The Concordance to 'The Tale of Genji' and the text analysis

『源氏物語』コンコーダンスとその応用

近藤泰弘 (青山学院大学) *

The Concordance to 'The Tale of Genji' and the text analysis

KONDO Yasuhiro (Aoyama Gakuin University)

要旨

日本語の古典語の語句の「総索引」、つまり、「コンコーダンス」をコンピュータで作成することによって、従来の手作業で作っていた総索引とは質的に異なったものを作成できることを述べ、その形式を工夫することで、各種の言語学的情報を引き出しやすくなることを述べる。

1. はじめに

日本語の古典文学作品の研究にあたっては、その中の出現単語をすべて網羅して配列した「総索引」と言われる種類の書物が多く作成されてきた。これは書物の中の重要語を検索する「索引」と区別するために作られた語である。日本で「総索引」を初めて名乗ったのは、『万葉集総索引』(正宗敦夫 1931) であるが、この書物は、その序文にあるように、その構成のすべてを橋本進吉の助言・指導によって作成したとある。そして、後で述べるように、この『万葉集総索引』の形式は、ヨーロッパで広く作られていたコンコーダンス (concordance) の形式をそのままなぞって作られている。おそらく橋本進吉はそれら欧州のコンコーダンスに対する知識があったため、それを日本語の古典作品においても作りたかったものと思われる。

これ以後、多くの「総索引」が作られるようになったが、このネーミングから、日本語としては、「総索引」が「コンコーダンス」、「索引」が「インデックス」というふうに対応することとなった。なお、「総索引」を「索引」と銘打って出版したものもあるなど、混乱もあるため、以下、本稿では、『万葉集総索引』のようなタイプのものをすべてコンコーダンスと呼ぶ。

そして、以下、まずヨーロッパにおけるコンコーダンス発達の歴史にふれ、手作業で作られていたコンコーダンスとコンピュータで作ることのできるコンコーダンスの相違を述べ、その後、コンピュータによるコンコーダンスとその応用としてのテキスト分析の方法について述べていきたい。

