0100:08:00:00:真静:しんせい:真静
PR-01-SH00-01-0200:08:00:00:真静法師:しんせいほうし:真静法師
PR-01-HH00-02-0100:08:00:00:遍照:へんじょう:遍照
PR-01-HH00-02-0200:08:00:00:遍照法師:へんじょうほうし:遍照法師
```

表 4: タグづけ済みの八代集シソーラス：左より、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 はコードを複数取るかどうかを示すフラグ、以降順に、シソーラスコード、漢字、よみ、代表形。

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

[11] このように下 2 桁で異形同語を区別し、同じ番号を与えなかったのには、「形が違えば同じ語ではない」という立場・考え方があるからである。

[12] 犬飼 (1988, p.38(271)–39(270)) は、その例として「例えば、「あたふ」には「与える」の分類番号 2.377 を与えればよい。また、「くらがる」という動詞は現代に生きていないが、名詞「くらがり」との連携等から「光」類の 2.501 を与えることができる」と述べている。また「厳密に処理するためには、各時代毎の「分類語彙表」を作成しなくてはならなくなる」とも述べている。

kh で単位分割品詞タグづけされたデータを t2c で処理するが、その際、該当データなしでエラーが返ってきた語については、各辞書にその不足情報を登録した。八代集データのすべてにエラーがなくなるまで繰り返し辞書登録作業を続け、その結果、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。以上で八代集に見られる語とその表記は網羅された。しかし、同形異語（語の形は同じだが、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」）の問題があるため、t2c で処理しただけでは、文脈から考えてふさわしくないコードや複数に該当するコードが出力される。そこで、１首ごと出力を確認し、不要なコードを排除した。ただし、１つの語に複数のコードも認められることもあるので、すべての語が１つのコードしかないというわけではない。以上の手続きを経て、八代集シソーラスが完成した。表 4 は八代集シソーラスの例（古今集２番、紀貫之）である。

3　検索例

シソーラスで語彙調査を実施すると、(1) 異形同語であっても、同じ語として集計できる、(2) 分類カテゴリで集計でき、カテゴリ毎の比を求めることができる、さらに (3) 存在しない語やそのカテゴリを指摘することができる。特に (3) は、テキストを分割し、文字列を検索するだけではむずかしい。シソーラスとの照合作業、ある体系を持ったリストと照合しない限り、欠落したカテゴリの指摘はできない。

まず、異形同語の検索であるが、「立田」のコード「CH-29-0000-00-1800」のうち上 16 桁を検索・集計すると「立田」「竜田」「龍田」の３種類の抽出され、八代集中でのそれぞれの表記と頻度、立田 (54)、竜田 (5)、龍田 (4) と合わせた頻度 (63) が簡単に出力できた。つぎに分類カテゴリによる検索だが、「BG-01-5520」は植物名のカテゴリを示すコードで、これでシソーラスを検索すると、「松」をはじめ、203 種類の植物名およびその頻度を出力することができた。

最後に欠落している分類カテゴリの検索例を示す。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ろ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」と述べている。ならば、和歌は人間のさまざまな心模様、喜怒哀楽を表現しているはずである。しかし「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」「今年も（食物名）の季節になったのだなあ」など美食の歌がないのである。

実は、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、食い物・飲み物それ自体が意識的に排除されている。俳諧の世界ではそのような規制はなくなり、芭蕉も蕪村も多くの食い物・飲み物の秀句を残している」と述べ、和歌において飲食物が見られないことを述べているのである。そこで、シソーラスを利用して飲食物の歌が本当にないのかどうか、確かめることを企む。検索は適切なキーワードをユーザが心得ているかどうかが成功の秘訣であるが、「含まれないもの／存在しないものを探せ」という課題を遂行するためのキーワードはない。思いつくままにあらゆる食物名を検索したとしても、古語の食物名まで漏らさず検索するには限度がある。

食料は、シソーラスコードでは、BG-01-4300 からはじまるが、このカテゴリには BG-01-4300（品目名以外、おかず、常食、飼料、餌など）、BG-01-4310（飯・そば・パン・汁など）、BG-01-4320（米・糠・小麦粉など）、BG-01-4321（乾物・漬物・煮物など）、BG-01-4322（梅干・豆腐・寒天・とろろなど）、BG-01-4323（さかな・鰹節・肉）、BG-01-4330（調味料・麹など）、BG-01-4340（菓子）、BG-01-4350（飲料・たばこ）、BG-01-4360（薬剤・薬品）、BG-01-4370（化粧品）などが含まれる。そこで食料だけを選ぶためには、4300、4360、4370 を除けばよい。grep で検索するなら、

```
% grep "BG-01-43[1-5]" hachidaishu.db
```

でよい。上記のコマンドで検索してみたところ、表 5 のように 13 首から「塩、蓼水、飯、餅、磯干鯛」の５品目が得られた。

表 5: 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。

```
 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 いゐ 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
```

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【古今集 708 番】
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【後撰集 1095 番】
7．しなてるや／かた岡山に／いゐにうへて／ふせるたひ人／あはれおやなし【拾遺集 1350 番】
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【後拾遺集 1203 番】
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【金葉集 501 番】

