語の出現傾向による『源氏物語』第一部各巻の分類
A Quantitative Study of "The Tale of Genji" via a Statistical Analysis of the Word Frequency

土山 玄
Gen Tsuchiyama
一橋大学 森有礼高等教育国際流動化センター, 東京都国立市中 2-1 第3研究館
Hitotsubashi University, 2-1 Naka, Kunitachi, Tokyo

概要: 『源氏物語』は平安時代に成立した 54 巻から構成される長編物語である。一般に『源氏物語』は三部構成であると考えられており、第一部は第 1 巻「桐壺」から第 33 巻「藤裏葉」までが該当し、これら 33 巻は「紫上系」と「玉鬘系」という 2 つの群に分類されるという見解がある。この見解は「紫上系」の登場人物は「玉鬘系」の諸巻にも現れるが、「玉鬘系」に初出の人物は例外なく「紫上系」の諸巻には現れないという事実に基づいている。しかし、第一部 33 巻が 2 群に分類されるとする見解を支持する根拠は登場人物の出現状況の他にない。そこで、本研究では語の出現傾向を統計的に分析することで、計量的な観点より両群の文体的特徴に相違が認められるか検討を加えた。分析の結果、出現頻度の高い語の出現頻度、出現頻度の高い機能語の出現頻度が両群に間において相違することが明らかになった。

Abstract: "The Tale of Genji" is one of the most famous classical works of Japanese literature and one of the oldest full-length stories. Most studies on Japanese literature have considered that the story comprises 54 volumes that are categorized into 3 parts, with volumes 1-33 categorized under the first part. However, the re is an idea that the order of these 33 volumes is different from the order of the number of volumes and there is a possibility that these volumes are classified into 2 groups: "Murasaki no Ue Goup" and "Tamaka zura Group." Furthermore, there are some objective datasets that support this idea. Herein, we statistically a nalyzed the writing style of the aforementioned groups by performing multivariable analysis and found a sta tistical difference between the writing style of these groups.

キーワード: 源氏物語, 計量文献学, 多変量解析, 主成分分析
Keywords: The Tale of Genji, stylometry, multivariable analysis, principal component analysis

1. はじめに

『源氏物語』は平安時代に成立した長編物語である。この物語は 54 巻から構成され、一般に『源氏物語』は三部に分割されると考えられている[1]。第一部には第 1 巻「桐壺」から第 33 巻「藤裏葉」までが属し、第二部は第 34 巻「若菜上」から第 41 巻「幻」の 8 巻が、第三部は第 42 巻「匂宮」から第 54 巻「夢浮橋」の 13 巻が属すとされる。この第一部の 33 巻には成立順序が現行の巻序と異なるとする見解があり、これら 33 巻は「紫上系」と「玉鬘系」という 2 つの群に分類されると論じられている[2]。「紫上系」及び「玉鬘系」の分類は表 1 に示す通りである。このように第一部 33 巻が 2 群に分類されるという見解の根拠として、登場人物の出現状況があげられる。紫上系に登場した人物は玉鬘系においても出現するが、反対に玉鬘系において初出となる登場人物は例外なく紫上系に登場することはない。武田 (1954) によれば、このような事実に基づき、紫上系の17巻が成立した後に玉鬘系が成立し年立に従い紫上系に挿入されたと論じられている[2]。しかし、第

一部 33 巻が 2 群に分類されるとする見解を支持する根拠は登場人物の出現状況の他にない。

そこで、本研究では『源氏物語』第一部の成立過程を解明するために、語の出現頻度について統計的に分析を行う。文学的文章のテキストデータを用いた計量的な研究は計量文献学と称される。文章の計量分析では、主に著者の文体に関わる習慣的特徴を統計的に解析する。文体という概念は多様であるが、計量文献学においては文体とは計数可能な記述形式のことであり、その内容は文字や語の頻度、語や文の長さなどの文章を構成する量的な要素である。