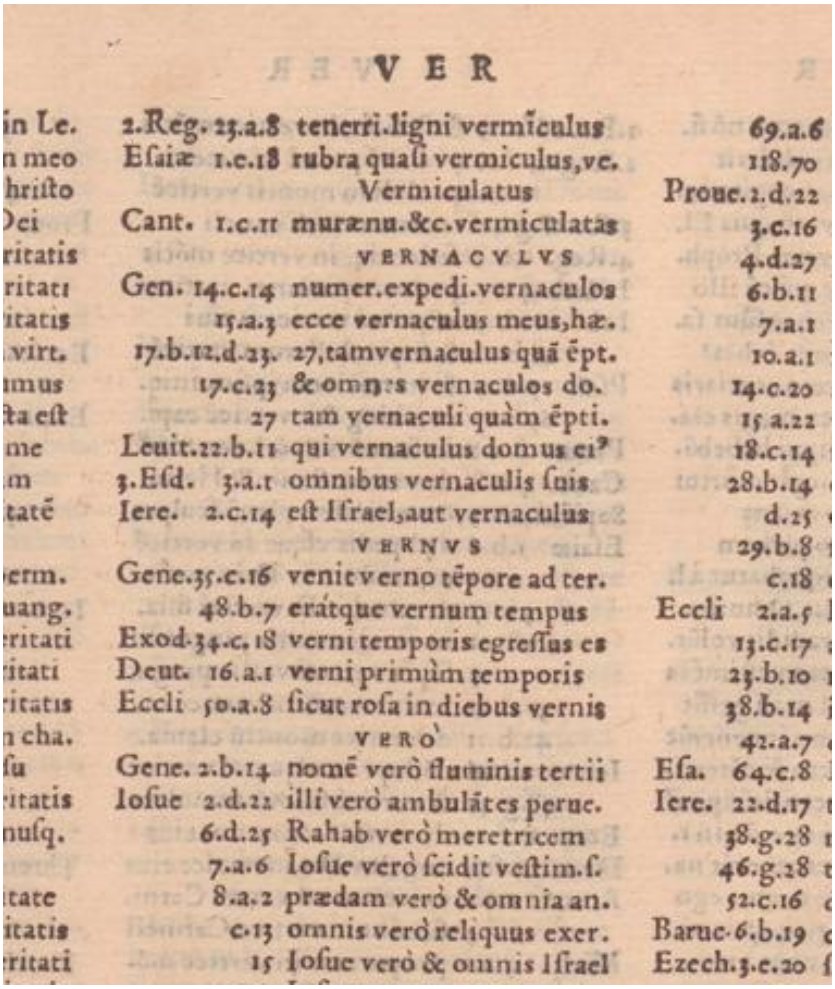
2. コンコーダンスとは

2.1 ヨーロッパにおけるコンコーダンス

コンコーダンスはいつ頃発明されたかであるが、13 世紀にラテン語聖書を対象として作り始められたものが最初のようだ。15 世紀になってグーテンベルクの印刷術の発明以後、コン

* yhkondo@cl.aoyama.ac.jp

コーダンスも印刷本の時代となるが、中でも現代のコンコーダンスにもっとも近いもので、印刷された最初は、1555 年に、フランス人のロベール・エティエンヌ（Robert Estienne）（出版者かつ学者）が刊行したラテン語旧新約聖書のコンコーダンスがある。これは、固有名詞を含む大半の語が用例付きで収められアルファベット順に配列されたもので、現代のコンコーダンスの直接の祖先と言ってもいいものである（Robert Estienne's Influence on Lexicography (Starnes 2014)）（コンコーダンスの正式名: Concordantiae Bibliorum vtriusque Testamenti, Veteris & Noui, nouae & Integrae.）



VER

2.Reg. 23.a.8 tenerri.ligni vermiculus
Esaiæ 1.c.18 rubra quasi vermiculus,ve.
Vermiculatus
Cant. 1.c.11 murænu.&c.vermiculatas
VERNACVLVS
Gen. 14.c.14 numer.expedi.vernaculos
15.a.3 ecce vernaculus meus,hæ.
17.b.12.d.23. 27,tamvernaculus quã ẽpt.
17.c.23 & omnes vernaculos do.
27 tam vernaculi quàm ẽpti.
Leuit.22.b.11 qui vernaculus domus ei⁹
3.Esd. 3.a.1 omnibus vernaculis suis
Iere. 2.c.14 est Israel,aut vernaculus
VERNVS
Gene.35.c.16 venit verno tẽpore ad ter.
48.b.7 erátque vernum tempus
Exod.34.c.18 verni temporis egressus es
Deut. 16.a.1 verni primùm temporis
Eccli 50.a.8 sicut rosa in diebus vernis
VERÒ
Gene. 2.b.14 nomẽ verò fluminis tertii
Iosue 2.d.22 illi verò ambulãtes perue.
6.d.25 Rahab verò meretricem
7.a.6 Iosue verò scidit vestim.s.
8.a.2 prædam verò & omnia an.
c.13 omnis verò reliquus exer.
15 Iosue verò & omnis Israel

69.a.6
118.70
Prouc.2.d.22
3.c.16
4.d.27
6.b.11
7.a.1
10.a.1
14.c.20
15.a.22
18.c.14
28.b.14
d.25
29.b.8
c.18
Eccli 2.a.5
13.c.17
23.b.10
38.b.14
42.a.7
Esa. 64.c.8
Iere. 22.d.17
38.g.28
46.g.28
52.c.16
Baruc.6.b.19
Ezech.3.c.20

しかも、この本にはひとつ大きな工夫が含まれている。それまでの聖書には、章しかなく、節がなかった。エティエンヌは、聖書の本文も別に、1551 年（ギリシャ語）、1555 年（ラテン語）と刊行しているのだが、それらに、初めて節番号をふっている（現在、「マタイによる福音書 5 章 3 節」のような形式で引用されるが、この節番号）。コンコーダンスもその節番号を使っているため、容易に本文箇所が検索できるのだ。前述の『万葉集総索引』では、国歌大観番号が万葉集につけられているため、それを使って検索できるだけだが、このように、コン

