2006－CH－71 (4)
2006／7／28

季語データベースの構築と「俳句投句鑑賞システム」の概要

吉岡 亮衛
国立教育政策研究所

本論は、季語データベースを構築するに当たって使用した4つの季語集について、データ作成のための処理手順と得られた季語からみた季語集の特徴について報告する。また、俳句データベースの構築のためにできるだけ多くの俳句を収集することを目的として開発した「俳句投句鑑賞システム」の概要を説明する。

The Development of the Kigo-database and the Outline of the "Haiku Entry and Appreciation System"

Ryoei Yoshioka
National Institute for Educational Policy Research

This paper reports on the processing procedure for data preparation and the characteristic by looking over the obtained kigos of four kigo books which were used for constructing kigo database. Also, this explains about the outline of the "haiku entry and appreciation system" that was developed for the purpose of that collects haiku as much as possible for constructing haiku database.

1．はじめに

これまでに、コンピュータを使った俳句研究の端緒として、季語データベースの構築とそれを利用して、俳句に詠み込まれた季語を機械的に判定する方法について検討してきた。

手始めとして、季語データベースとしては主要な季語があれば十分であると考え、角川書店の「新版季寄せ」[1]と成星出版の「現代歳時記」[2]を使用し、見出し季語をデータベース化した。そして、「新版季寄せ」収録の4,704 語と「現代歳時記」収録の2,371 語のうち、両方に共通し、かつ同じ季節を表すもの 1,542語を用いて、俳句に詠み込まれた季語を機械的に判定する方法について検討した[3]。

次に季語を増やすことにより、季語の判定率が改善できるかを検討した。季語の増補を容易に行う方法は別の歳時記や季寄せ（以下、季語集と称する）のデータを追加することである。そこで、文藝春秋「季寄せ」[4]（以下「山本季寄せ」と呼ぶ。）のデータを追加した。見出し語の総数は、3,790 語であった。これと先の「新版季寄せ」と「現代歳時記」との共通語を求めると、新たに1,463 語が増補できた。また、季語の増補により俳句の季語の判定率を65.4%から76.8%へと向上させることが出来た[5]。

季語集に共通する季語、つまり季語の積集合は、庶民の持つ季節感をピンポイントで理解する上にお

いて、有益であると考える。一方でどのような言葉が季語として取り上げられているのか季語の和集合に焦点を当てることは、世界観の広がりを実感するために重要であろうと考える。

そこで今回は、日本で一番季語数の多い季語データベースを構築することを目指し、これまでに行ってきた収集のプロセスと結果について報告する。

季語のデータベースは、単に季語の研究のためのみではなく、本来的には俳句研究（ここでの俳句研究は、俳句の文学的研究や歴史的研究ではなく、数量的な分析的研究を意図する）のための基礎データである。一方、俳句研究のためには、俳句のデータが不可欠である。そこで今回俳句データを収集する方策として、「俳句投句鑑賞システム」を開発したので、合わせてこれを紹介する。

2．季語データベースの構造と季語収集の過程

季語を収集した材料は、これまでの研究で扱った3タイトルに講談社の「大歳時記」[6]を加えた4つの季語集を用いた。今回は季語をできるだけ多く収集するために見出し季語のみを扱うのではなく、各見出し季語の下に記された異称・別名・異形・別字・同類季語あるいは他季や分類の異なる関連季語をすべて収集の対象とした。

すべての材料ははじめに電子化する必要がある。その際、できるだけ入力の手間を省くために様々な工夫を行ったが、基本データとして、見出し季語、季節、分類、類語及び関連語を1レコードするデータを作成した。このデータを基にして、表1の季語データベースのフォーマットに合わせて項目情報の

表1　データベース項目内容一覧

タグ番号	項目名	備　　考
1001	ID	半角数字8桁（出典コード3＋通番5）
1002	季語	出典コード
1003	仮名読み	現代仮名遣い　101 新版季寄せ
1004	歴史仮名読み	歴史的仮名遣い　102 山本季寄せ
1005	季節	複数生起　103 現代歳時記
1006	分類	複数生起　104 大歳時記
1007	意味	
1008	類語	複数生起、同じ意味に用いられる語など
1009	類語主見出し語	季語が類語の場合の親見出し季語
1010	関連語	複数生起、季が異なる語など
1011	関連語主見出し語	季語が関連語の場合の親見出し季語
1012	起源フラグ	見出し季語1、類語2、関連語3
1013	例句	複数生起
1014	備考	