このような文章の計量分析において、著者について議論の余地がある文章を対象に、計量的な手法による研究は従来から行われている。計量的に著者の推定あるいは識別を行う場合、文中において語彙的意味ではなく文法的機能を担う助詞や助動詞などの機能語が広く分析に用いられ、欧米諸語で記述された文献や日本の近現代の文献を対象に、その成果が報告されている。一方で、テキストの電子化が容易ではないことから、日本の古典文学作品を対象とした研究は十分に展開されているとは言えない。これに加えて、『源氏物語』などの日本の古典文学作品の多くはオリジナル原稿が散逸しており、書写によってのみ受け継がれており、このため著者の特徴的な文体が希釈されている可能性が予想される。しかし、土山・村上 (2011) では、『うつほ物語』と『源氏物語』を対象とし、現代文や欧米文学と同様に古典文においても、語の出現率、あるいは語の頻度について計量分析を行うことで著者の識別が可能であることを論じている[3]。

このような著者不詳の文献を対象とした著者の識別や推定を目的とした計量的な研究では、機能語と称される文中において語彙的意味ではなく文法的機能を担う語彙が分析項目として取り上げられることが一般的である。日本語の場合は、助詞や助動詞などが機能語に該当し、このような機能語が計量分析に取り上げられる背景には、名詞や動詞などの語彙的意味を担う実質語に比べ、機能語は研究対象となる文献の内容、すなわちストーリーの影響を受けにくく、出現する語彙に記述内容に起因すると考えられる出現傾向の偏りが生じにくく、著者の識別に適していると考えられることがある。

表 1 紫上系の諸巻及び玉鬘系の諸巻

紫上系		玉鬘系	
01桐壺	13明石	02帚木	24胡蝶
05若紫	14澪標	03空蝉	25蛍
07紅葉賀	17絵合	04夕顔	26常夏
08花宴	18松風	06末摘花	27篝火
09葵	19薄雲	15蓬生	28野分
10賢木	20朝顔	16関屋	29行幸
11花散里	21少女	22玉鬘	30藤袴
12須磨	32梅枝	23初音	31真木柱
	33藤裏葉		

よって、本研究ではこのような計量文献学の方法を用いて、『源氏物語』の第一部について、紫上系と玉鬘系との間に計量的な相違が認められるのか検討を加える。

2. 関連研究

第一部の成立過程については、すでに計量的な検討が加えられている。村上・今西 (1999) は『源氏物語』の全文を電子化したテキストデータを用い、多変量解析を行った初の本格的な研究である。村上・今西 (1999) では、各巻の助動詞の出現率を求め数量化 III 類を行っている。分析の結果、紫上系は玉鬘系よりむしろ第二部の 8 巻と助動詞の出現傾向は類似していることを明らかにした。従って、第 1 巻「桐壺」を起筆の巻と仮定するのであれば、紫上系が成立した後に第二部が成立し、その後に玉鬘系が成立した可能性を論じている[4]。

次いで、小野 (2015) は村上・今西 (1999) と同一のデータを用いて、分散安定化変換を行い、階層的クラスター分析を用いて『源氏物語』の成立過程について検討を加えている。分析の結果、『源氏物語』の諸

巻は紫上系及び第二部、玉鬘系及び宇治十帖という2群に分類されることを報告した[5]。

また、土山 (2016) では助動詞以外の品詞についても巻別の出現率を求め主成分分析を行っている。その結果、紫上系と玉鬘系との間において出現傾向が異なっている品詞は助動詞のみであることを指摘し、紫上系の諸巻は玉鬘系の諸巻に比べて使役・尊敬の助動詞が頻出していることを明らかにした。またこれに加えて、動詞の未然形に接続する受身・尊敬・可能・自発の助動詞である「る」は紫上系に、「らる」は「玉鬘系」に多く出現することを報告した。

3. 分析

3.1 データ

本研究では青表紙本系の大島本を主な底本とする『源氏物語語彙用例総索引 自立語編』[6]及び『源氏物語語彙用例総索引 付属語編』[7]を電子化したテキストデータを分析に利用した。『源氏物語語彙用例総索引』は源氏物語の本文すべてについて、形態素解析を行ったものである。すべての語に、表記形、表記形仮名読み、終止形、終止形仮名読み、品詞コード、活用形コード、意味コードなどの情報が付与されている。なお、単語認定については、『源氏物語大成索引篇』[8]の単語認定基準に準拠している。

