人文科学とコンピュータ 49-3
(2001.1.19)

# 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)
## －季語の増補と判定率の向上－

吉岡　亮衛
国立教育研究所

本論は、コンピュータによる俳句の研究を行うために必要な俳句データベースと、俳句を分析するために必要な季語データベースの本格的な構築に先立ち、季語データベースの構造とデータベースに収録すべき季語の数を検討した結果を報告するものである。

具体的には、3種類の季語を集めた本を材料として、①共通に存在する季語、②すべての見出し語である季語、③見出し語の異称・別名・同類・対象語を含めたもの、の3通りの季語集合を作り、それらを用いて、サンプルとして抽出した俳句の季語を特定することを試みた。その結果、先の研究で 1,542語の季語で 448句の俳句を分析した結果、全体の約65%の俳句の季語を特定することができたものが、2,901 語の季語により344 句（76.8%）、6,709 語で 399句（89.1%）、約2万語で 420句（93.8%）まで、判定率が向上することが見いだされた。また、最後まで季語が特定出来なかった俳句について、改善の方策を検討した。

To Build a Kigo-database and a Trial to Specify the Kigo for the Haiku Automatically (2)
- Improving the Specification Rate by Enlarging the Kigo-database -

Ryoei Yoshioka
National Institute for Educational Research

This paper reported the investigated results about the amount of Kigo in Kigo-database. This investigation is needed for building Kito-database, that is useful to analyse the Haiku. A Haiku-database and a Kigo-database are both needed to study Haiku by computer.

Concretely, tree types of Kigo-group are made from three different books of "Kigo". The first one is the common Kigos in books, the second one is the all different Kigos in books, and in the third one are included more broader terms in the book "Shinhan-Kiyose". As a result of the Kigo matching test of all 448 Haikus, at the last time 1,542 Kigos hit about 65%. Then 2,901 Kigos hit 344 Haikus(76.8%), 6,709 Kigos hit 399 Haikus(89.1%), and about 20,000 Kigos hit 420 Haikus(93.8%). The specification rate of Kigo is improved by Enlarging the Kigo-database. At last the reasons of unmatched Haikus are discussed.

## 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見つけ出す方法について検討してきた。

第一段階として、季語データベースの構築に角川書店の「新版季寄せ」[1)]と成星出版の「現代歳時記」[2)]を取り上げその共通語を基に、俳句の季語の判定率を検討した。具体的には「新版季寄せ」収録の見出し語　,704語と「現代歳時記」収録の見出し語　,371語に共通して採られており、かつ同じ季節を表すものとされた　,542語を季語データベースとした。これを用いてサンプルとした第52回芭蕉祭献詠俳句一般の部448 句[3)]を分析したところ、最終的に293 句（65.4%）の季語を特定することができた[4)]。

そこで今回は、季語数を増やすことにより、季語の判定率がどのように改善されるかを検討することを目的として研究を行った。

## 2　季語の増補 I

### 2.1　文藝春秋「季寄せ」

前回と同様、まずより一般的と考えられる季語を増補することを目指し、今回は文藝春秋の山本健吉編「季寄せ」[5)]と先の角川「新版季寄せ」との共通語を追加することとした。

「季寄せ」は1973年の初版であるが、平成12年 7 日 1 日に第十三刷を重ねているロングセラーである。本書の特徴は、季語の分類について、旧来の時候・天文・地理・人事・宗教・動物・植った分類法を廃除していること、一年を春・夏・秋・冬・新年とするカテゴリと、更に四季を三春・初春・仲春・晩春のように下位区分するカテゴリを設けていることである。ここで、三春とは初春、仲春、晩春の三月にまたがる季語が該当する。また、冬には特に歳末という分類が設けられている。ただし、季節以外のカテゴリが定められていないため、本書のみでは高度な分析を行うことは不可能である。そこで、季節以外のカテゴリは角川書店「新版季寄せ」に習うこととし、二種類の季寄せの共通語を取り出すこととした。

見出し語の分類毎の頻度は、表 1 の通りである。見出し語の総数は、3,790 語である。季節ごとの季語の数でみると夏が最も多く、 3 割強を占め、春、秋、冬がそれぞれ 2 割、新年が 1 割弱となっている。これは、角川書店「新版季寄せ」とほぼ同じ分布であった。