充足を図った。4つの季語集はそれぞれ季節の区切りや分類の仕方、オリジナルの情報項目に関して充足度が異なるため、1季語集ずつ処理を行い、その後全データを統合して最終調整を図ることとした。次に各季語集毎の処理について記す。

2．1　講談社「カラー版新日本大歳時記」

5巻本の大判でカラー写真等をふんだんに取り入れたもので、4人の監修者の下、数十人の編集委員が関わって作成されたものである。

原本では、各ページのヘッダー情報として季節と分類が、各見出し季語について、季語、読み、歴史かな遣い（一部のみ）、季節（下位区分）、関連季語、関連季語読みの各情報が存在する。すべてのレコードについて、歴史かな遣い以外のすべての情報項目は満たされていた。データ入力作業は、まず見出し季語単位に各項目情報を入力した。次に関連季語を見出し語として独立させた。その際には、関連季語の項目情報は親見出し季語の項目情報を継承するものと仮定して可能な情報を付与した。見出し季語数は 4,100語、関連季語数は11,551語であった。

ここで、見出し季語は索引からも明らかなように語の重複がないことは証明されている。一方、関連季語は、いずれかの見出し季語である場合や、複数の見出し季語に採られている可能性がある。したがって重複チェックを行い、重複する季語はひとつにする作業が必要である。今回は、見出し文字列と読みのどちらかがユニークであれば単独の季語であるとし、両方とも同じであるものは重複チェックの対象とした。重複チェックの対象とした語は、その他の項目情報を見比べて、同じ語であるか独立した語であるかを判断することとした。合計15,651語のうち重複チェックの対象となったのは、628語（見出し季語起源：125語、関連語起源：503語）であった。

重複であるかどうかの判定基準と判定結果は次の通りである。

前提条件：見出し文字列と読みが同じ

判定条件1：季節同じ＋分類同じ＋共に関連季語の場合
・・・・1語立てて、見出し語起源を2つにする。（77語）

判定条件2：季節同じ＋分類同じ＋見出し季語と関連季語の場合
・・・・分類を見て、見出し季語を活かし、関連季語の見出し季語を関連季語に追加する。
（37語）

判定条件3：季節同じ＋分類異なる＋見出し季語と関連季語の場合
・・・・見出し季語を活かし、関連季語の見出し季語を関連季語に追加する。（35語）

判定条件4ａ：季節下位区分が別＋分類同じ＋見出し季語と関連季語の場合
・・・・見出し季語を活かし、関連季語の見出し季語を関連季語とする
判定条件4ｂ：季節下位区分が別＋分類同じ＋共に関連季語の場合
・・・・1語立て、季節を決定し見出し起源を2つにする。（29語）

判定条件5：季節同じ＋分類異なる＋共に関連語の場合
・・・・1語立て、分類を決定し、見出し起源を2つにする。（42語）

判定条件6ａ：季節下位区分が別＋分類異なる＋見出し季語と関連季語の場合
・・・・見出し季語を活かし、関連季語の見出し季語を関連季語とする。
判定条件6ｂ：季節下位区分が別＋分類異なる＋共に関連季語の場合
・・・・1語立て、季節を決定し、見出し起源を2つにする。（30語）
※2語（月氷る、秋容・・時候）の分類を変更。

