人文科学とコンピュータ　48－8
（２０００．１０．２７）

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

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、２種類の季語を集めた本を材料として、共通に存在する季語を取り出し、それらの季語を用いて、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。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。

加えて、季語データベースと俳句データベースを用いて可能となる数量的研究の一部を紹介する。

To Build a Kigo-database and a Trial to Specify the Kigo for the Haiku Automatically

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, at first the common appeared Kigo in two books are selected. And it is tried to specify Kigos of sample Haikus with those Kigos. About 65% of 448 Haikus are specified with 1,542 Kigos.

Additionally an example of a possibility of quantitative research about Haiku by using a Kigo-database and a Haiku-database is introduced.

1　はじめに

久しく俳句はブームである。俳句年鑑[1)]によれば、俳誌数は832 誌にのぼっている。これらは結社や同人が定期的または不定期に発刊する俳句集である。また、主要新聞には俳句のコラムがありそこに寄せられる俳句も多い。

松尾芭蕉の生誕地、三重県上野市では毎年秋の芭蕉忌に由来して芭蕉祭を催している。芭蕉祭の行事のひとつに献詠俳句がある。第53回の芭蕉祭には、一般13,934句、児童・生徒の部には33,458句が寄せられている[2)]。このように多くの俳句が寄せられる背景には、市民生活に根ざした大衆芸術としての側面と、表現力と理解力を培う国語教育の目標にかなう教育的な側面があるからだと思われる。

一方、文学研究の観点では、これまでに故人となった俳人の紀行文や句集の研究が多くなされてきて

いる。また、良い句を詠むための俳句作法の研究も行われている。さらに最近では、数量的分析手法を用いて特定の観点を検証する研究も見られる。しかしながら、俳句は既に膨大な量が存在し、かつ毎年毎年何万何十万と産み出されるということもあり、データベース化してそれを研究しようという試みはまだ無いようである。

そこで本研究は、俳句データベースを構築するための予備的な研究として、季語データベースの構築のための要件を明らかにすることを目的とした。また、試験的に構築した季語データベースを用いた俳句研究の可能性を示す。

2　季語データベース

2.1 「季寄せ」と「歳時記」

簡単に専門用語の解説をしておく。俳句に読み込まれる『季語』は、同義語として『季題』とも言われる。季語の集成は「季寄せ」と呼ばれ、「歳時記」は季語の解説を主としていると言われる。本論では断りの無い限り、俳句に読み込まれた季節のテーマを表す言葉を『季語』と呼ぶことにする。

2.2 季語データベースの構想

季語データベースは、俳句に詠み込まれた季語をデータベース化するものである。データベース化に当たっては、それを用いての様々な研究利用が可能となるよう、季語に付属する情報項目を同時にデータベース化する必要がある。

当面考えられる項目として、季語の漢字見出し、かな見出しの他に、それが表す季節は必須の項目である。そこで季語データベースの材料として、これらの項目を備えた角川書店の「季寄せ」[3)]を材料とした。「季寄せ」は、一般的な歳時記本に比べてコンパクトで語数が多く、また、既にある程度の構造化がなされているためデータベース化し易いと考えた。

ただし、「季寄せ」や「歳時記」本は、編者によってそれぞれ分類カテゴリに特徴があり、分類カテゴリについて検討しておく必要がある。入手出来た「季寄せ」や「歳時記」本についてのカテゴリについては、付録「季語の分類カテゴリ」に示す。

「季寄せ」のデータベース化は、1997年度科研費研究成果公開促進費「人文科学研究データベース」（人文科学研究データベース作成委員会，委員長：及川昭文）の一部として行った。

3　季語の検討

言葉は生き物であるというのは、季語にも当てはまる。今ではほとんど季語として使われなくなった言葉がある一方で、新しく季語と認められる言葉が出てくる。したがって、季語データベースの構築には終わりがないように思われる。そこで、当面の到達目標として収集季語数の目標値を定めたいと考え、一定数の季語を用いて、一定数の句の季節を特定することを通して、季語データベースに登録すべき季語について検討を試みた。

「季寄せ」データベースの季語は、編者の採録意図等によりある程度の偏りがあると考えられる。そこで、別途星成出版の「現代歳時記」[4)]（以下「歳時記」と表記する）の季語を入力し、両者に共通の季語を検討材料とすることで、検討材料の平均化・一般化を図ることとした。