3.2 方法

本研究では語のn-gramを分析に採り上げた。n-gram は文中において隣接する n 個の要素を 1 つの単位とする。よって、単語の n-gram は隣接する n 個の単語を 1 つの単位として頻度を集計した特徴量である。n = 1 のときは unigram、n = 2 のとき bigram、n = 3 のときは trigram と称され、本研究では unigram 及び bigram を分析に用いた。『源氏物語』の冒頭の文章である「いつれの御時にか女御更衣あまたさふらひ給けるなかにいとやむことなききはにはあらぬかすくれて時めき給ありけり」を例とすると、unigram は「いつれ (代名詞)」「の (助詞)」「御時 (名詞)」「に (助詞)」「か (助詞)」などの各語の頻度を集計する。他方、bigram では「いつれ (代名詞)─の (助詞)」「の (助詞)─御時 (名詞)」「御時 (名詞)─に (助詞)」「に (助詞)─か (助詞)」という語の隣接共起した組の頻度が集計される。

また、文章の計量分析では品詞別に語の unigram を集計し、品詞毎に分析を行うことが多い。しかし本研究では、まず品詞別に語を集計せずに出現頻度上位の unigram 及び bigram を用いて分析を行った。次いで、先にふれたように、研究対象となる文章の内容の影響を受けにくいと考えられる機能語の unigram 及び bigram を用いて分析を行った。機能語とは文章中に

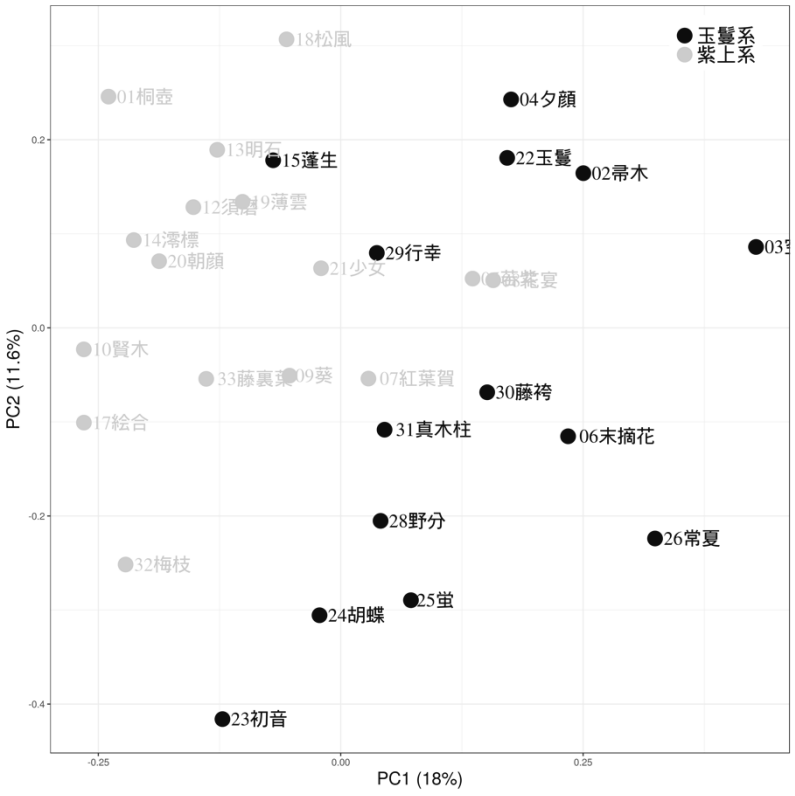

図 1　unigram の出現頻度上位 50 変数についての主成分分析の結果

おいて語彙的意味を担わず文法的機能を担う語であることから、本研究では補助動詞、助詞、助動詞の3品詞を機能語として、頻度を集計した。上掲の『源氏物語』冒頭の文章を例とすると機能語は「の (助詞)」「に (助詞)」「か (助詞)」「給(補助動詞)」「ける (助動詞)」が該当する。また、bigram は隣接する語であることから機能語のbigram は「に (助詞)─か (助詞)」「給 (補助動詞)─ける (助動詞)」などとなる。このような品詞及び単語の n-gram の集計を巻別に行った。ただし、『源氏物語』の各巻の延べ語数は一様ではないことから、各 n-gram の頻度を分析には用いず、n-gram の総度数に対する割合、すなわち各 n-gram の出現率を求め、これを分析に用いた。また、n-gram の集計に際しては、補助動詞と助動詞の活用する 2 品詞は終止形に直した。このように集計した n-gram の出現率について主成分分析を行った。なお、主成分分析では相関係数行列を用いた。