判定条件7：季節異なる＋分類異なる＋見出し季語と関連季語の場合

・・・・見出し季語を活かし、関連季語の見出し季語を関連季語とする。（9語）

判定条件8：季節異なる＋分類異なる＋共に関連季語の場合

・・・・1語立て、季節と分類を決定し、見出し起源を2つにする。（15語）

判定条件9：季節異なる＋分類同じ＋共に関連季語の場合

・・・・見出し季語としたときの季節を決定できないので除外。（2語（初見草・切子））

判定条件10：上記判定条件に当てはまらず意味が異なる場合

・・・・2語を別見出季語とする。（73語（ながしは3語））

この結果、季語数は、15,373語となった。

2．2　文藝春秋「季寄せ」

山本健吉編集の上下2巻本の小冊子で携帯の利便性を考えられている。

原本では、各ページのヘッダー情報として季節（下位区分）と分類が、各季語には、季語、読み、歴史かな遣い（一部）、関連季語（ゴシック体で表記されている）、関連季語読み（一部）の情報が存在する。歴史かな遣いについてはごく一部の該当する漢字に対して付いているのみで欠落が大きい。関連季語の読みもすべての語に付いているのではなく、難解な単語または漢字についているのみで情報の欠落がある。また、本文中には季節の区切りのみで分類の区切りは存在しないので、一つのページに複数の分類の季語が跨がる場合には、ヘッダー情報を見ながら適宜分類の切れ目を判断する必要があった。データ入力作業は、まず索引の部に上げられた季語とその読み及び掲載頁を入力し（7,791 語）、掲載ページ情報から項目情報を付与した。次に本文中の関連季語に当たり、未入力の語をチェックして追加入力した（5,263 語）。読みの記載が無い語については、辞書等を駆使して読みを入力した後、見出し文字列と読みの一致で抽出した季語の重複チェックを行った。今回は関連季語の追加入力時点で重複をチェックしながら入力したため、重複のチェックを要する語数は少なかった。 312語がチェック対象となり、関連季語の親見出し季語を見出し季語の関連季語に集約するなどした結果、見出し季語3,791 語、関連季語9,144 語を得た。

2．3　成星出版「現代歳時記」

1巻本のハードカバーのもので編者は3名からなる。特徴として、現代俳句に対応するよう季節によらない物象感を表す言葉を集めた雑の部を設けている。今回はこの雑の部の言葉も収集対象とした。

原本では、各季節各分類ごとに項立てがある。見出し季語には現代仮名遣いで読みが付されており、各見出し季語の下に、関連する季語や関連語を類語として記載している。類語の読みはすべてについている訳ではなく、すべて漢字に読みが付けられてはいない。そこで、はじめに見出し季語について各項目情報とともに入力した（2,366 語）。次に類語を入力し、ページ番号により見出し季語と関連づけして見出し季語の項目情報を類語に反映させた。この時点での季語数は、11,421語となった。類語の読みを入力した後で、見出し文字列と読みの一致により重複チェックを行った。その結果、重複候補の128語について吟味した結果、入力ミスによる重複語の削除や季節や分類の判断による統合を行って75語を残した。最終的には見出し季語2,363 語、関連季語9,004 語、合計11,367語を得た。

2．4　角川書店「新版季寄せ」

角川書店編集の1冊もので、ポケット辞書のような体裁である。本文は季節・分類ごとに区分されている。漢字を含む見出し季語には読みがあり、見出し季語と読みは必ず存在している。各見出し季語の下には、異称・別名・同類の季語が記載され、さらに季節のマークと合わせて他の季で関連する季語等を載せている。

まず見出し季語と各項目情報を入力した。次に類語や関連語に、見出し季語の項目情報を継承させて

独立させた。見出し季語 4,704語のうちの 3,902語に延べ15,256語の類語が存在し、 1,906語に延べ 2,922語の他の季の関連語が付属していた。延べ季語数は、都合22,882語となった。先と同様に重複チェックを行ったところ、重複候補となった季語は、見出し季語 1,576語、類語季語 883語、関連季語 2,805語の合計 5,264語にのぼった。これは見出し季語が、他の季節の季語の関連語とされていたものを、両方を季語として独立させたために生じたと考えられる。そのため次の条件に照らして、最初に機械的に重複語を削除した。

1）見出し文字列が同じ＋季節・分類が同じ＋見出し季語と関連季語の関係にある場合

・・・・見出し季語の関連語と関連季語の親見出し季語が一致する場合に、関連季語を削除(500語)

