

複合的な語彙に関する情報による平安時代の文献の分類

土山 玄
同志社大学大学院 文化情報学研究所
日本学術振興会 特別研究員 DC

村上 征勝
同志社大学 文化情報学部

計量文献学において、研究対象となる文献の文体について統計的な手法を用い分類する研究は、従来から為されているが、現代文を対象とした研究に比べ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。そこで本研究では、平安時代に著された『源氏物語』および『宇津保物語』を対象とし、語彙に関する情報を用い、作者の相違に起因すると考えられる文体的相違を明らかにし、両作品の諸巻を作品別に分類する。分析に用いた語彙に関する情報とは語の頻度、語の長さ、品詞構成比率、語彙の豊富さの4つの指標である。分析の結果、これらの指標を単独に用いるのではなく、複合的に分析することで、両作品の文体的相違が顕著になる。

Quantitative Analysis of Style Differences in Classical Japanese Literatures using Multi-Lexical-Level Textual Measurements

Gen Tsuchiyama
Graduate School of Culture
and Information Science
Doshisha University

Masakatsu Murakami
Faculty of Culture and Information
Science
Doshisha University

Research Fellow of Japan Society
for the Promotion of Science

We investigate some style differences between authors of Japanese classical literatures by using statistics method. To this end, we analyze four lexical-level textual measurements: word frequency, word-length distribution, part-of-speech frequency, and vocabulary richness. For the analysis, we consider two works, i.e., “Tale of Genji” and “Tale of Utsubo,” which were written in the Heian period (794–1185); both original (and signed) manuscripts were lost and hence only passed on as transcriptions. Therefore, we use the revised scripts for the analysis. The results reveal that there are noticeable differences of word usage between classical authors.

1. はじめに

文献を、その文体に基づいてジャンル別、あるいは著者別に分類するとき、研究方法はおおよそ2つに大別される。1つは国文学や国語学といった人文学の方法に基づく研究であり、このような研究では対象となる文献の記述内容を精読および考察することにより文献の特徴に検討を加え、文献を分類する。これを換言すると、人文学の方法は対象文献の質的側面に着眼していると言える。これに対して、研究対象となる文献の文体、特に文体の量的側面に焦点を当てる計量文献学は、計数可能な文体的要素(文体素)を分析対象とし、主に統計学や情報学の手法を用いることで、文献の分類を行う。

このような計量文献学において、現代文を対象とした文体研究は広く行われており、多くの研究成果が報告されている[1]。しかし、その一方で、古典文を対象とした計量的な文体研究は、十分に展開されているとは言いがたい。これは校定本文の取り扱いに十分な注意が必要なこと、および現

代文に比べて形態素解析が困難であることに起因すると考えられる。

また、古典文学作品は『源氏物語』を筆頭に、人文学領域では非常に古くから研究されており、1000年に及ぶ研究史を有している。これらの研究の中に、計量文献学の手法が有効であると考えられる研究テーマも散見され、その1つに原著者の推定がある。これは、作者未詳の文献あるいは複数作者説が論じられている文献を対象とし、計量的な手法を用いて作者の推定を行うことを目的とする。このような研究の嚆矢として、欧米では Charles Dickens (1812-1870)、William Makepeace Thackeray (1811-1863)、John Stuart Mill (1806-1873) の3人の文章を対象とし、計量的な分析を行うことで、作者によって用いる語の長さ(文字数)の分布が異なることを示した研究[2]が著名である。

しかし、日本における古典文学作品をこのような計量文献学の研究対象とするとき、現代文の研究や欧米の研究と同列に扱うことは難しい。これは、古典文学作品の多くにおいて、作者の自筆原

稿が散逸し書写によってのみ現代に継承されていることによる。日本の中世において、書写とは厳密な行為ではなく、書写者によってオリジナルの文章が改変されていることが十分にあり得る[3]。それゆえ、『源氏物語』のように書写が繰り返された作品は、作者自身の文体的特徴が書写によって希釈され、書写者の文体的特徴が混入している可能性が考えられる。つまり、このような作品に計量的な分析手法が、どこまで有効か検討する必要が想定される。

そこで本研究は、このような日本の古典文学作品が有する事情を考慮し、作者の相違することが明らかな『源氏物語』と『宇津保物語』を対象とし、計量文献学的手法が、古典文学作品においても文体の相違から作者を識別することが可能であることを示すことを目的とする。分析対象として用いる『源氏物語』および『宇津保物語』は平安時代に成立し、現存最古の長編物語に類せられる和文体の物語である。『源氏物語』については、一部の諸巻について複数作者説が論じられているが、平安時代の女流作家として有名な紫式部(973-1014)の手によるものであるとされる。他方、『宇津保物語』の作者については詳細が詳らかにされていないが、源順(911-983)が作者として想定されることが多い。

