平安時代古典語古典文学研究のための N-gramを用いた解析手法

近藤泰弘（青山学院大学）・近藤みゆき（千葉大学）

1　はじめに

近年、平安時代の日本語で書かれた『古今和歌集』や『源氏物語』を代表とする、いわゆる古典語古典文学のテキストデータベースが数多く作成され、CD-ROMやWebで公開されるようになってきている。そこで、本研究ではそのための手法として、文字を対象としたN-gram統計を用いる方法を提案し、それを実際の問題に応用し、その手法の妥当性を検証した。N-gramを高速に算出するためには、長尾・森 [4] の方法を用いた。

用いるテキストデータとしては、『古今和歌集』（梅沢本を底本として、漢字部分も平仮名にしたもの）および『源氏物語』（大島本を底本としたもの。同様に平仮名にしたもの[1]）を用いる。古典語の場合、漢字仮名交じり文の表記が現代語のように均質なものとして体系化されていないため、表記レベルでそのまま処理するよりも、いったん含まれる漢字すべてを平仮名に開いて、仮名文字レベルに統一されたテキストデータを用いるのがよい。これは、音声言語のレベルでは、モーラ（音韻的音節）を対象としてN-gramを採取することに相当するものであり、言語分析の態度としても妥当である。

2　N-gram統計による古典語処理

2.1　方法の概要

本研究においては、文字を対象としたN-gramを2グラムから10数グラム[2]程度まで作品から採取し、その採取した異なり文字列群を、他の作品（あるいは同じ作品の別の個所）から採取した同様な文字列群と相互比較して、文字列のマッチングにより集合演算するという方法を用いた。

従来、古典語研究は、手作業で形態素解析されたデータについてKWICを作成するか、あるいは単純な文字列検索を行うか、いずれにせよ限られた解析方法で分析されてきた。いずれの方法も、データの量や正確さに問題があり、また、掛詞や古典語の複雑な慣用句などの分析には適していなかった。今回の提案はそのような問題点をある程度解決できるものである。

2.2　古典語におけるN-gram統計の状態

まず文字（モーラ）でのN-gramの分布を調査してみた結果である。『古今集』から見てみる。数値は各グラム内の異なり文字列数、括弧内がグラムである。

67(1)　2664(2)　15879(3)　26860(4)　31792(5)　34088(6)　35126(7)　35707(8)　35951(9)　36063(10)

[1]但し作業中につき、平仮名に開いたものは前半部のみ。
[2]同グラム数で頻度が2以上であるグラム。最大有一致グラム

36117(11) 36151(12) 36177(13) 36187(14) 36193(15) 36196(16) 36199(17) 36202(18) 36204(19) 36205(20) 36206(21) 36207(22) 36208(23) 36209(24) 36210(25) 36210(26)

次に『源氏物語』の原写本表記（漢字がわずかに混じった平仮名）の場合の N-gram の分布である。

663(1) 11897(2) 95378(3) 319835(4) 550652(5) 706041(6) 790657(7) 830940(8) 848089(9) 854855(10) 857493(11) 858494(12) 858889(13) 859047(14) 859113(15) 859142(16) 859158(17) 859165(18) 859168(19) 859170(20) 859171(21) 859172(22) 859173(23) 859174(24) 859175(25) 859176(26) 859177(27) 859178(28) 859179(29) 859180(30 859181(31) 859182(32) 859183(33) 859184(34) 859184(35)

3 『源氏物語』中の『古今和歌集』からの「引歌」の抽出

3.1 異なった作品の比較

最初に扱うのは、異なった作品同士の比較である。表現の差や作者の差などを調べることも可能であるが、今回注目したいのは、引用関係の調査である。古典文学のひとつの特色として、相互の引用関係が複雑であり、それぞれの時代の作品は前代の作品や、他者の作品を自由にその時代のコンテクストとして引用するということがある。散文に和歌が引かれる場合、その修辞技巧を「引歌」と呼ぶが[3]当然のことながら、後代の研究者が平安時代のコンテクストを抽出するのは容易な作業ではない。例えば次のようなものが引歌である。