表1　文藝春秋「季寄せ」のカテゴリ別季語数

| | 春 | 夏 | 秋 | 冬 | 新年 | 計 |
|---|---|---|---|---|---|---|
| 三＊ | 203 | 510 | 204 | 386 | 258 | 1,561 |
| 初＊ | 74 | 182 | 179 | 84 | | 519 |
| 仲＊ | 191 | 208 | 134 | 41 | | 574 |
| 歳末 | | | | 76 | | 76 |
| 晩＊ | 303 | 313 | 286 | 158 | | 1,060 |
| 計 | 771 | 1213 | 803 | 745 | 258 | 3,790 |

## 2.2　見出し語の重複

漢字見出しによる重複は4語認められた（表2）。そのうち、『初春』は読みが異なり別の語の誤一致である。『事始』と『針供養』は、地域により時期が異なるためにわざと見出し語が立てられたものである。唯一『防風の花』は、編者の責任に帰されるべきもので、このままでは季節が特定できないため除外されるものである。

表2　文藝春秋「季寄せ」における見出し語の重複

| 見出し語 | 読み | 季節 | 意味、備考 |
|---|---|---|---|
| 事始 | ことはじめ | 春<br>冬 | 東日本では、旧12月8日の事納に対する旧2月8日<br>関西では12月13日 |
| 初春 | はつはる<br>しょしゅん | 新年<br>春 | 正月の意味<br>二月 |
| 針供養 | はりくよう | 春<br>冬 | 二月八日：関東<br>十二月八日：関西 |
| 防風の花 | ぼうふうのはな | 夏<br>秋 | 浜防風の花<br>浜防風の花 |

## 2.3　「季寄せ」と「新版季寄せ」のマッチング

文藝春秋の「季寄せ」と角川の「新版季寄せ」の季語のマッチングを取り、共通の語を増補語彙とすることにした。そのため、語の見出しのマッチングを取ったところ、文藝春秋の方は旧字体を使用しており、多くの語について直接マッチングにかからないことが分かった。そのため、読みでマッチングした語について、目で吟味して拾い上げることとした。

最初に、見出し、読み両方でマッチした語は1,831 語あり、そのうち季節が一致しないもの44語を除いた残り1,787 語を抽出した。次に読みで一致した語は、759 語あった。そのうち季節が異なるものをまず機械的にふるい落とした。さらに同音異義語がある場合に誤マッチングが生じる可能性があるため、同音異義語をリストアップしてチェックした。その結果、表3の19の読みについて、同音異義語が見いだされた。両方に存在する語以外は削除して追加は723 語となった。これらについては目によるチェックで同じ意味を表す漢字が用いられていることを確認している。

表3　角川「新版季寄せ」と文藝春秋「季寄せ」の同音異義語

| 読み | 角川 | 文藝春秋 |
|---|---|---|
| あきのうみ<br>あきのひ<br>いなだ | 秋の海、秋の湖<br>秋の灯、秋の日<br>稲田、いなだ | 秋の海<br>秋の燈、秋の日<br>いなだ |

| | | |
|---|---|---|
| かき | 柿、牡蠣 | 柿、牡蠣 |
| かんごえ | 寒肥、寒声 | 寒肥、寒声 |
| かんとう | 竿燈 | 寒燈 |
| かんらん | 甘藍、寒蘭 | 寒蘭 |
| たこ | 凧、章魚 | 凧 |
| なつのひ | 夏の日、夏の灯 | 夏の日 |
| はつごえ | 初声、初肥 | 初肥、初声（はつこえ） |
| はるかぜ | 春風邪 | 春風 |
| はるのうみ | 春の海、春の湖 | 春の海、春の湖 |
| はるのかぜ | 春の風 | 春の風邪 |
| ふすま | 襖、衾 | 襖 |
| ふゆのうみ | 冬の海、冬の湖 | 冬の海 |
| ふゆのひ | 冬の日、冬の灯 | 冬の日 |
| ほととぎす | 時鳥、杜鵑草 | 時鳥 |
| れいか | 零下、冷夏 | 冷夏 |
| わた | 綿、棉 | 綿 |

以上の結果、文藝春秋の「季寄せ」と角川の「新版季寄せ」の共通語は2,510 語となった。カテゴリ別の語数は表 4 の通りである。

表4　角川「新版季寄せ」と文藝春秋「季寄せ」の共通語

| | 春 | 夏 | 秋 | 冬 | 新年 | 計 | |
|---|---|---|---|---|---|---|---|
| 時候 | 37 | 30 | 33 | 40 | 14 | 154 | 6.1 |
| 天文 | 33 | 53 | 52 | 31 | 9 | 178 | 7.1 |
| 地理 | 22 | 27 | 17 | 20 | 4 | 90 | 3.4 |
| 人事 | 119 | 232 | 109 | 218 | 113 | 791 | 31.5 |
| 宗教 | 35 | 35 | 38 | 40 | 29 | 177 | 7.1 |
| 動物 | 85 | 156 | 70 | 62 | 5 | 378 | 15.1 |
| 植物 | 180 | 266 | 216 | 75 | 5 | 742 | 29.6 |
| 計 | 511 | 799 | 535 | 486 | 179 | 2510 | |
| | 20.4 | 31.8 | 21.3 | 19.4 | 7.1 | | |