分析においては、語彙に関する情報に注目し、語の頻度、語の長さの分布、品詞構成比率、語彙の豊富さの4つの指標を用いる。これらの指標の詳細については後述する。

また、結論を先に述べると、語の頻度、語の長さの分布、品詞構成比率、語彙の豊富さと言った4つの指標を単独に用い、それぞれを分析することより、これら4つの指標を複合し、分析を行ったときに、作者の相違に起因すると考えられる文体的相違が顕著になる。加えて、『源氏物語』は相対的に長い語、すなわち文字数の多い語を多用し、形容詞および形容動詞の2品詞については構成比率が高く、また形容詞については語彙も豊富である。一方、『宇津保物語』は漢文訓読文体において使用される機能語が特徴的に用いられ[4]、相対的に文字数の少ない動詞を多用し、名詞や動詞の構成比率が高く、名詞や動詞の語彙が豊かであるという傾向が明らかになった。

2. 資料

上述したように、『源氏物語』の自筆原稿は藤原定家(1162-1241)の時代にはすでに散逸しており、現在においては写本によって受け継がれるのみである。『源氏物語』の写本系統は青表紙本系、河内本系、別本と3系統に大別される。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』[5]および『源氏物語語彙用例総索引 付属語編』[6]を電子化したデータベースを分析に利用した。『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。なお、形態素解析については、『源氏物語大成』[7]の単語認定基準に準拠している。

同様に、『宇津保物語』の自筆原稿も散逸しており、写本によって継承されている。本研究では尊経閣蔵前田家十三行本を底本とする『宇津保物語 全 改訂版』[8]を『源氏物語』と同一の単語認定基準を用いて構築されたデータベースを分析に使用した。

3. 関連研究

『源氏物語』については計量的な研究が為されており、安本[9]や村上・今西[10]が著名である。安本[9]は、日本における文章の計量分析としては草創期の研究であり、『源氏物語』において他作者説が論じられる「宇治十帖」と称される10巻について、仮説検定を行い「宇治十帖」の作者は他44巻と同一作者であるとは言い難いと結論づけている。次いで、村上・今西[10]では、『源氏物語』の助動詞の出現率について多変量分析を行い、『源氏物語』の成立順序の推定している。

また、土山・村上[4]では『源氏物語』および『宇津保物語』に用いられている語の出現率について統計的に分析することで、現代文や英文の研究における指摘[11][12][13]と同様に、両作品の間に助詞、助動詞といった機能語の使用傾向に顕著な相違があることを報告している。また、これに加え、漢文訓読文体の形式の1つである「して」や漢文訓読文体の直喩の表現形式である「ごとし」が『宇津保物語』に相対的に多用されることも指摘している。

4. 分析

4.1 分析対象

分析において、『源氏物語』および『宇津保物語』の1巻を1つの対象とした。『源氏物語』は全54巻、『宇津保物語』は全20巻であるので、本研究で扱う総対象数は74となる。

4.2 分析項目

先述のように本研究では、語彙に関する情報に注目し、これを分析する。分析において採り上げる語彙に関する情報は語の頻度、語の長さの分布、品詞構成比率、語彙の豊富さの4つの指標である。

まず、語の頻度については、土山・村上[4]と同様に、助詞および助動詞といった機能語(付属語, **function words**)を採り上げた。機能語とは文中にあって文法的機能を担う語彙を意味する。一般に、名詞や動詞といった文中において語彙の意味を担う実質語(自立語, **content words**)の出現率は、対象となる文献の記述内容の影響を受けやすいとされるが、機能語の出現率は作者の識別に有効であると考えられている[11][12]。また、本研

究において分析対象とする各対象、つまり両作品の各巻は延べ語数が一様ではなく、各巻におけるばらつきが大きいいため各見出し語の出現頻度ではなく、各巻の延べ語数で基準化した各見出し語の出現率を分析に使用した。

次に、語の長さ分布については、文章を対象とした研究の嚆矢とされる Mendenhall[2]において採り上げられた指標であり、研究方法における問題点が指摘されているが[14]、分析対象となる作品のジャンルが同一である場合、作者の識別に有効であることが報告されている。また、後述する理由から本研究では、機能語は分析に使用せず、実質語の語の長さの分布のみを用いた。また、語の頻度の分析において、実質語は記述内容の影響を受けると考えられることから分析に使用しなかったが、語の長さを分析対象とすることで、語彙的意味を考慮する必要がなくなり、実質語については語の頻度に比べ、記述内容に影響を受けにくい、より抽象度の高いデータとして扱うことが可能となる。