2）見出し文字列が同じ＋季節・分類が同じ＋見出し季語と関連季語の関係にある場合

・・・・関連季語の親見出し季語を見出し季語の関連語に代入し、関連季語を削除(500語)

以上の処理で丁度千語チェック候補を減らすことができた。

独立した異なる季語であるということは、見出し文字列、読み、季節、分類のすべてが異なる場合は明らかである。そうではない場合、人による判断が必要になる。今回の処理のプロセスでは、見出し文字列と読みがユニークな場合にはすべて独立した季語として処理をするので、問題は見出し文字列と読みが一致する場合のみである。その場合に重複かどうかを判断する拠り所はその他の季節や分類などの項目情報である。それらを使って吟味した結果、見出し季語から4,703 語、類語と関連語から15,102語の合計19,805語を得た。

3．季語の統合と季語集の特徴

今回使用した4つの季語集から得られた季語の延べ合計は、表2の右下欄に示す通り 59,480 語であった。次にここから重複する季語を取り除く作業を行う。単独の季語である条件は、ここでも見出し文字列と読みがユニークであることを条件とした。条件に照らした季語の数を表2の単独季語、重複季語欄に示す。全体数に占める単独の季語数の割合が多い季語集は「現代歳時記」であった。この季語集の特徴として他の季語集では取り上げていない雑の部が約5,000 語あったので、この結果は理解できるものである。その次に単独の季語の割合が高い「新版季寄せ」は、掲載されていた季語数が一番多く、2番目に季語数の多い「大歳時記」よりも4千語以上多いためこれも想定内の結果と言える。他方、「山本季寄せ」と「大歳時記」は単独の季語の割合が少ないが、これは言い換えれば共通する季語を収録しているということで標準的な季語集としては望ましいことと言えよう。

表2　季語集別統計

季語集		見出し語起源	関連語等起源	重複季語	単独季語 (%)	合計
1	新版季寄せ	4,703	15,102	10,602	9,203 (46.5)	19,805
2	山本季寄せ	3,791	9,144	10,455	2,480 (19.2)	12,935
3	現代歳時記	2,363	9,004	5,705	5,662 (49.8)	11,367
4	大歳時記	4,100	11,273	11,737	3,636 (23.7)	15,373
延べ合計		14,957	44,523	38,499	20,981 (35.3)	59,480

さて、残る重複の可能性のある語についてであるが、4つの季語集を統合したために4つの季語集す

べてに取り上げられている季語もあるはずであり、3つあるいは2つの季語集に取り上げられていた季語もあるはずである。これが重複の可能性のひとつである。それとは別に同音異義語としてたまたま他の季語と見出し文字列と読みが一致したために、ここに混在しているものがあると考えられる。ただし、後者はそれぞれの季語集の季語の重複を吟味した際に既に同音異義語あることは分かっている。そのため異なる季語集に採られている限りにおいてはそれは前者の単純重複であると言える。そこで同じ季語集に採られている同音異義語が混じる重複候補のみを抽出し、目で見て判断することにした。444 語の候補が抽出され、その内 101語が単独の季語で、残りはいずれか複数の季語集に採られた122 語と判明した。複数の季語集に採られた季語が何語存在したかを表3にまとめた。うち見出し季語と書かれた欄には、見出し季語のみについて重複していた語数を示す。さらに季語集間の重複語数を表4に示す。

表3　重複採録のパターンと季語数

採録季語集				全	見出し
1	2	3	4	季語数	季語数
○	○	○	○	3,255	1,199
○	○	○	×	237	52
○	○	×	○	2,909	980
○	×	○	○	566	125
×	○	○	○	524	167
○	○	×	×	1,050	290
○	×	○	×	513	95
○	×	×	○	1,809	494
×	○	○	×	158	37
×	○	×	○	2,114	432
×	×	○	○	280	51
合計				13,415	3,922

表4　季語集間で重複する季語数

	1	2	3	4	
1	——	7,451	4,571	7,973	全季語
2	2,521	——	4,174	8,802	
3	1,471	1,455	——	4,625	
4	2,798	2,778	1,542	——	
	見出し季語				