### 2.4　季語の増補による判定率の向上

文藝春秋の「季寄せ」と角川の「新版季寄せ」の共通語2,510 語のうち、前回使用した角川の「新版季寄せ」と成星「現代歳時記」の共通語を除いたものを季語の増補分として、前回の季語判定用サンプ

ル俳句の中の季語未判定句を調べることとした。

「新版季寄せ」と「現代歳時記」の共通語は、下図のＦ＋Ｇに当たる1,542 語あった。これと文藝春秋の「季寄せ」と角川の「新版季寄せ」の共通語との共通語は、３種類の季語集の共通語となるが、これは1,151 語（Ｇ）であった。したがって、2, 510－1,151 の1,359 語（Ｄ）が増補分となる。他方、前回448 句のうちの未判定句は155 句であった。

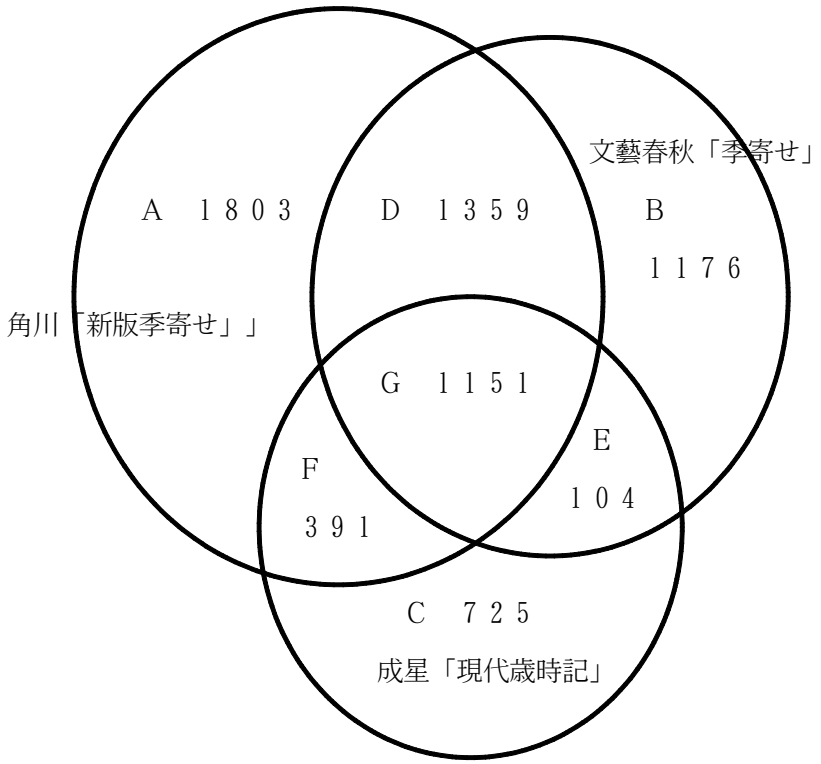


図1　角川「新版季寄せ」、文藝春秋「季寄せ」、成星「現代歳時記」の季語の集合関係

増補分の季語を使用し155 句に対して、前回と同様に見出し語と読みでマッチングを取った結果、見出し語では、１語マッチした句が35句、２語マッチした句が６句、24語マッチした句が１句であった。見出し語ではマッチしなかった句のうち読みでマッチした句は、１語：31句、２語：23句、３語：23句、４語：17句、５語：３句、６語：４句、12語：１句であった。その他にまったくなにもマッチしなかった句が11句あった。

マッチした語が意味をなす語であるかどうかを目で調べたところ、見出し語でマッチした句42句の内41句が、読みでマッチした 102句のうちの15句が残され、季語の可能性のある語が57語あった。さらに一語ずつ句に照らして吟味した結果、38句はまさしく季語に該当しており、13句は語が季語の一部をなしていると判断された。残り５句の６語については季語には該当しないことが明白であった。

したがって、この段階では、155 句のうちの51句の季語を特定できたと考えられる。これを前回の結果と合わせて考えるならば、2,901 語の季語により344 句（76.8%）の季語が特定できたことになる。