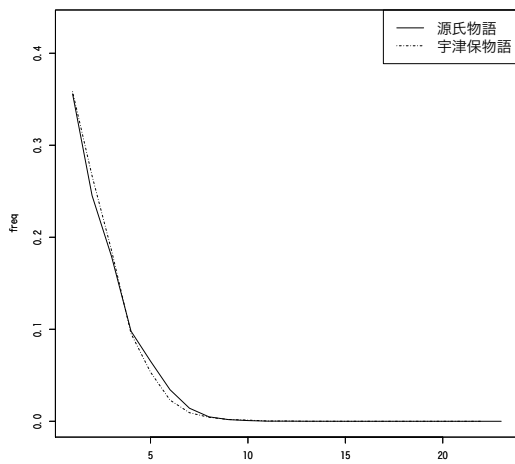


図 1 全語彙の語の長さの分布

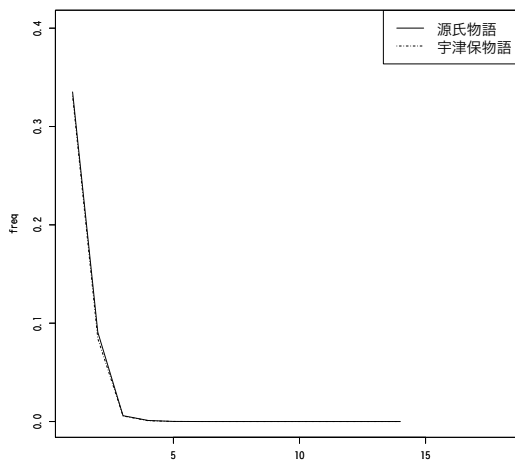


図 2 機能語の語の長さの分布

本研究において対象となる作品は書写によって継承されていることより、写本間あるいは同一写本内においても、漢字表記と仮名表記というように語の表記にゆれが認められることを考慮し、すべての語彙を仮名に直し、仮名の文字数を単語の長さとした。また、図 1, 2 に示すように機能語の 98%以上の語の長さが 1 文字あるいは 2 文字であり両作品の分布に大きな相違が認められないことから、機能語を分析から除外した。なお、図 3 に示すように実質語の語の長さの分布は、両作品の分布の間に相違が認められる。

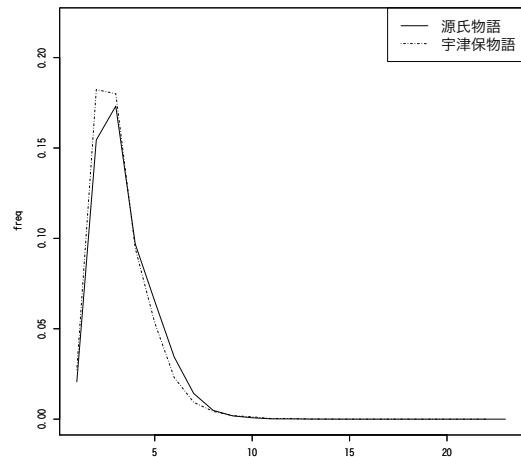


図 3 実質語の語の長さの分布

第 3 に、品詞構成比率については作品によって相違することが知られている[15][16]。本研究においては名詞、代名詞、動詞、補助動詞、形容詞、形容動詞、副詞、接続詞、感動詞、連体詞、助詞、助動詞といった 12 品詞を分析に用い、これらの構成比率を巻ごと、つまり対象ごとに求めた。

最後に、語彙の豊富さについては多くの指標が提案されており、なかでも K 特性値 (characteristic K)[17]は著名である。本研究では作者の識別に最も有効な指標とされる Tuldava の LN を用いた[18][19]。

LN は下記の数式によって求められる。

$$LN = (1 - V^2) / (V^2 \log N)$$

V : 異なり語数, N : 延べ語数

分析に用いた品詞は名詞、代名詞、動詞、補助動詞、形容詞、形容動詞、副詞、接続詞、感動詞、連体詞の 10 品詞であり、品詞構成比率と同様に巻ごとに語彙の豊富さを求めた。ここにおいて、助詞および助動詞を除外した理由については語の長さの分布の分析と同様である。

4.3 分析結果と考察

4.3.1 語の頻度

語の頻度の分析では、助詞および助動詞の各見

出し語の出現率を変数とし、相関係数行列に基づく主成分分析を行った。助詞と助動詞の異なり語数は順に 87 語、38 語である。

まず、助詞についての分析の結果、図 4 に示すように出現頻度上位 38 語を主成分分析に用いたとき、『源氏物語』と『宇津保物語』の諸巻は分離して付置される。なお、横軸が第 1 位主成分、縦軸が第 2 主成分である。出現頻度上位 38 語は両作品を通じた出現頻度が 193 回以上の語である。また、第 2 主成分までの累積寄与率は 30.8% であり、図 4 における実線と破線はそれぞれ『源氏物語』と『宇津保物語』の諸巻についての正規分布を仮定した 95%信頼区間を表す領域である。『源氏物語』の 95%信頼区間と『宇津保物語』の 95%信頼区間はわずかに重複するものの、『源氏物語』の 95%信頼区間に『宇津保物語』の諸巻は 1 つも付置されず、逆もまた同様である。なお、主成分分析に用いる変数を 38 語より増やし、両作品を通じた出現頻度が 3 回以上の語、つまり出現頻度上位 62 語まで変数を増やしても分析結果は大きく変わらない。しかし、変数を 38 語より減らしたとき、および 62 語より増やしたとき、どちらかの作品の 95%信頼区間にもう一方の作品の諸巻が付置される。