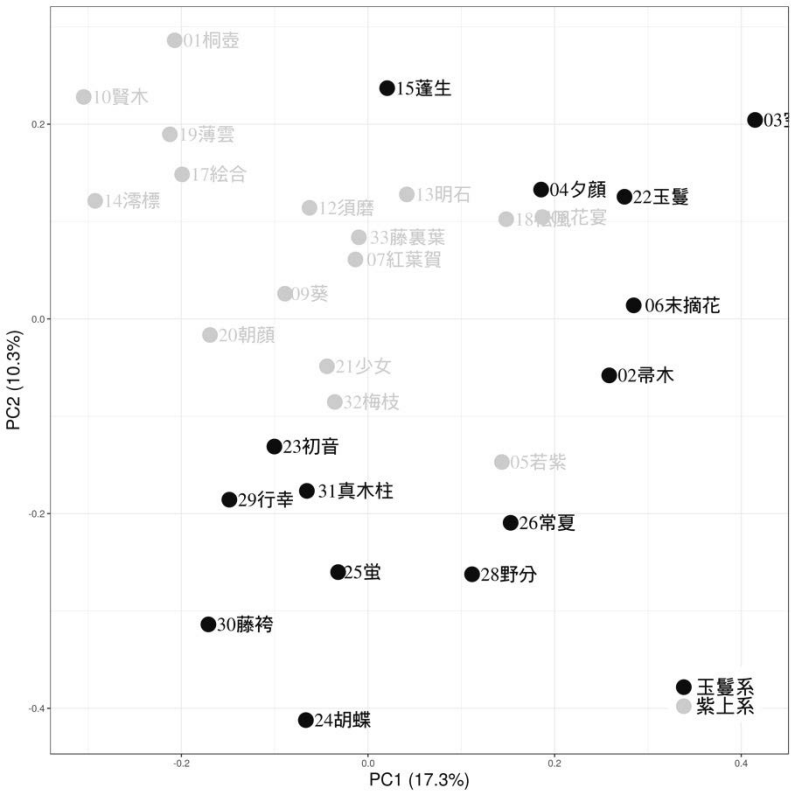

図2　bigram の出現頻度上位 50 変数についての主成分分析の結果

また、本研究においては延べ語数が 1000 語を下回る第 11 巻「花散里」(724 語)、第 16 巻「関屋」(934 語) 、第 27 巻「篝火」(653 語) を分析から除外した。

3.3 分析結果

まず、unigram の出現頻度上位 50 変数について主成分分析を行った。上位 50 変数は頻度が 422 以上の語が該当し、上位 50 変数までの累積度数は総度数の 54.0%である。この出現頻度上位 50 変数の unigram には名詞や動詞と言った出現傾向が物語の内容に影響を受ける語彙も含まれている。しかし、上位 50 変数に含まれる名詞は「こと」「ひと」「ほど」「こころ」「さま」「もの」「よ」の 7 語であり、一般的かつ抽象的な語であると考えられることから、物語の筋によって出現頻度が大きく影響を与えられないことが推測される。動詞についても、「あり」「おもふ」「おぼす」「みる」「す」などの語が含まれ、副詞は「いと」が含まれるのみである。図 1 は主成分分析によって求められた第 1 主成分と第 2 主成分の主成分得点の散布図である。図 1 において、紫上系に属する第 5 巻「若紫」及び第 8 巻「花宴」の 2 巻が玉鬘系に接近し、玉鬘系に属する第 15 巻「蓬生」及び第 29 巻「行幸」の 2 巻が紫上系に接近して位置しているが、紫上系の諸巻と玉鬘系の諸巻は第 1 主成分において概ね分離され付置されていることが分かる。紫上系においては「若紫」と「花宴」に第 7 巻「紅葉賀」を加えた 3 巻の第 1 主成分得点が正になるが他の巻は負になり、その一方で、玉鬘系の諸巻は「蓬生」、第 23 巻「初音」、第 24 巻「胡蝶」の 3 巻の第 1 主成分得点は負になるが他の巻は正になる。なお、第 1 主成分