歌を通して各食品を吟味する。【古今集 708 番】「塩」は塩焼煙の「塩」であるので、食塩ではあるが、ごちそうではない。【後撰集 1095 番】「ただみ」は蓼の葉のしぼり汁に味噌を加えた冷たい汁物であるから、確かに食物である。ただし「辛い（からい／つらい）」の意味を添えた上で、作者（壬生忠見）の名前を掛けている歌である (片桐, 1990, p.325)。【拾遺集 1350 番】の聖徳太子の歌は「飯に飢へて」なので美食の歌ではない。【後拾遺集 1203 番】「三日の夜の餅は喰はじ」は「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)。【金葉集 501 番】「磯干鯛」はこの語を見ただけでもおいしそうな感じがする。しかし、川村他 (1989, p.142) の注釈によると「いそひたひ」を「磯干鯛」と認めながらも「歌意不詳」とある。以上、食物名はまったくないわけではないことがわかった[13]。しかし、9500 首中の 13 首 (0.14 %)、「蓼水」「餅」「磯干鯛」を認める立場なら、わずか 3 首 (0.03 %) であるから「食い物・飲み物それ自体が意識的に排除されている」という主張がいかに正確なものなのかがよくわかる。あらかじめ和歌に出現した語彙をシソーラスにあてはめる作業をしておくと、存在しないカテゴリを探すことができる。

4 おわりに

本稿では、八代集の語彙のシソーラスとその処理ツールの開発について報告した。これらは「歌ことばのモデリングシステム (山元, 2006)」（ネットワーク表現による歌ことばの可視化システム）の内部に用いるデータベースとして開発されたものであるが、語彙の体系的情報を提供するため、他の研究にも利用できるものと思われる。

シソーラスコードを用いることによって、異形同語の語も一括して検索、同形異語のコードも抽出できるようになった。人手による確認作業は必要ではあるが、コードづけの作業は軽減された。照合すべきコードの桁数によって、上位カテゴリの検索や集計も可能になった。シソーラスを参照することによって、和歌に出てくるカテゴリと出てこないカテゴリを確認することができた。上位–下位の関係は文化や状況によって固定的ではないため、上位カテゴリの検索・集計には洩れや矛盾があり、まだ完全ではない。しかし、人手でこの作業を行うのは大変であるので、完全ではないことを承知の上であれば、参考になる情報は得られよう。

[13] 他にも物名（歌に物の名前を詠み込んだもの）に食物名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、また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登録できない。

既存の分類語彙表の構造に即して古語データを追加していった。すなわち、「即して追加した」ことにより、基礎研究として国立国語研究所で行われた数々の研究成果（たとえば、語彙体系、語彙単位、類義語判別に関する研究など）が本研究のシソーラスにも反映されているものと思われる。しかし、語彙を時代別、カテゴリ別に見たときに、八代集シソーラスから得られる体系的な語彙構造がどの程度の妥当性・信頼性のあるものなのか、今後も検証する必要があろう。

現段階では本シソーラスは八代集をカバーするだけである。今後、他の古文についても調査し、充実させ、和歌だけでなく連歌や散文においても利用できるものにしていきたい。機能的には、データ作成上の注釈や問題点も併せて記述していくために、現在のデータをＸＭＬに拡張する予定である。これらのデータは他の研究においても有益であると思われるので、諸方面に問い合わせ著作権およびその他の権利の問題がクリアされるならば、インターネットで公開したいと考えている。

引用文献

浅見徹 (1986)「八代集における季節」，国語語彙史研究会（編）『国語語彙史の研究』，第 7 巻，和泉書院，111–31 頁.

Halpern, Jack (2002) "Lexicon-based orthographic disambiguation in CJK intelligent information retrieval", in *COLING '02: Proceedings of the 3rd workshop on Asian language resources and international standardization*, pp. 1–7, Morristown, NJ, USA: Association for Computational Linguistics.

犬飼隆 (1988)「平安末期複合動詞の意味構造」，国語語彙史研究会（編）『国語語彙史の研究』，第 9 巻，和泉書院，272–258 頁.

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

川村晃生 (1991)『摂関期和歌史の研究』，三弥井書店.

久保田淳・平田喜信 (1994)『後拾遺和歌集』，新日本古典文学体系，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.

宮島達夫 (1977)「語彙の体系」,『語彙と意味』，第 9 巻，岩波講座日本語，岩波書店，東京，1–41 頁.

宮島達夫・野村雅昭・江川清・中野洋・真田信治・佐竹秀雄（編） (1982)『図説日本語: グラフで見ることばの姿』，第 9 巻，角川小辞典，角川書店，東京.

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

西端幸雄 (1994)「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—」，国語語彙史研究会（編）『国語語彙史の研究』，第 14 巻，和泉書院，318–303 頁.

西尾寅弥 (1988)『現代語彙の研究』，明治書院.

NTT コミュニケーション科学基礎研究所（編） (1999)『日本語語彙大系 CD − ROM 版』，岩波書店，東京.

阪倉篤義 (1977)「ことばの古さと新しさ」，阪倉篤義（編）『日本語の歴史』，第 6 巻，日本語講座，大修館書店.

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

——— (1999)『比較語彙研究序説』，笠間書院.

土屋信一 (1978)「高校教科書の同語異語判別システム」,『電子計算機による国語研究 IX』，第 61 巻，国立国語研究所報告，秀英出版，東京，1–16 頁.

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.

上野理 (1976)『後拾遺集前後』，笠間書店.

山田進 (2002)「意味分類辞書」,『国語学』，第 53 巻，第 1 号，30–43 頁.

山元啓史 (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 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.