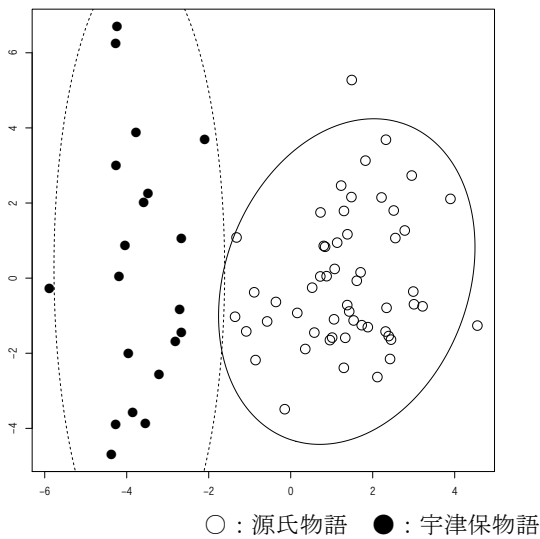


図 4 助詞出現頻度上位 38 語の主成分分析結果

次に、助動詞についての分析結果を図 5 に示す。図 5 は助動詞の出現頻度上位 24 語についての主成分分析の結果である。出現頻度上位 24 語は両作品を通じた出現頻度が 23 回以上の語である。第 2 主成分までの累積寄与率は 28.5% であり、図 4 と同様に実線と破線はそれぞれの作品の 95% 信頼区間である。助詞の出現頻度上位 38 語についての分析結果と同様に、それぞれの 95% 信頼区間にもう一方の諸巻は付置されない。これは助動詞の出現頻度上位 25 語、26 語についての分析

結果でも同様である。しかし、それ以外においてはどちらかの作品の 95%信頼区間にもう一方の作品の諸巻が付置される。

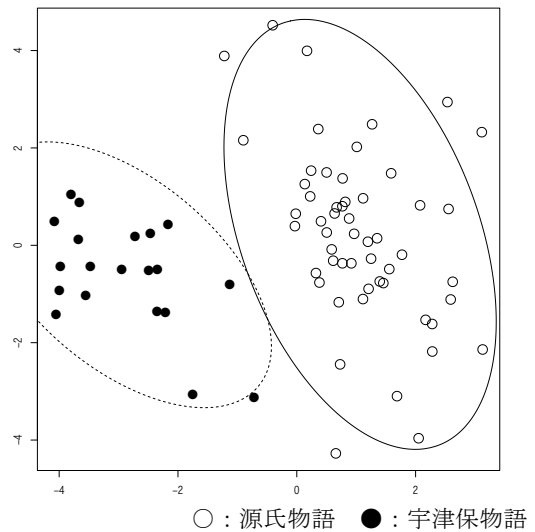


図 5 助詞出現頻度上位 24 語の主成分分析結果

助詞および助動詞に対する分析の結果、『源氏物語』と『宇津保物語』の作者の相違に起因すると考えられる文体的相違が明らかになった。また、両作品の間において顕著に出現傾向が相違する語彙、つまり特徴語については土山・村上[4]において報告されているように、一般に両作品ともに和文体の物語とされるが、漢文訓読文体の文献に多用される「して」や「ごとし」が『宇津保物語』の特徴語として抽出される。

4.3.2 語の長さの分布

語の長さの分布の分析においては、上述のように機能語を分析から除外し、巻ごとに名詞、代名詞、動詞、補助動詞、形容詞、形容動詞、副詞、接続詞、感動詞、連体詞といった 10 品詞において品詞別に語の長さを集計したデータ、およびこれら 10 品詞を区別することなく実質語として語の長さを集計したデータを加え、合計 11 のデータについて主成分分析を行った。

これらのデータに対する分析の結果、動詞を除く 9 品詞および実質語についての主成分分析においては、図 6 に示すように『源氏物語』あるいは『宇津保物語』の 95%信頼区間にもう一方の作品の諸巻が付置されたが、動詞の語の長さ分布についての分析においては両作品が分離して付置された。作者の相違が動詞の長さの分布についての分析において顕著に現れることについては、現代文に対する分析結果と合致する[20]。