3.1 「季寄せ」の分析結果

季語数：４，７０４

ＪＥＦにない漢字を含む季語数：６９
漢字見出しで複数回採られた季語：９

ながし	夏・天文	九州で梅雨のころに吹く西南風
	夏・人事・遊楽	夏の夜、花街や料亭のあたりを流して歩く新内ながし
山王祭	春・宗教・神道	今は四月十四日、滋賀県日吉（ひえ）神社の祭礼
	夏・宗教・神道	六月十五日ごろ、東京日枝神社の祭礼
初春	春・時候	しょしゅん（春の初め）
	新年・時候	はつはる（新年のこと）
針供養	春・人事・行事	二月八日：関東
	冬・人事・行事	十二月八日：関西
切山椒	春・人事・食	生菓子
	新年・人事・食	蒸した正月用の菓子
雪祭	春・人事・行事	二月に札幌で行われる
	新年・宗教・神道	一月十四日長野県伊豆神社の神事
草石蚕	夏・植物・野菜	
	新年・人事・食	正月料理
稗蒔き	夏・人事・農耕	
夏・人事・遊楽	水盤などに綿を敷き水に浸して稗を蒔き、発芽を観賞	
橙	秋・植物・果樹	
新年・植物	新年の飾り	

『初春』は、漢字見出しでは重複して採録されているように見えるが、読みと意味は異なる別々の季語であった。

3.2 「現代歳時記」の分析

季語数：２，３７１
ＪＥＦにない漢字を含む季語数：２６
複数回採られた季語：５

一月	一月・時候	いちがつ
	雑・天文	ひとつき
花	四月・植物	桜の花のこと
	雑・植物	植物の有性生殖の器官
競馬	五月・行事	くらべうま、京都賀茂別雷神社の神事
	雑・文化	けいば
針供養	二月・行事	関東
	十二月・行事	関西
橙	十月・植物	果実
	新年・植物	正月の飾り

同書の凡例部分には、目次には見出し語合計2,364 語を配したと記載されているが、新年の季語で、「四日　五日　六日」となっているものを３つの見出し語として数えたこと、本文中の見出し語で目次に漏れているものがあり、本来の季語数は上に挙げた通りである。

3.3 「現代歳時記」と「季寄せ」の一致度分析

見出しでマッチングした季題数：１５７０（６６%）

季節が異なるもの：２８

見出し語	「季寄せ」	「歳時記」
えんぶり	新年（行事）	二月（生活）
かまくら	新年（行事）	二月（生活）
なまはげ	新年（行事）	十二月（行事）
襖（ふすま）	冬（住）	雑（社会・生活）
夏蜜柑（なつみかん）	春（植物）	七月（植物）
駒鳥（こまどり）	夏（動物）	三月（動物）
恵比寿講（えびすこう）	冬（宗教）	十月（行事）
鯨（くじら）	冬（動物	雑（動物）
御命講（おめいこう）	冬（宗教）	十月（行事）
山雀（やまがら）	夏（動物）	十月（動物）
四十雀（しじゅうから）	夏（動物）	十月（動物）
障子（しょうじ）	冬（住）	雑（社会・生活）
石楠花（しゃくなげ）	夏（植物）	四月（植物）
赤潮（あかしお）	春（地理）	七月（地理）
雪割草（ゆきわりそう）	夏（植物）	三月（植物）
草蜉蝣（くさかげろう）	夏（動物）	九月（動物）
日雀（ひがら）	夏（動物）	十月（動物）
白蟻（しろあり）	夏（動物）	雑（動物）
髪洗う（かみあらう）	夏（人事）	雑（社会・生活）
風船（ふうせん）	春（人事）	雑（社会・生活）
頬白（ほおじろ）	春（動物）	十月（動物）
木天蓼（またたび）	秋（植物）	六月（植物）
冷蔵庫（れいぞうこ）	夏（住）	雑（社会・生活）
藁仕事（わらしごと）	秋（人事）	十二月（生活）
屏風（びょうぶ）	冬（住）	雑（社会・生活）
筍（たけのこ）	夏（植物）	三月（植物）
絨毯（じゅうたん）	冬（住）	雑（社会・生活）
鮑（あわび）	夏（動物）	四月（動物）

見出しでマッチングした実語数は、1,571 語であったが、初春（しょしゅん）は読みが異なるため除外した結果、1,570 語となった。さらに、マッチングした語の中で上記28語は、両書で季節が異なるため除外することとした。

次に「歳時記」における分類にしたがい、両書に共通の季語がカテゴリによって偏りがあるかを調べた。その結果、『雑』というカテゴリは「歳時記」に独特のものであるため、共通語が少ないのは当然のことと考えられる。また、「季寄せ」には『行事』と『生活』カテゴリに分類される季語の割合が低いようである。

カテゴリ別該当語数

	共通語／総　数	(%)
行事	114／161	(70.8)
時候	158／181	(87.3)
雑	9／383	(2.3)
植物	535／629	(85.1)
生活	299／442	(67.6)
地理	65／ 74	(87.8)
天文	128／157	(81.5)
動物	262／343	(76.4)