- げに、この世は、短かめる**命待つ間**も、つらき御心は見えぬべければ…（源氏・宿木・5-408）
- ありはてぬ**命待つ間**のほどばかり憂きこと繁く思はずもがな（古今・雑下・965・平貞文）

[3]鈴木 [3] 等参照。
[4]近藤泰弘 [2] による

N-gram 分析によって得られた文字列の集合の共通集合を見るという前処理を行うことで、この引歌の認定をある程度簡単にしようというのが本手法のねらいであり、具体的には『源氏物語』における『古今和歌集』の引歌の研究に適用してみる[4]。

3.2 比較方法

和歌と散文との比較であるため、あまり長大な N を求めても意味はないため、2 グラムから 15 グラムを対象とする。すると、『源氏物語』では、異なりで 930 万種類の文字列が存在する。同じく『古今集』では、43 万種類となる。この相互の文字列をマッチングして共通文字列を抽出する。

このようにして比較すると、この 2 グラムから 15 グラムの範囲で、共通文字列は、30090 個存在する。その中で注目されるのは先の「いのちまつま」などのように、比較的長いやや特殊な意味のある文字列が一致しているものである。そこで、5 グラム以上のもので相互に一致する文字列をすべて検討してゆくと、偶然ではなく、有意に一致が見られると考えられるものが、見つかってくる。

3.3 新規の「引歌」の発見

1. あひみむとたのめ
   - 「逢ふまでのかたみに契る中の緒のしらべはことに変らざらなむ。この音たがはぬさきにかならず**あひみむ**」と**頼め**たまふめり（源氏・明石・2-267）
   - 今ははや恋ひ死なましを**あひ見むと頼め**しことぞ命なりける（古今・恋2・613・清原深養父）

源氏と明石の君との別れの場面であるが「あひみむとたのめ」というかなり長い文字列の一致により『古今』613 番歌が導き出されてくる。

2. くれなゐふかき

- 十月中の十日なれば、神の斎垣にはふ葛も色変はりて、松の下紅葉など、音にのみ秋を聞かぬ顔なり。（中略）**紅深き**袖の袂の、うちしぐれたるにけしきばかり濡れたる、松原をば忘れて、紅葉の散るに思ひわたさる。（源氏・若菜下・4-171）
- もみぢ葉の流れてとまる水門には**紅深き**波やたつらむ（古今・秋下・293・素性）

若菜下の巻での、住吉社での秋の舞楽の描写であり、岸辺で舞う人々の衣装を「紅深き」とする。古今素性歌における「紅深き」との用法と強い類似（季節・水辺・植物）がある。

他の例としては「いつしかとのみ」（源氏・蜻蛉・6-232・古今・183）「いろいろのはな」（源氏・少女・3-81・古今・245）「にそほちつつ」（源氏・夕霧・4-411・古今・422）などがある。引歌研究は中世以来の伝統があり、新しく見つけることはきわめて困難とされており、「あひみむとたのめ」のような文中の中間的場所にある文字列をもすべて総比較するこの手法によるならば、新規の発見が容易に実現するのである。今後比較対照の和歌集を増やすならば大量の引歌の発見を可能にすると予測されるのである。

4 『古今和歌集』中の作者の性差による言葉の型の抽出

4.1 同一作品中の比較

ここでは、同一作品の中の部分をその属性によっていくつかのグループに分類して、表現の差などを調査するという方法について考える。属性としては、部分的に作者が異なるのではないかと考えられている作品の一部分、和歌集のように、作者が複数である場合、その作者の属性によって分類して区分するなどの方法が考えられるが、ここでは作者の性別を用いる[5]。

『古今集』は、勅撰集の最初のものであるが、他の勅撰集と同様に、それぞれの歌には作者が明記されている。そのために、歌を原則として作者の性別によって分類することができる。なお、その上に「読人不知（よみびとしらず）」として、作者が不明な伝承的な歌が掲載されている。その結果「男性歌人」「女性歌人」「読人不知（性別不明歌人）」の3種類に作者を分類できることになる。