動詞の長さの分布に対する分析では、図 7 に示すように語の長さが 1 文字から 6 文字までの語を集計した、つまり 6 変数による主成分分析におい

て、両作品が分離して付置され、両作品の 95% 信頼区間は重複しない。第 2 主成分までの累積寄与率は 69.9% である。主成分分析に用いる変数を 6 変数から増やし、1 文字から 10 文字までの語を集計しても分析結果は大きく変わらないが、それ以外ではどちらかの作品の 95% 信頼区間にもう一方の作品の諸巻が付置される。6 変数による主成分分析の主成分負荷量は表 1 に示す通りである。

また、動詞の語の長さについて、11 文字以上の語の頻度は極端に低くなり、11 文字の語は 74 の対象うち 65 の対象において、頻度が 0 回である。したがって、11 文字以上の語についてはスパースであると言える。

表 1 1 から 6 文字までの動詞の主成分負荷量

	PC1	PC2	PC3	PC4	PC5	PC6
3	0.346	-0.093	0.844	-0.207	0.276	-0.200
2	0.434	-0.267	-0.155	0.756	0.378	-0.040
4	-0.241	-0.947	0.031	-0.123	-0.122	0.117
5	-0.468	-0.009	-0.188	-0.173	0.760	-0.371
1	0.457	-0.040	-0.299	-0.493	0.370	0.565
6	-0.455	0.145	0.370	0.312	0.225	0.698

よって、語の長さの分布についての分析から、『源氏物語』は 5 文字あるいは 6 文字の動詞を多用し、『宇津保物語』は 1 文字から 3 文字までの動詞を多用していると言える。

4.3.3 品詞構成比率

品詞構成比率に対する主成分分析では、上述のように 12 品詞を分析に用いた。主成分分析においては出現頻度に応じて変数を増減させたが、語の頻度や語の長さの分布についての分析と異なり、『源氏物語』および『宇津保物語』の 95% 信頼区間が常に重複した。最も両作品が分離して付置された分析結果は図 8 に示す 12 品詞すべてを用いた分析である。第 2 主成分までの累積寄与率は 57.4% であり、主成分負荷量は表 2 に示す通りである。

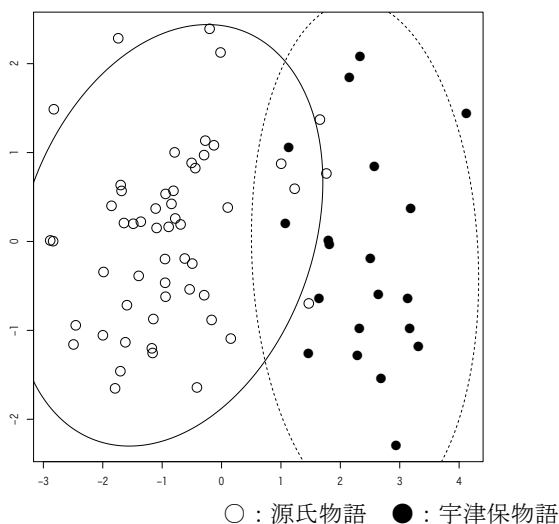


図 6 実質語の長さの分布の主成分分析結果

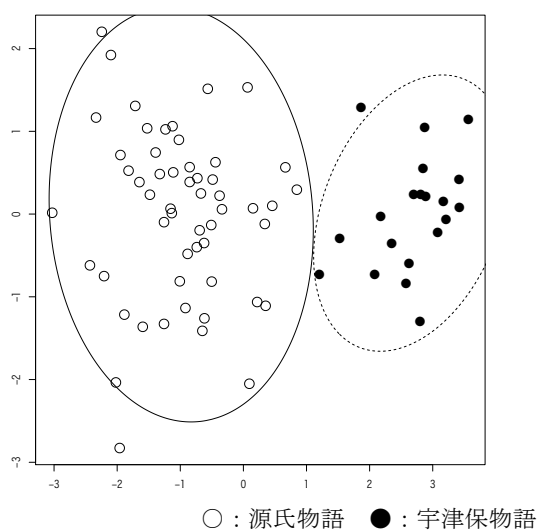


図 7 1 から 6 文字までの動詞の主成分分析結果

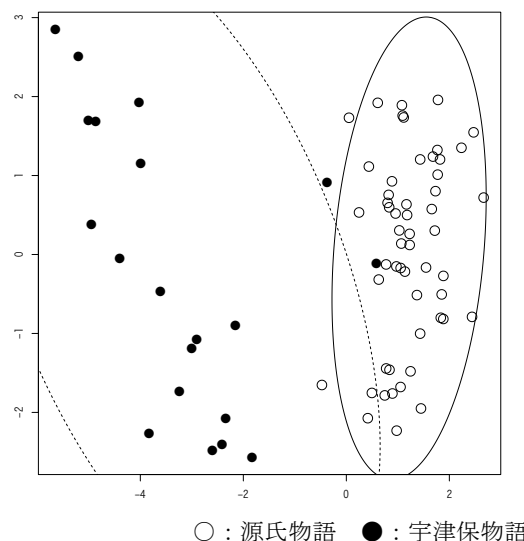


図 8 品詞構成比率の主成分分析結果

品詞構成比率についての分析においては、明確に『源氏物語』と『宇津保物語』の諸巻は分離して付置されなかったが、主成分分析の結果から『源氏物語』は形容詞、形容動詞、副詞と言った相の類とされる品詞[21]の構成比率が高く、他方、『宇津保物語』は名詞、代名詞、動詞、接続詞といった品詞の構成比率が高いと言える。