### 3　季語の増補Ⅱ

ここまでは、数多くある季語のうちで誰もが季語と認め、かつ、季語が表す季節に揺らぎが無いもの

を一般的な季語として収集してきた。次に、季語をできるだけ多く集めて判定率が向上するかを調べることにする。

### 3.1　入力済全見出し語による判定

まず、既に入力が終わっている、角川「新版季寄せ」、文藝春秋「季寄せ」、成星「現代歳時記」に採録されている季語すべてについて見てみる。すでに、ＦＧの季語での判定は前回に報告しており、Ｄの増補による結果は、上述の通りである。そこでここでは更にＡＢＣＥの季語を用いての判定結果を調べる。追加される季語の数は、3,808 語となる。

また、季語判定の対象となる俳句は、先の判定によっても季語が特定できなかった104 句である。Ａの季語は読みが付いているため見出し語のみでなく読みでのマッチングも求めることができるが、その他は読みが付いていないため見出し語によるマッチングのみとなる。その結果、マッチした季語数の分布は次のようになった。見出し語によるマッチでは、１語（34句）、２語（34句）、３語（７句）、４語（14句）、５語（１句）、６語（２句）。読みでは、１語（１句）、２語（３句）、３語（１句）、４語（１句）、５語（１句）、６語（２句）、７語（１句）、10語（１句）。

それぞれの語を俳句に照らして吟味した結果、季語と見なされるもの24語（24句）、季語の一部と見なされるもの31語（31句）の55句に季語を特定できたと考えられる。結局、さらに3,808 語の増補（都合6,709 語）により55句の季語が特定され、判定率は89.1％となった。

### 3.2　角川「新版季寄せ」の類語による判定

角川「新版季寄せ」は、見出し語4,704 語に対して、それぞれの季語の①異称・別名・異形・別字、②応用形・活用形、③同類・相似・関連または対照的な季語、④他の季における季語、⑤植物の季語については、花を主とする季語には果実・結実の期の季語を、果実・結実を主とする季語については花の期の季語、が記載されており、これらをすべて季語として扱うことができる。そこで、これを見出し語の季語とそれぞれの注釈に基づいて分類した結果、19,964語の季語を得た。そのうち既に分析に使用した本見出し4,704 語を差し引いた15,260語については、表のように分類される。

表5　角川「新版季寄せ」の類語のカテゴリ別季語数

| | 春 | 夏 | 秋 | 冬 | 新年 | 計 |
|---|---|---|---|---|---|---|
| 時候 | 205 | 233 | 216 | 273 | 128 | 1055 |
| 天文 | 190 | 307 | 385 | 333 | 38 | 1253 |
| 地理 | 142 | 162 | 128 | 163 | 11 | 606 |
| 人事 | 710 | 1298 | 769 | 1291 | 770 | 4838 |
| 宗教 | 236 | 288 | 282 | 233 | 120 | 1159 |
| 動物 | 566 | 925 | 510 | 337 | 2 | 2340 |
| 植物 | 1037 | 1493 | 1095 | 339 | 45 | 4009 |
| 計 | 3086 | 4706 | 3385 | 2969 | 1114 | 15260 |

これを用いて、残る49句の季語判定を行った。判定方法は、見出し語のみのマッチングによった。その結果、マッチした季語の分布は、1語（16句）、2語（8句）、3語（5句）、23語（1句）、19句はマッチングする語はなかった。季語の候補を吟味した結果、最終的には、22語（21句）の季語が特定された。

ここまでで、全448 句のうちの420 句（93.8%）の季語が特定されたことになる。ここまでの俳句の季語の増補による季語の判定率は表6のようにまとめられる。最終段階は、角川の「新版季寄せ」に採録されている季語総数が、19,964語であったことから、他の文藝春秋「季寄せ」（B）及び成星「歳時記」（C）の季語との重複季語を調査していないため、約20,000語としている。

表6　季語数と判定句数の関係

| 季語数 | 判定句数（%） | 未判定句数 |
|---|---|---|
| 1,542 | 293 (65.4) | 155 |
| 2,901 | 344 (76.8) | 104 |
| 6,709 | 399 (89.1) | 49 |
| 約20,000 | 420 (93.8) | 28 |

## 4　季語の未特定句と考察

次に挙げる28句が今回最後まで季語が特定できなかった句である。便宜的に五七五に区切っており、下線部が季語と見なされる語である。また、括弧内には選者を示す。2)、6)、24) 、27) の季語は今もって不明であり、無季の句とみなされるものである可能性がある。