4つの季語集のすべてに採録された数が最も多く、次が「現代歳時記」を除く3つに採録された季語、3番目が「山本季寄せ」と「大歳時記」に採られた季語が多い。見出し季語についてもすべての季語集に採録された季語が最も多く、次に多いのは「現代歳時記」を除く3つに採られたものであった。各季語集間での季語の重複（表4）をみても、「現代歳時記」は他の季語集とは一線を画している様子がうかがえる。

最終的に単独の季語20,981語（これには最終チェックで単独の季語と判定した 101語の同音異義語を含む）と重複を整理した13,415語の合計34,395語の季語データベースを構築することができた。

4．「俳句投句鑑賞システム」の開発

俳句を研究するためには、材料となる俳句が必要がある。しかもできるだけ多くの俳句を収集したいと考えた。一つの方法は既に電子化されたデータを利用することである。例えば、集英社の定本である古典俳文学大系のＣＤ－ＲＯＭ版[7]は、室町、江戸期の主要な撰集等を網羅した約25万句が納められており、同梱の独自の検索システムで検索が可能になっている。しかしながらデータだけを取り出して二次的に加工・分析することはできない。ホトトギス電子新歳時記[8]、俳句嚢[9]にも１万数千の俳句が納められており、検索表示が可能であるが、これを一括して取り出して利用することはできない。

簡単に利用できる電子データが無いのであれば、自ら入力することになる。その場合に、他人の作品を入力するとなると著作権の問題に引っかかる可能性がある。没後50年が経過している作者であれば問題はない訳であるが、1956年以降に亡くなった方並びに存命の方の作品は、著作権に対する配慮を要する。

そこで一計を案じ、作者に自発的に俳句を入力してもらいそれを蓄積するシステムを開発し公開することを考えた。現状を調査してみたところ、インターネットの普及に伴って俳句を愛好する人のホームページは沢山できていた。いくつかのサイトでは投句を受け付けるところもある。ホームページを作っているのは個人ばかりではなく、例えば現代俳句協会やいくつかの俳句結社でも、ホームページに投句を受け付ける仕組みを持っているところがあった。その多くはＢＢＳを使って投句するもの[10]であったり、メールで投句するもの[11]もあった。中には自前のシステムを作っていると思われるところ[12]もあった。ただし、そういうところは会員登録が必要で、投句料を課するところ[13]が多いのは必然的な理由があるものと思われる。

今回開発するシステムの目的は、俳句の収集である。そのため開発のコンセプトとして、(1) 利用者が容易に入力できること、(2) 入力した俳句の削除が簡単に行えること、(3) 俳句の閲覧が簡単に行えることの３つの要件を満たした上で、俳句の分析の際の利用価値を高めるために必要と考える項目を付加し、また、遊び心と競争心を高める仕掛けを付けることとした。システムの画面遷移は図１のようになる。

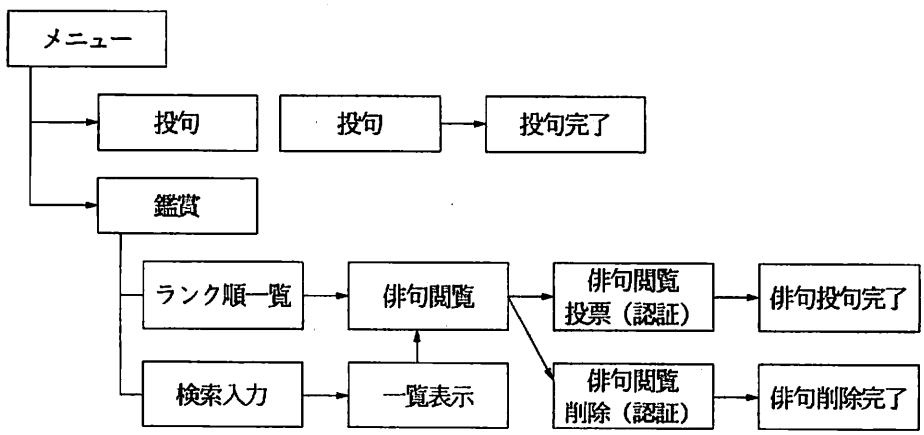

図１　俳句投句鑑賞システム画面フロー

投句画面での入力項目は、俳句、カナ読み、季語、四季、詞書きとし、投句者の情報として作者名、メールアドレス、年齢、男女の別を入力する。入力された情報は、自動的にサーバ上のPostgreSQLに蓄