表 2 品詞構成比率の主成分負荷量

	PC1	PC2	PC3	PC4	PC5	PC6
助詞	0.03	-0.38	-0.29	-0.43	0.71	-0.07
名詞	-0.39	0.31	-0.02	-0.11	-0.07	-0.18
動詞	-0.18	-0.17	0.59	0.37	0.20	0.10
助動詞	0.28	-0.37	0.34	0.03	0.05	0.28
形容詞	0.41	0.10	0.12	0.00	-0.11	-0.19
補助動詞	0.25	0.09	-0.51	0.23	-0.09	0.60
副詞	0.35	-0.27	-0.10	0.08	-0.26	-0.20
形容動詞	0.40	0.11	0.00	0.12	0.02	-0.19
連体詞	-0.08	-0.47	0.04	-0.46	-0.57	-0.07
代名詞	-0.30	-0.37	-0.07	0.19	-0.09	0.37
接続詞	-0.34	0.00	-0.19	-0.01	-0.18	0.03
感動詞	-0.13	-0.37	-0.34	0.59	0.02	-0.50

4.3.4 語彙の豊富さ

上述のように語彙の豊富さの分析においては、機能語を除く 10 品詞を分析の対象とした。また、語の頻度、語の長さの分布、品詞構成比率の分析では、それぞれ該当する分析項目の出現頻度上位から順に変数として分析に用いたが、語彙の豊富さに対する分析では『源氏物語』および『宇津保物語』の全巻、つまり分析に用いる全対象を通じて算出した語彙の豊富さの指標の値が大きい品詞から順に分析に用いた。両作品の全巻を通じた語彙の豊富さについては表 3 に示す。

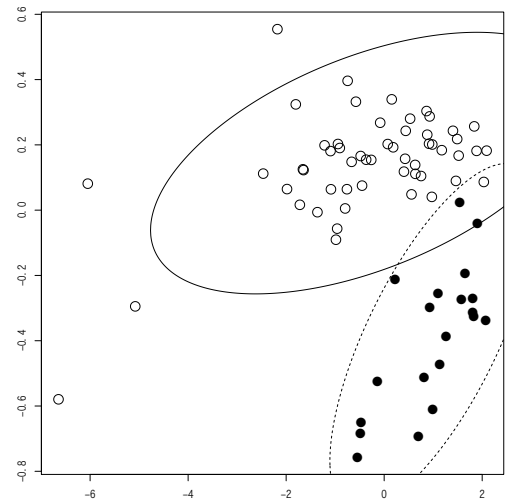
表 3 全対象を通じた語彙の豊富さ

	Tuldava's LN
名詞	-0.059
動詞	-0.060
形容詞	-0.067
補助動詞	-0.067
副詞	-0.069
形容動詞	-0.074
連体詞	-0.081
代名詞	-0.081
接続詞	-0.104
感動詞	-0.108

主成分分析の結果、品詞構成比率の分析と同様に『源氏物語』および『宇津保物語』の 95%信頼区間が常に重複した。最も両作品が分離して付置された分析結果は図 9 に示す名詞、動詞、形容詞の 3 品詞を用いた分析である。第 1 主成分の寄与率は 96.8%と非常に高く、第 2 主成分までの累積寄与率は 99.7%である。

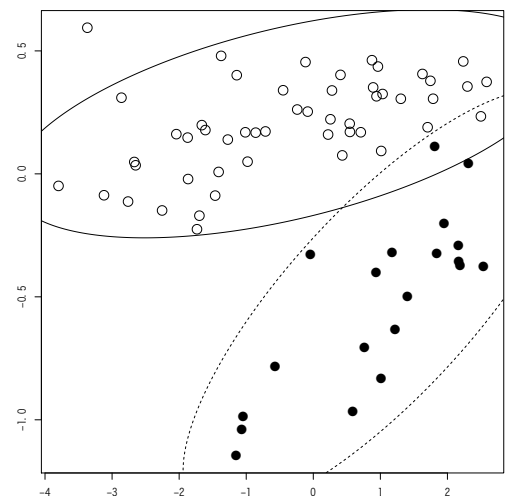
また、図 9 において、横軸の負の領域に外れる 3 巻は『源氏物語』の第 11 巻「花散里」、第 16 巻「関屋」、第 27 巻「篝火」であり、これらは他の諸巻に比べ、大きく延べ語数が少ない巻である。