1)燃え尽きて　闇深めけり　送り舟　（井澤正江）
2)子育ての　一区切せし　ひからかさ　（稲畑汀子）
3)悴かみて　貼りし切手の　歪みたる　（稲畑汀子）
4)虚子舐めし　ひやし飴とや　吾も口に　（稲畑汀子）
5)終戦忌　語り部呆けの　淵に立ち　（金子兜太）
6)終の家　芭蕉掛軸　吾れと古り　（金子兜太）
7)ブラジルに　老のともしび　芭蕉祀る　（金子兜太）
8)死地に赴く　最後の遺稿　敗戦忌　（金子兜太）
9)終戦忌　遺児と呼ばれて　髪白く　（金子兜太）
10)噛締める　ビーチバレーの　奥歯かな　（金子兜太）
11)母の忌の　トンボよ母に　似て小さし　（金子兜太）
12)黒豆を　土産に丹波　杜氏来る　（草間時彦）
13)心経を　写せる筆を　洗ひけり　（澤木欣一）
14)曲がり家に　馬の匂はず　はせをの忌　（鷹羽狩行）
15)出穂の田に　激しき雨の　伊賀盆地　（鷹羽狩行）
16)除草剤　撒くや巣籠る　鳥翔たせ　（鷹羽狩行）
17)空海の　飛錫に裂けし　袈裟曝す　（早崎明）

18)曝す中に　香象着座　なす一基　　（早崎明）
19)流灯の　帯となる灯に　笠置山　　（早崎明）
20)静かなる　日の続きをり　田は出穂に　（早崎明）
21)田の神を　祭る（べつばう）ゐる暗さ　（早崎明）
22)蒼求を　焚く刀匠の　白たすき　　（堀口星眠）
23)敗戦を　知らぬ遺影や　終戦忌　　（松崎鉄之介）
24)浜風や　潮が染めたる　チカの艶　　（松崎鉄之介）
25)赤ワイン　飲み赤バラの　闇迫り　　（丸山海道）
26)合歓閉じる　眠れば夢の　展がらむ　（丸山海道）
27)名を呼べば　鰐がかお出す　池の端　（丸山海道）
28)威銃　伊賀の旅人　おどろかす　　（森田峠）

これらの句に含まれる季語から、コンピュータによる季語の特定についての課題や難点を見いだすことが出来る。ひとつは、3)、13）、16）に見られるように活用語の活用形を考慮する必要がある。次に5)、8)、9)、23）に見られるものは、通常の忌日ではなく、『終戦記念日』が季語として採られている。これを誤用とするか、バリエーションとするかは専門家の判断を仰ぐひつようがある。3点目として、11）、25）はカタカナ書きされていたためにマッチングがとれなかったもので、見出し語の読みはひらがなとカタカナで持つ必要がある。また、26）の『合歓』は、『合歓の花』『合歓の木』は季語として存在するが、その省略語がマッチングされなかった例となる。この点については、季語を一語ずつ吟味していかざるを得ず、またそのための専門的知識も必要となる。4)の『ひやし飴』の『ひや』あるいは『ひやし』は、夏の季語をつくる接頭辞的な語がくっついた語であり、1）の『送り舟』の『送り』は盆に関係する季語を作る接頭辞と見なせる。この類のことばを整理して季語データベースに持つ必要がある。10）のビーチバレーは新語であり、機械的には判断できないものであり、臨機応変に追加していく必要がある。21）は、JIS にない漢字の熟語が該当し、そこに括弧書きしたるびがついていたが、これが誤りであったためにマッチングが取れなかったものであり、本来の『まくなぎ』がるびとして振られておれば、読みでマッチングが取れていたものである。同様に、22）は『蒼朮を焚く』が正しく、テキストの誤植のためにマッチングが取れなかったものである。

活用形の扱いとカタカナ表記等は、対応が可能であると考えられるが、テキストの誤植や新語については、最終的に目で見て判定する以外にはなさそうである。季語データベースの語彙としては、判定率も93% まで上げることができ、取り敢えずこれまでに収集した季語を整備することで俳句研究の準備段階としては十分であろうと考えられる。

## 文　献


1）角川書店編，「新版季寄せ」，角川書店，1985年 5 月10日
2）金子兜太，黒田杏子，夏石番矢編，「現代歳時記」，星成出版，1998年10月18日
3）芭蕉記念館，「第五十二回芭蕉祭献詠句集」，㈶芭蕉顕彰会，1998年10月10日
4）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
5）山本健吉編，「季寄せ　上巻・下巻」，文藝春秋，1973年10月5 日



※　本研究は、科学研究費補助金萌芽的研究（№11878028）の助成に負っている。