積される仕組みになっており、表示画面では、PostgreSQL上のデータベースを検索して結果を表示する。

表示機能は、ランク順表示と検索画面を用意した。ランク順表示は、データベース中に蓄えられた俳句を投票数の多いもの順に表示するものである。検索画面では、作者、キーワード、四季の3つの項目を組み合わせて検索が可能である。ランク順表示画面及び検索結果一覧画面から、一つの俳句を詳細表示する俳句閲覧画面に移り、投票と削除を可能にしている。投票は気に入った俳句に一票を投じる仕掛けで一人で一句に複数票の投票はできないようにしている。ただし一人で投票できる句の数は制限していない。俳句の削除に関しては、投句した本人の認証が必要でる。

5．おわりに

季語データベース構築のためのデータ整理作業では各季語集の特徴が分かるなどいくつかの発見があった。戦前より前に作られた俳句を鑑賞あるいは分析するためには、歴史仮名遣いが必要となる。そのため、季語の歴史仮名読みを項目として採用したがオリジナルに付与された語数は多くなかったため、すべての季語に付けられてはいない。これに関してはできるだけ早い時期に整備したいと考える。また、同じ見出しで読みが異なる場合、今回はこれを機械的に別の季語としたが、専門家の意見を伺って対応を考えたいと思っている。

「俳句投句鑑賞システム」については、仕掛けはできたので、今後はいかにＰＲをして多くの人の投句を促すかが課題である。当面は魅力的なホームページを作って知名度を上げていく方向である。

以上の準備によって俳句研究の進展の一助となることを念願している。

【引用文献・注】

1) 角川書店編，「新版季寄せ」，角川書店，1996年2月20日第13版

2) 金子兜太，黒田杏子，夏石番矢編，「現代歳時記」，成星出版，1998年10月18日改訂版

3) 吉岡亮衛，「季語データベースの構築と俳句の季語の自動判定の試み」，情報処理学会研究報告，Vol. 2000，No. 100，pp. 57-64，2000年10月27日

4) 山本健吉編，「季寄せ　上巻・下巻」，文藝春秋社，1973年10月5日

5) 吉岡亮衛，「季語データベースの構築と俳句の季語の自動判定の試み(2)ー季語の増補と判定率の向上ー」，情報処理学会研究報告，Vol. 2001，No. 6，pp. 17-24，2001年1月19日

6) 飯田龍太，稲畑汀子，金子兜太，沢木欣一監修，「カラー版　新日本歳時記　全5巻」，講談社，1999年12月～2000年6月

7) CD-ROM版編集委員会編，「古典俳文学大系(CD-ROM 版)」，集英社，2004年

8) 稲畑汀子編，「ホトトギス電子新歳時記」，三省堂，2000年

9) 俳句發編集委員会編，「俳句發」，日外アソシエーツ，2002年

10) 例えば、俳句雑誌「水煙」投句箱(http://www3.ezbbs.net/14/suien8/)、俳俳本舗　本店(http://otd11.jbbs.livedoor.jp/1105557/bbs_plain)、など。

11) 例えば、俳句ランド(http://www.asint.jp/~fuchi/)、ネット句〔北側松太のホームページ〕(http://www5b.biglobe.ne.jp/~ matu0909/kukai/touku.html)、など。

12) 例えば、現代俳句協会(http://www.gendaihaiku.gr.jp/haikukai/index.htm)、句会桃季(http://www.interone.jp/~ touri/cgi-bin/kukaibody.htm)、など。

13) インターネット俳句会(http://www.haiku.jp/new_haikukai/index.html)、俳句ステーション(http://www.haiku-st.co.jp/index.html)、など。