これら 3 巻の延べ語数は順に 724 語、934 語、653 語であり、両作品を通じ延べ語数が 2000 語を下回るのはこの 3 巻のみである。ゆえに、これら 3 巻を分析から除外し、改めて主成分分析を行った結果を図 10 に示す。分析の結果、図 9 に示した結果と同様に名詞、動詞、形容詞の 3 品詞を用いた分析結果において、両作品の 95%信頼区間は重複するものの、両作品の諸巻は比較的に分離して付置された。また、表 4 に示した主成分負荷量から第 2 主成分において、名詞および動詞は負の値であり、形容詞は正の値であることから、『源氏物語』は相対的に形容詞の語彙が豊富であり、『宇津保物語』は名詞および動詞の語彙が豊富であると考えられる。これは品詞構成比率の分析における考察の類似する。



○：源氏物語 ●：宇津保物語

図 9 3 品詞の語彙の豊富さの主成分分析結果



○：源氏物語 ●：宇津保物語

図 10 3 巻を除外した主成分分析結果

表 4 語彙の豊富さの主成分負荷量

	PC1	PC2	PC3
名詞	0.576	-0.620	0.533
動詞	0.585	-0.143	-0.798
形容詞	0.571	0.772	0.280

4.3.5 考察

以上の分析結果から、計量的観点から『源氏物語』および『宇津保物語』の語彙に関わる文体的特徴を考察すると、『源氏物語』は文字数の長い動詞を相対的に多用し、形容詞および形容動詞の2品詞については構成比率が高く、また形容詞については語彙も豊富である。一方、『宇津保物語』は漢文訓読文体において使用される機能語が特徴的に用いられ[4]、相対的に文字数の少ない動詞を多用し、名詞や動詞の構成比率が高く、名詞や動詞の語彙が豊かであると言える。

5. 分析結果の応用

本研究では語の頻度、語の長さの分布、品詞構成比率、語彙の豊富さといった4つの語彙に関わる情報を分析に用い、『源氏物語』と『宇津保物語』の諸巻について、作者による分類を行った。語の頻度と語の長さの分布においては、両作品の間における顕著な相違が認められ、品詞構成比率と語彙の豊富さにおいては、両作品が相違するという傾向は認められるが、語の頻度と語の長さの分布についての分析に比べ、その相違は顕著ではない。

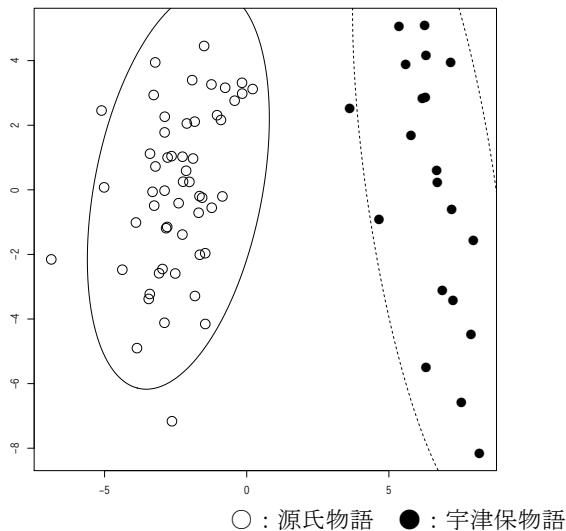


図 11 複合データの主成分分析結果

ここで、これら4つの指標を複合し、分析を行った。すなわち、助詞の出現頻度上位38語、助

動詞の出現頻度上位24語、動詞の1文字から6文字までの語の分布、12品詞の品詞構成比率、名詞、動詞、形容詞の3品詞の語彙の豊富さを複合したデータについて主成分分析を行った。図11に示すように、両作品の95%信頼区間は重複することなく、分析対象は作品別に明確に分離し付置される。

図11に示した分析結果は、『源氏物語』と『宇津保物語』という分類基準が既知であり、それに準じ、語彙に関わる4つの指標について検討した結果を応用したものである。古典文学作品においては、『源氏物語』の終わりの10巻である「宇治十帖」に代表されるように、作者がそれ以前の諸巻と異なるのか否かが明確ではない場合が多い。これはすなわち、分類基準が未知の場合、上述の分析方法を応用することには問題がある。よって、本研究では分類基準が未知であっても対応し得ると考えられる以下の基準を設ける。

- (1) 語の頻度：助詞 全語彙・助動詞 全語彙
- (2) 語の長さの分布：動詞の長さの分布においてスパースにならない文字数
- (3) 品詞構成比率：全品詞
- (4) 語彙の豊富さ：全対象を通じ、語彙が豊富な上位3品詞