4　季語データベースによる俳句の季語の特定

4.1 俳句データベースとサンプルの抽出

現在収集構築中の俳句データベースは、先の芭蕉祭献詠俳句集第40巻から第52巻までの選者及び一般の部の6,912 句をすでに収録している。また、児童・生徒の句については、第40巻から第53巻までの6,833 句を別途データベース化している。俳句データは、可能な限り五七五のリズムに合わせて上の句、中の句、下の句に三区分している。また全文かな読みくだしを付与している。

今回選んだ季語とのマッチングテスト用のサンプルとして、一般の部の第52巻分448 句を抽出した。一般の部から抽出した理由は、後にも述べるようにかな文字列でのマッチングではその後に要する人的な作業が多くなることが想像されたためである。

4.2 サンプルの季語の特定

季語を長さの順に並べ、長いものから順に一語ずつ、ひとつの俳句の頭から一文字ずつずらして比較していき、マッチしたものを季語として特定した。ただし、最長マッチング法では、俳句のさらに後半部分により短い文字列だが重要な季語を見落とす可能性があるため、すべての季語のマッチングをとった。次に、漢字見出しで季語が抽出出来なかった句に対しては、ＪＥＦ以外の漢字を使用している場合には、ゲタコードとなってマッチングがとれなかったり、作者の趣味によってひらがなや旧字を使用していることが考えられるため、かな読み下し文に対して同様にして読みによるマッチングを行った。

その結果、サンプルとした448 句のうち、漢字見出しで季語が特定出来た句は、295 句(65.8%) であった。ひとつの句に対してマッチした季語の数の分布は、１語（197 句）、２語（77句）、３語（18句）、４語（１句）、５語（１句）、15語（１句）であった。15語マッチしたのは季語がゲタコードであったためのものである。

漢字見出しで季語の特定が出来なかった153 句のうち、かな見出しでマッチした句は144 句、まったくマッチングがとれなかった句は９句であった。ひとつの句に対するかな見出しでマッチした季語の数の分布は、１語（28句）、２語（47句）、３語（32句）、４語（17句）、５語（12句）、６語（７句）、９語（１句）であった。

4.3 マッチした季語の検査

季語のマッチングは機械的に行っているため、不適切な文字列とのマッチングが発生していることが考えられる。そこで、目による適合の検査を行った。この段階の検査では、長い季語に含まれる文字が

短い季語として採られている場合、例えば「稲の花」と「稲」「花」、に生じる重複マッチングを削除した。また、かな見出しで意味のない部分とマッチしたものを削除した。

その結果、かな見出しでマッチした144 句のうち意味のある文字列とマッチしていたものは16句のみで、歩留りは11.1%であった。一方、漢字見出しでマッチした295 句は、すべて検査に合格した。したがって、1,542 語の季語でマッチした句は311 句 (69.4%) となった。

また、一つの句で複数の季語とマッチした句は35句あり、うち1句は季語が3つ含まれていた。

4.4 季語の吟味

先の検査の結果残った 311句の347 語の季語について、その句の季語として適切なものであるかを一つずつ吟味した。その際の観点は、明らかにその語が季語として認められるもの、その語を含む熟語あるいは語句が季語と同じ季節を表すと見なせるもの、その語を含む熟語や語句がまったく異なる意味であるもの、あるいは季語とは見なされないもの、の3通りに分けることにした。

その結果、明らかに季語であるものは、176 語、その語を含むものが季語であるもの118 語、適切な季語と認定できないものが53語あった。俳句で数えると、293 句は季語を特定できたことになる。ただしこの内2句は季語を2語持っていた。残る18句は、文字列としては可能性はあったが、季語としては適切な言葉ではなく、季語が特定できなかった句ということになる。

4.5 季語の特定結果

「歳時記」と「季寄せ」に共通する季語1,542 語を用いて、サンプルとした俳句 448句の季語の特定を試みた結果、季語と文字列がマッチングした句は439 句 (98.0%) であったが、そのうち季語が特定できたと考えられるものは293 句(65.4%) であった。

季語の読みでマッチングした句は 144句あったが、旧字あるいは平仮名表記であるものを救済できたものは、14句 (9.7%) で、90パーセント以上が篩落とされたことになる。一方、漢字見出しでマッチングしたものについては、295 句のうちの279 句(94.6%) が季語が特定された。

したがって、俳句の季語を特定するためには漢字見出しが有効であること、かな見出しを用いることについてはさらに工夫が必要であることがわかった。また、1,542 語で448 句の65.4% の季語が特定できたことから、季語データベースとして収集すべき季語の数は、それほど多くなくとも機能するであろうということがわかった。具体的な目標数値は、今後季語数を倍に増やして再度分析を試みることと、俳句の数が増えた場合に季語の数がどのような増え方をするのかを分析して決定することになる。これは、今後の課題である。