コーダンスには、検索の手がかりとなる出典の場所表示が不可欠である。

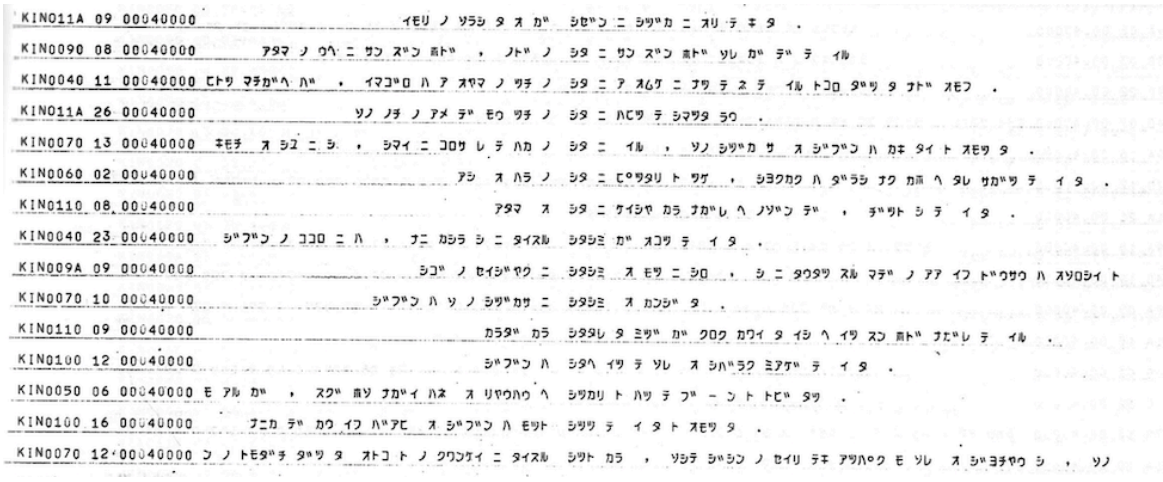
なお、エティエンヌは、その後ギリシャ語新約聖書についても同様のコンコーダンスを 1624 年に出しており、これはギリシャ語聖書コンコーダンスの印刷物としては最初期のものである（Concoredantiae Testamenti Novi, Graecolatinae. ）

以上、最初期のコンコーダンスに立ち戻って考えると、コンコーダンスに不可欠な要素は、単語の一覧性と規則的な配列、文脈表示、容易な検索のための出典表示の 3 つである。

2.2 日本におけるコンコーダンス

日本におけるコンコーダンスは、前述『万葉集総索引』以後、数多く出版された。文脈表示がない、出典表示だけの簡易なものが多かった中で、文脈までつけたものとしては、『竹取物語総索引』(山田忠雄 1958)、『土左日記総索引』(日本大学文理学部国文学研究室 1967)、『後撰集総索引』(大阪女子大学国文学科国文学研究室 1965) などが代表的なものである。また、総索引と銘打ってはいないが、『源氏物語辞典』(北山谿太 1957) は、語釈がつけてあるという点以外はコンコーダンスとしての条件を満たしており、実質的に源氏物語コンコーダンスである。

以上は、手作業によるものであるが、1960 年代後半から、コンピュータによるコンコーダンスが作られるようになった。日本で最初の市販されたコンピュータによるコンコーダンスは『平家物語総索引』(金田一春彦・近藤政美（編） 1973) であるが、これは文脈がついていなかった。用例付きのコンコーダンスとして公刊されたものとしては、『志賀直哉『城の崎にて』用語索引ー電子計算機による用例集作成の一実験ー』(国立国語研究所（石綿敏雄） 1971) がおそらく最初のものだと思われる。



このコンコーダンスは、当時の計算機の制約から、半角片仮名の表示にはなっているが、用例付き、しかも、当時日本でも取り入れられたばかりの KWIC 形式を採用しているなど先進的なものである。ちなみに、出典表示は、章と文の連番によっている。これ以後、国語研でのコンコーダンスは『森鷗外『寒山拾得』用語索引』(国立国語研究所（鶴岡昭夫） 1974)、『牛

店雑談安愚楽鍋用語索引』(国立国語研究所（飛田良文） 1975) があるが、その後は作られていない。

今回の発表ではこの部分を問題としてみたい。

2.3 コンコーダンスの必要性

国語研究所では、上記コンコーダンスの作成の後しばらく間隔があってから、コーパス事業が開始され、『太陽コーパス』を皮切りとして現在の日本語歴史コーパス (国立国語研究所 編) に至る古典語コーパスの開発が進められている。そして、KWIC 型出力が可能なコンコーダンサである「中納言」の能力も相まって、多くの古典文学作品から KWIC 形式で単語の検索結果、文脈、出所情報などが得られるようになっている。しかしながら、大型の歴史コーパスの KWIC 形式のビューアーと、特定の作品のコンコーダンスとではすこし機能が違っている。