の寄与率は 18.0%、第 2 主成分の寄与率は 11.6%である。

次いで、bigram の出現頻度上位 50 変数について主成分分析を行った。上位 50 変数は頻度が 173 以上の bigram が該当し、上位 50 変数までの累積度数は総度数の 9.4%である。図 2 は主成分分析の結果であり、図 1 と同様に紫上系に属する第 5 巻「若紫」及び第 8 巻「花宴」が玉鬘系の諸巻に接近して付置されている。また、unigram の出現頻度上位 50 変数についての分析結果とは異なり、図 2 においては第 18 巻「松風」も玉鬘系の諸巻に接近して位置している。しかし、紫上系の諸巻は玉鬘系の諸巻とは異なり主に図中の第 2 象限に付置されていることから、bigram の分析においても第一部は 2 群に分類され得ると考えられる。なお、第 1 主成分の寄与率は 18.0%、第 2 主成分の寄与率は 11.6%である。また、上述したように bigram の出現頻度上位 50 変数までの累積度数の総度数に対する割合は 9.4%と低いが、変数を増加させても分析結果は大きく変わらない。

次に、機能語について分析を行った。先にふれたように、本研究では補助動詞、助詞、助動詞を分析に用いた。機能語の unigram の出現頻度上位 50 変数について主成分分析を行った。上位 50 変数は頻度が 179 以上の語が該当し、上位 50 変数までの累積度数は総度数の 98.3%である。主成分分析の結果は図 3 に示す通りであり、紫上系の属する多くの巻は第