4.2 分類の方法

この3種にわけた上でN-gram分析を行うと、次の例のようにそのグラム数ごとに、文学表現としては異なった種類のものが析出される。

女性歌人2グラム 27　ひと・21　おも・20　もの・19　なり

男性歌人5グラム 23 ほととぎす・18　さくらばな・16 をみなへし

読人不知5グラム 17　ほととぎす・12　あしひきの・8　はるがすみ

男性歌人8グラム 4 かみなづきしぐれ・3　あらじとおもへば

男性歌人12グラム 2 あかずしてわかるるなみだ・2　かとのみぞあやまたれける

基礎語的な単語、いわゆる枕詞などの歌語、類句表現や歌風を感じさせる表現などいろいろのレベルのものが存在する。このようなものを相互に集合演算してみて、男性歌人特有文字列、女性歌人特有文字列等を選び出すことができるが、今回はその中から、男性歌人特有文字列を見てゆく。分類すると次のようになる[6]。

景物 をみなへし 16・うつせみ 5・しらくも 9

動詞 なびく 5・わたらむ 5

[5]基本的な方法については、近藤みゆき [1] に詳しい。
[6]掲出したものはごく一部である

枕詞 あしひきの 7・ちはやぶる 7・あらたまの 5

特徴的連語・「飽かず」を核とする語 あかず 6・あかずして 2・あかで 2・あかぬ 4

同上・「恋ふ」を核とする語 こひしかり 6・こひしかりける 2・こひしかる 6・こひしかるべき 2・こひしきものを 3・こひしと 2（以下略）

4.3 「飽かず」を核とする語

先の語群のうち、「飽かず」を核とする語を見てみる。

1. よそにのみあはれとぞ見し梅の花**あかぬ**色香は折りてなりけり（春上・37・素性）
2. 春霞たなびく山のさくら花見れども**あかぬ**君にもあるかな（恋 4・684・紀友則）
3. むすぶ手のしづくににごる山の井の**あかで**も人にわかれぬるかな（離別・404・紀貫之）

いずれも対象への尽きることのない愛着を表す言葉であるが、男性だけが用いる用語となっている。このような対象をいとおしむという発想によることばが、男性の専用の領域として存在し、逆に女性側にはほとんど存在しなかったことが判明する。この「飽かず」に対して、女性専用文字列として「あかれやはせぬ」（古今・61）がある。「飽く」の受身形が反語によって否定となり「十分には満足できない」という意味になったものであり、この「飽かず」とは性差の上で対極にある語であると言える。

4.4 ことばの型と性差の N-gram 分析による発見

以上のように『古今集』の表現には、性差を反映したことばの型があることがこの研究により判明した。このことは、従来、判別がほとんど不可能だった古代日本語における男性と女性のことばの位相差を網羅的に発見し、その発生と展開を研究することが可能になることを意味している。このような分析は従来文学研究で用いられてきた単語索引 (KWIC) などでは行うことができなかった。つまり、付属語を含めた述語を構成する特徴的な文字列を抽出することは、形態素解析してしまった後ではきわめて困難なのであり、N-gram で文字列（モーラ列）の範囲で分析した方が網羅的となるわけである。これは、よく知られているように、日本語の述語が、ある程度、形態素の連続という正規表現で近似できること[7]もその大きな理由になっているものと考えられる。

参考文献

[1] 近藤みゆき「ｎグラム統計処理を用いた文字列分析による日本古典文学の研究–『古今和歌集』の「ことば」の型と性差–」（千葉大学『人文研究』29,2000）

[2] 近藤泰弘「《文化資源》としてのデジタルテキスト–国語学と国文学の共通の課題として–」（『国語と国文学』77-11,2000）

[3] 鈴木日出男『古代和歌史論』（東京大学出版会,1990）

[4] 長尾眞・森信介「大規模日本語テキストのｎグラム統計の作り方と語句の自動抽出」（『自然言語処理』96-1,1993）

[5] 水谷静夫『国語学五つの発見再発見』（創文社・1974）

[7] 水谷 [5] などによる