具体的には、コーパス検索の KWIC 形式出力では、ターゲットとなった語についての情報は得られるが、その語と語形が近い語や意味の近い語がそのコーパス（文献）に入っているかどうかの情報はいっさい得ることができないことがあげられる。それに対して、特定の文献のコンコーダンスでは、単語が五十音順にソートされていれば似た語形を順に調査することが可能であるし、もし意味的に関連ある語でソート（シソーラス的な配列）がされていれば、似た意味の語を調査してその文献の性格を調査する手がかりとすることができる。1 作品に限ることではなく、例えば、シェークスピアの全作品の、シェークスピア・コンコーダンスというようなものもあり、その作家について調査するというような方法もある。

以上のように、「中納言」等でコーパスが検索できれば十分というわけではなく、人間の目で見て特定の文献の語彙の概要をつかむためのツールとして、コンコーダンスは必要である。『万葉集総索引』の編纂には十数年の年月がかかったようであるが、コーパスデータから、コンピュータによってコンコーダンスに改編するのは極めて短時間にできる。また、このような作業はいわゆる理系の研究ではあまり需要がないが、文系の研究では従来から必要であるし、今後もその必要性は高い。

前回の発表 (近藤泰弘 2017b) では、平安時代語語彙集という形での、特定時代のコンコーダンスの必要性について述べたが、ひじょうに重要度の高い文献についての、個別のコンコーダンスを作成することも重要である。近代語で言えば、夏目漱石コンコーダンスなどが必要であるし、古代では、八代集コンコーダンスなどが必要である。今回は、そのような考えのもとに源氏物語コンコーダンスを素材として、どのような形態が有用であるかを述べてみたい。

2.4 コンピュータによるコンコーダンスの分類

2.4.1 キーワードの位置

コンピュータで作成する場合には、文脈情報は必須である。それと直接の検索対象の語（以下では「キーワード」と呼ぶ）との関係には大きくわけて 2 種類がある。ひとつは、キーワードが文脈中にあるいわゆる KWIC(Key Word In Context）の形式、もうひとつはキーワードを文脈の外に特出する KWOC(Key Word Out of Context）の形式である。どちらも一長一短であるが、今回はなるべく古来からのコンコーダンスの形に似せるため、KWOC の形を主に採用した。

2.4.2 ソート順

キーワードの五十音順・出所順・キーワードの前後文脈順（正順、逆順）・キーワードのシソーラス状の順番などが考えられ、また、これらの複数のキーのソート順位（どれからまずソートするか）も問題となる。これについても、今回は、古来のコンコーダンスに習い、キーワードの五十音順を第 1 キーとして、次に出所情報（桐壺・帚木などの 54 帖順）を第 2 キーとして配列した。

2.4.3 キーワードの形態論的性質

キーワードを切り出すためには、その言語単位を決めておく必要がある。通常は単語をキーワードとするが、今回の日本語歴史コーパスの XML データの場合には、長単位か短単位かという問題がある。あるいは、文節による方法もある。過去には手作業によるものであるが『今昔物語集文節索引』(馬渕和夫 1970) というものがあった。また、連語の状態を見るために、短単位の N-gram による KWIC も考えられる。

3. 源氏物語コンコーダンス

以上の様なことを前提に、コンピュータによってコンコーダンスを試作する。以下では、キーワードの形態論的性質によって分類して示す。

3.1 短単位コンコーダンス

まずは、短単位で作成することが一番使いやすいコンコーダンスとなる。これは何が長単位であるかがわかりにくいため、もれなく見ることが簡単だからである。しかし、複合動詞や複合辞も分散して配置されるため、その点は使いにくい。しかし、今回はこれを主力として、作成した。また、コンコーダンスに補助情報をつけるため、前回の発表と同様に word2vec の skip-gram モデル（ウィンドウサイズは今回は 10）で値が近い短単位を採集し、それぞれのキーワードの用例の末尾に付載した。