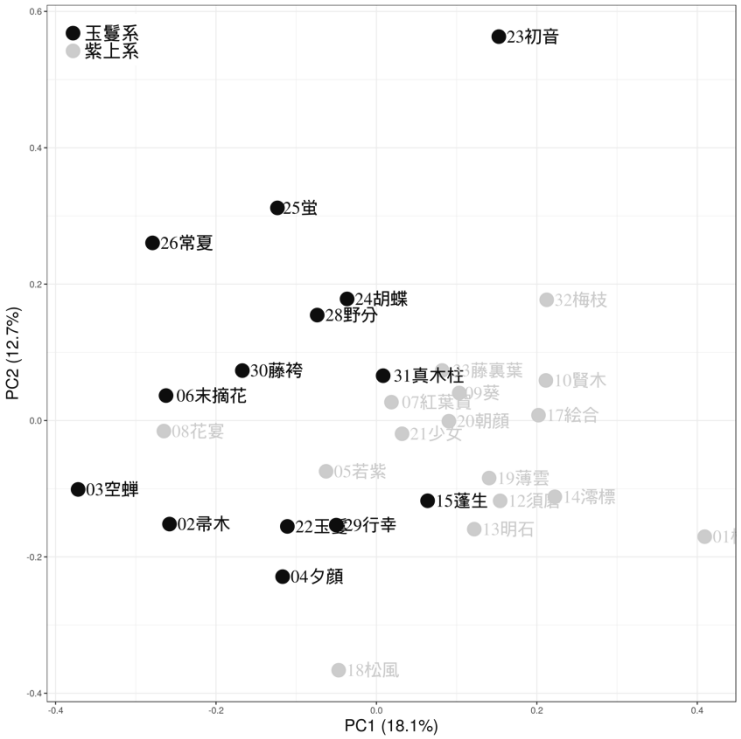

図 3　機能語 unigram の出現頻度上位 50 変数についての主成分分析の結果

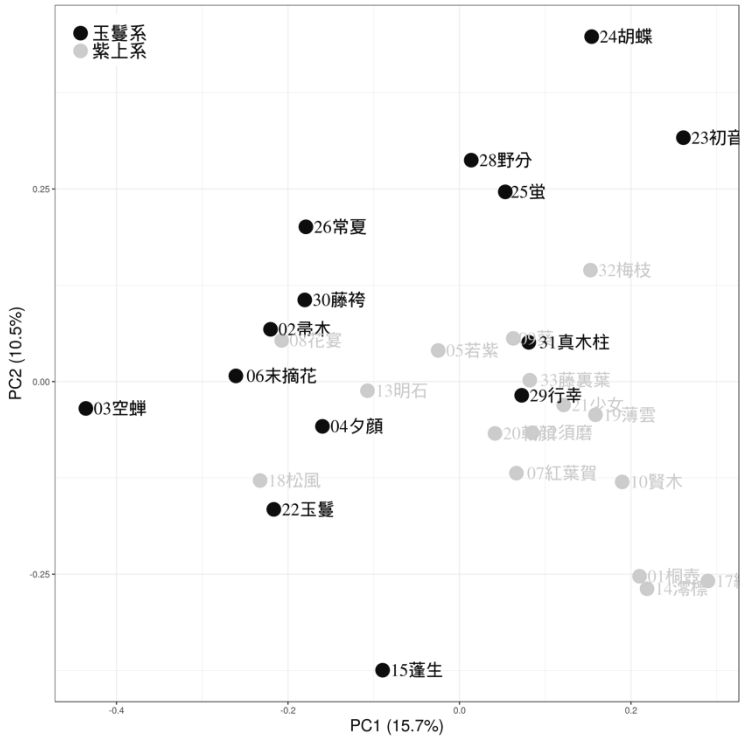

図 4　機能語 bigram の出現頻度上位 50 変数についての主成分分析の結果

1 主成分の正の領域に、玉鬘系の諸巻は負の領域に付置されており、第一部の諸巻は概ね紫上系と玉鬘系で分離して位置していると考えられる。なお、第 1 主成分の寄与率は 18.1%、第 2 主成分の寄与率は 12.7%である。

最後に、機能語の bigram の出現頻度上位 50 変数について主成分分析を行った。上位 50 変数は頻度が 110 以上の bigram が該当し、上位 50 変数までの累積度数は総度数の 49.9%である。図 4 は主成分分析の結果であり、紫上系の多くの巻は主に図中の第 4 象限に位置している。よって、機能語の bigram の分析結果から紫上系の諸巻と玉鬘系の諸巻は異なる量的傾向を有していると考えられる。ただし、紫上系の第 8 巻「花宴」、第 13 巻「明石」、第 18 巻「松風」は玉鬘系の諸巻と混在し、玉鬘系の第 29 巻「行幸」及び第 31 巻「真木柱」は紫上系の諸巻と混在している。なお、第 1 主成分の寄与率は 15.7%、第 2 主成分の寄与率は 10.5%である。

4. 考察

作家が文章を執筆する上でどのような語を多く用いるか、ということはその作家の習慣的特徴の現れであると考えられる。本研究において用いた出現頻度上位の語の unigram 及び語の bigram は作家が文章を執筆する際の習慣的特徴であり、文体的形式的特徴を規定する要素の 1 つであると思われる。本研究において検討を加えた項目は『源氏物語』にあらわれる出現頻度上位の unigram 及び bigram、そして機能語に限定した unigram 及び bigram である。

各 n-gram の出現頻度上位 50 変数について主成分分析を行った結果、どの分析結果においても紫上系の諸巻と玉鬘系の諸巻は一部の巻が混在して付置されるものの概ね異なる傾向を有していると考えられる。従って、計量的な判断に基づくならば、文体的形式的特徴は相違していると考えられる。

参考文献

[1] 池田亀鑑. 源氏物語の構成 (新講源氏物語（上）所収). 至文堂, 1951.
[2] 武田宗俊. 源氏物語の研究. 岩波書店, 1954.
[3] 土山玄・村上征勝. 源氏物語と宇津保物語における語の使用傾向について. じんもんこん 2011 論文集, 2011(8), 125-132, 2011.
[4] 村上征勝・今西祐一郎. 源氏物語の助動詞の計量分析. 情報処理学会論文誌, 40(3), 774-782, 1999.
[5] 小野洋平. 源氏物語成立論の統計科学的再考察 : 村上・今西(1999)を中心に. 計量国語学, 29(8), 296-312, 2015.
[6] 上田英代・村上征勝・今西祐一郎・樺島忠夫・上田裕一.源氏物語語彙用例総索引―自立語編―. 勉誠出版, 1994.
[7] 上田英代・村上征勝・今西祐一郎・樺島忠夫・上田裕一. 源氏物語語彙用例総索引―付属語編―. 勉誠出版, 1996.
[8] 池田亀鑑.源氏物語大成 索引篇, 中央公論社, 1985.