5　季語データベースを利用した俳句の数量的研究

最後に、今回サンプルとした俳句について、季語のカテゴリによりどのようなことがわかるか。数量的分析の結果を示す。次の表は、季語が特定された293 句の選者別、カテゴリ別俳句数を示す。この年の14人の選者は、それぞれ32句ずつを選んでいる。アスタリスクのついた数値は、ひとつの句に2つの季語が存在するため重複して数えていることを示す。シャープは新年の句を1句含むことを示す。ただし、新年の句はそれ1句であったためカテゴリ列は省略した。

表から明らかなように、芭蕉祭が秋催されることもあり、夏の句の投稿が多くそのため夏の句が多く選ばれる結果になっていると思われるが、選者毎にが好む季節感が異なるということが窺われる。また、今回のマッチングをとった季語を一般的な季語とするならば、詠み込まれる季語の一般性と特殊性につ

いても選者によって違いが見られる。事柄については、植物、動物を詠み込んだ句が多く選ばれているのは、このカテゴリに属する季語も多く妥当な結果であると思われる。一方、生活のカテゴリに属する季語は動物に属するものよりも多かったにもかかわらず、選ばれた俳句の数は少ない。また、選者によって事柄に関する嗜好も異なるようである。このことは投句者が選者を指名して投句することと密接な関係があると思われる。

選者名	季節					事柄							
	春	夏	秋	冬	計	時候	植物	生活	地理	天文	動物	行事	計
井澤　正江	4	12	6	1	23	7	5	1	1	1	8	0	23
稲畑　汀子	1	11	6	1	19	2	7	4	1	4	0	1	19
塩田藪柑子	0	12	8	6	26	8	4	4	4	3	3	0	26
丸山　海道	2	15	4	1	22	3	4	2	1	4	7	1	22
金子　兜太	4	9	4	3	20	3	6	3	1	1	4	2	20
鍵和田柚子	3	8	7	2	#21	4	4	2	1	3	7	0	21
古舘　曹人	4	9	7	3	23	1	9	2	0	2	8	1	23
松崎鉄之介	0	11	7	0	18	4	5	3	0	3	3	0	18
森田　峠	1	13	3	1	18	4	4	2	1	2	3	2	18
早崎　明	0	11	5	2	18	2	2	4	0	3	5	2	18
草間　時彦	1	8	8	3	20	3	*10	3	1	1	2	* 1	21
鷹羽　狩行	0	11	6	4	21	3	4	4	0	1	5	4	21
堀口　星眠	2	14	4	2	22	2	4	6	1	4	3	2	22
澤木　欣一	2	9	11	0	22	3	12	0	1	0	* 4	* 3	23
計	24	153	86	29	293	49	80	40	13	32	62	19	295

以上のように、季語データベースと連動させて俳句を分析することで、様々な分析が行えると予想される。

文　献

1）「二〇〇〇年版俳句年鑑」（俳句１月号増刊），角川書店，2000年１月１日

2）芭蕉翁記念館編，「第五十三回芭蕉祭献詠句集」，(財)芭蕉翁顕彰会，1999年10月10日

3）「新版季寄せ」，角川書店，1985年５月10日

4）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

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

【付　録】　　季語の分類カテゴリ

入手した「季寄せ」「歳時記」本の季語の分類カテゴリについて表にして示す。

角川・新版季寄せ・1985年5月10日

春 夏 秋 冬 新年	時候	
	天文	
	地理	
	人事	行事 衣 食 住 農耕・狩漁 遊楽 情緒
	宗教	神道 仏教 キリスト教 忌日
	動物	四足動物 鳥 魚介 虫
	植物	花木 果樹 樹木 草花 野菜・作物 野草 茸 海藻

※新年には網かけ部はない。

成星出版・現代歳時記・1997年2月12日

新年（冬） 一月（冬） 二月（冬・春） 三月（春） 四月（春） 五月（春・夏） 六月（夏） 七月（夏） 八月（夏・秋） 九月（秋） 十月（秋） 十一月（秋・冬） 十二月（冬）	時候
	天文
	地理
	生活
	行事
	動物
	植物
雑	天文・時間 地理・空間 人間 社会・生活 文化・宗教 動物 植物 物質・物理 固有名

※一月は地理がない。

文藝春秋・季寄せ・1973年10月5日

春	三春 初春 仲春 晩春
夏	三夏 初夏 仲夏 晩夏
秋	三秋 初秋 仲秋 晩秋
冬	三冬 初冬 仲冬 歳末 晩冬
新年	

角川春樹事務所・合本現代俳句歳時記・1998年7月8日

春 夏 秋 冬 新年	時候
	天文
	地理
	人事
	宗教
	動物
	植物

角川書店・合本俳句歳時記第三版・1997年5月30日

春 夏 秋 冬 新年	時候
	天文
	地理
	生活
	行事
	動物
	植物