以下は、「あ」の部分の例である。ちなみに、『源氏物語』コンコーダンスをこの形で作ると、A4 で、1332 ページとなる。

あかいろ【赤色】［名詞-普通名詞-一般］「なり。童六人、赤色に桜襲の汗衫、(17 絵合)」「を着たまふ。帝は赤色の御衣奉れり。(21 少女)」「たまふに、帝の、赤色の御衣奉りてうるはしう (29 行幸)」「容貌すぐれたる四人、赤色に桜の汗衫、(35 若菜下)」《連想語》汗衫・藤襲ね・表袴・山吹襲・浮紋・袿・唐綾・腰差・無文・葡萄染め

あかぎ【赤木】［名詞-普通名詞-一般］「て、よしある黒木、赤木の籬を結ひまぜ (28 野分)」《連想語》黒木・色種・萌え木・ハコヤ・枝差し・刀自・古巣・植木・稲・唐守

あかぎぬ【赤絹】［名詞-普通名詞-一般］「なき気色にて、おどろおどろしき赤衣姿いときよげなり (14 澪標)」《連想語》物鮮やか・あいだれる・朝明け・侍童・腰つき・聳える・縒る・上衆・匂わしい・面様

あかし【アカシ】［名詞-固有名詞-地名-一般］「所には、播磨の明石の浦こそなほこと (5 若紫)」

「もえ出でざりけり。明石の浦は、ただ (12 須磨)」「。まことや、かの明石には、返る波 (13 明石)」「。まことや、かの明石に心苦しげなりし (14 澪標)」「ついでにも、かの明石の家居ぞ、まづ (17 絵合)」「しおかせたまへり。明石には御消息絶え (18 松風)」「引き寄せて、かの明石にて小夜更けたりし (19 薄雲)」「にこそ、いと頼もしけれ。明石の入道の例に (21 少女)」「が艶やかなる重ねて、明石の御方に、(22 玉鬘)」「暮れ方になるほどに、明石の御方に渡り (23 初音)」「にて明かし暮らしたまふ。明石の御方は、(25 蛍)」「かし。劣り腹なれど、明石のおもとの産み出で (26 常夏)」「やがて北に通りて、明石の御方を見 (28 野分)」「御方はしつらひたり。明石の御方、今 (34 若菜上)」「を」などのたまへば、明石の君は、いと (35 若菜下)」「はこなたにおはすれば、明石の御方も渡り (40 御法)」「をかしきほどなれば、やがて明石の御方に渡り (41 幻)」「なりけりと見えて、明石の御方は、(42 匂兵部卿)」「ことなりけるこそあやしけれ、明石の浦は心にくかりける (52 蜻蛉)」《連想語》戌亥・対・丑寅・斎宮・夕方・秋頃・町・西・戌・舎

あかし【証し】[名詞-普通名詞-一般]「にさやうに触ればひぬべき証やあると尋ねとぶらひ (26 常夏)」《連想語》五師・江・不意・落ち葉・囃す・側々・ミシマ・等し並み・心得・かじける

あかしぶみ【明かし文】[名詞-普通名詞-一般]「大徳呼びて言ふ。御あかし文など書きたる心ばへなど (22 玉鬘)」《連想語》真名・若やぐ・クマノ・ショウクン・僻覚え・コセ・オウ・女文・中定・和らぐ

3.2 長単位コンコーダンス

長単位コンコーダンスの場合は、定義上、敬語要素がひとつの長単位となる。次は、長単位でデータを採集中の例である。

桐壺,32, より, ヨリ, 助詞-格助詞,, て時めきたまふありけり。はじめより我はと思ひあがりたまへる

桐壺,33, 我, ワレ, 代名詞,, 時めきたまふありけり。はじめより我はと思ひあがりたまへる御方々

桐壺,34, は, ハ, 助詞-係助詞,, ありけり。はじめより我はと思ひあがりたまへる御方々、

桐壺,35, と, ト, 助詞-格助詞,, けり。はじめより我はと思ひあがりたまへる御方々、めざましき

桐壺,36, 思い上がり給う, オモイアガリタマウ, 動詞-一般, 文語四段-ハ行,。はじめより我はと思ひあがりたまへる御方々、めざましきもの

桐壺,37, り, リ, 助動詞, 文語助動詞-リ, はじめより我はと思ひあがりたまへる御方々、めざましきものに

桐壺,38, 御方々, オオンカタガタ, 名詞-普通名詞-一般,, より我はと思ひあがりたまへる御

方々、めざましきものにおとしめそねみたまふ

このように「思い上がり給う」や「御方々」などの長単位語彙素がキーワードとなってくる。ただ、長単位の定義から、『源氏物語辞典』などが採用する「明かし暮らしわぶ」のような長大な複合動詞は採用していない。