本研究においてこれらの基準を満たす複合データは、助詞87語、助動詞38語、1文字から10文字までの動詞の長さの分布、12品詞の品詞構成比率、名詞、動詞、形容詞の語彙の豊富さである。これについての主成分分析の結果は図12に示すように、図11に示した結果と同様に作品別に対象が分離して付置される。また、語彙の豊富さについては複合する変数を増加させても分析結果は大きく変わらない。

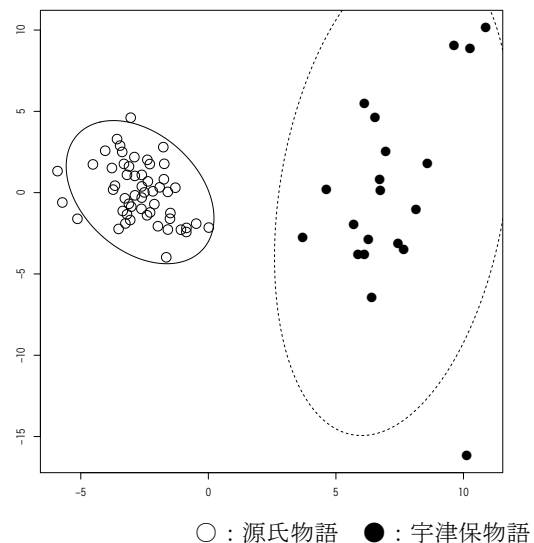


図 12 基準を満たすデータの主成分分析結果

以上の分析から、分類基準が未知であっても語彙に関する情報を複合的に用いる分析は有効であると言える。

6. おわりに

古典文学作品において、上述した『源氏物語』の「宇治十帖」に代表されるように物語の一部の諸巻に他作者説が提起されている古典文学作品は少なくない。このような問題について、本研究で用いた語彙に関する情報についての複合データに対する計量的な分析は有効な方法となることが期待され、仮に検討の対象となる作品において複数作者説が正しいのであれば、作者の相違に応じて、語彙に関わる文体的相違を明らかにすることが可能であると言える。

謝辞

本研究の成果の一部は、日本学術振興会科学研究費補助金(特別研究員奨励費)「『源氏物語』の成立論および多作者説についての計量的研究」(代表：土山玄)(2013-2014年度)によるものである。

参考文献

- [1] 計量国語学会(編)：計量国語学事典，朝倉書店(2009)。
- [2] Mendenhall, T. C. : The characteristic curves of composition, *Science*, (214S), pp.237-246 (1987)。
- [3] 池田亀鑑：古典学入門，岩波書店(1991)。
- [4] 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，*人文科学とコンピュータ論文集*, Vol.2011, No.8, pp.125-132 (2011)。
- [5] 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，勉誠社(1994)。
- [6] 上田英代，村上征勝，今西祐一郎，樺島忠夫，上田裕一，藤田真理：源氏物語語彙用例総索引 付属語編，勉誠社(1996)。
- [7] 池田亀鑑：源氏物語大成 索引篇，中央公論社(1985)。
- [8] 室城秀之：うつほ物語 全 改訂版，おうふう(1995)。
- [9] 安本美典：宇治十帖の作者-文章心理学による作者推定-，*心理学評論*, Vol.2, No.1, pp.147-156 (1957)。
- [10] 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*, Vol.40, No.3, pp.774-782 (1999)。
- [11] 金明哲：助詞の分布における書き手の特徴に関する計量分析，*社会情報*, Vol.11, No.2, pp.15-23 (2002)。
- [12] 金明哲：助詞の n-gram モデルに基づいた書き手の識別，*計量国語学*, Vol.23, No.5, pp.225-140 (2002)。
- [13] Holmes, D. I. : Authorship attribution, *Computers and the Humanities*, Vol.28 pp.87-106 (1994)。
- [14] Williams, C. B. : Mendenhall's studies of word-length distribution in the works of Shakespeare and Bacon, *Biometrika*, Vol.62, No.1, pp.207-212 (1975)。
- [15] 大野晋：基本語彙に関する二三の研究-日本の古典文学作品に於ける，*国語学*, Vol.24, pp.34-46 (1956)。
- [16] 水谷静夫：大野の語彙法則について，*計量国語学*, Vol.35, pp.1-13 (1965)。
- [17] Yule, G. U. : The Statistical Study of Literary Vocabulary, Cambridge University Press (1944)。
- [18] Grieve, J. : Quantitative authorship attribution: An evaluation of techniques. *Literary and linguistic computing*, Vol.22, No.3, pp.251-270 (2007)。
- [19] Tuldava, J. : Quantitative Relations between the Size of the Text and the Size of Vocabulary, *Journal of Linguistic Calculus*, Vol.4, pp.28-35 (1977)。
- [20] 金明哲：日本語における単語の長さの分布と文章の著者，*社会情報*, Vol.5, No.2, pp.13-21 (1996)。
- [21] 国立国語研究所(編)：分類語彙表 増補改訂版，大日本図書(2004)。