3.3 短単位 2gram コンコーダンス

短単位で 2 グラムで作成すると次のようになる。これは作成途中であるが、「我は」「と思う」などの 2 単位連続のコーパスを作成することになる。「より我」のような意味のない連続もあるが、大半は意味のある連続であり、その様子を大局的に観察できて有用である。

桐壺,35, 始めより, ハジメ, 名詞-普通名詞-副詞可能,, 時めきたまふありけり。はじめより我はと思ひ
桐壺,36, より我, ヨリ, 助詞-格助詞,, めきたまふありけり。はじめより我はと思ひあがり
桐壺,37, 我は, ワレ, 代名詞,, たまふありけり。はじめより我はと思ひあがりたまへ
桐壺,38, はと, ハ, 助詞-係助詞,, ありけり。はじめより我はと思ひあがりたまへる
桐壺,39, と思う, ト, 助詞-格助詞,, けり。はじめより我はと思ひあがりたまへる御
桐壺,40, 思う上がる, オモウ, 動詞-一般, 文語四段-ハ行,。はじめより我はと思ひあがりたまへる御方々
桐壺,41, 上がる給う, アガル, 動詞-一般, 文語四段-ラ行, はじめより我はと思ひあがりたまへる御方々、
桐壺,42, 給うり, タマウ, 動詞-非自立可能, 文語四段-ハ行, より我はと思ひあがりたまへる御方々、めざましき
桐壺,43, り御, リ, 助動詞, 文語助動詞-リ, 我はと思ひあがりたまへる御方々、めざましきもの
桐壺,44, 御方々, オオン, 接頭辞,, はと思ひあがりたまへる御方々、めざましきものに

4. おわりに

本稿ではコンコーダンスをコーパスから作成することについての考察を行った。紙面の関係で、word2vec による情報だけを掲載したが、その他の機械学習の結果や統計による情報も更に拡充したものをスライドでは提示したい。

謝　辞

本研究は JSPS 科研費 JP25284086（「平安時代における言語リソースの構築に関する研究・研究代表者・近藤泰弘・連携研究者・近藤みゆき）の助成を受けたものである。

文　献

正宗敦夫 (1931).『万葉集総索引』　白水社.

DeWitt T. Starnes (2014). *Robert Estienne's Influence on Lexicography.*: University of Texas Press.

山田忠雄 (1958).『竹取物語総索引』　武蔵野書院.

日本大学文理学部国文学研究室 (1967).『土佐日記総索引』　日本大学.

大阪女子大学国文学科国文学研究室 (1965).『後撰和歌集総索引』　大阪女子大学.

北山谿太 (1957).『源氏物語辞典』　平凡社.

金田一春彦・近藤政美（編） (1973).『平家物語総索引』　学習研究社.

国立国語研究所（石綿敏雄）(1971).『志賀直哉『城の崎にて』用語索引ー電子計算機による用例集作成の一実験ー』　国立国語研究所.

国立国語研究所（鶴岡昭夫）(1974).『森鴎外「寒山拾得」用語索引』　国立国語研究所.

国立国語研究所（飛田良文）(1975).『牛店雑談安愚楽鍋用語索引』　国立国語研究所.

国立国語研究所 (編) (2017a).『日本語歴史コーパス』, (バージョン 2017.3, 中納言バージョン 2.2.2) https://chunagon.ninjal.ac.jp/ (2017 年 7 月 26 日確認).

近藤泰弘 (2017b).『古典語コーパスからの語彙集作成について』, 通時コーパスシンポジウム 2017 口頭発表資料.

馬渕和夫 (編) (1970).『今昔物語集文節索引』　笠間書院.

関連 URL

日本語歴史コーパス『中納言』	https://chunagon.ninjal.ac.jp/
上記科研費成果公開サイト（Japanese.gr.jp）	http://japanese.gr.jp/