語の bigram による『源氏物語』の分類

土山 玄　　村上 征勝
同志社大学 文化情報学研究科　　同志社大学 文化情報学部

平安時代に著された『源氏物語』が現行の巻序の通りに執筆されたとすることについては，いくつかの疑問が論じられている．一般に、第 1 巻「桐壺」から第 33 巻「藤裏葉」の成立順序は現行の巻序と相違し、「紫上系」および「玉鬘系」と称される二つの系統に分類されると考えられている．そこで、本研究では語の bigram を対象に両系統の文体に相違があることを計量的なアプローチにより示す。くわえて、ランダムフォレストを用いて、両系統の間において顕著に出現傾向が相違する bigram を抽出し、指摘する。

Clustering of *The Tale of Genji* Based on Word Bigram

Gen Tsuchiyama
Graduate School of Culture and Information Science
Doshisha University

Masakatsu Murakami
Faculty of Culture Information Science
Doshisha University

Some researchers of Japanese Literature have considered that the current order of chapters of *The Tale of Genji* is different to original order, and the first of a three-part series on the story is divided 2 groups; '*Murasaki no Ue Group'* and '*Tamakazura Group'*. Thus, we will investigate difference in style between 2 groups using Principal Components Analysis and Random Forest about word bigram. We will conclude that there is a difference in usage of auxiliary verb.

1． はじめに

1.1. 背景

『源氏物語』は平安時代に著され，各時代を通じてひろく読み継がれてきた古典作品である．『紫式部日記』の記述から，紫式部（973 頃～1014 頃）の手により執筆されたと考えられ，現存最古に類する長編物語である．『源氏物語』は全 54 巻によって構成されるが，現行の巻序の通りに執筆されたとすることについては，いくつかの疑問が論じられている．

また、『源氏物語』は三部構成の物語であると考えられており[1]，第 1 巻「桐壺」から第 33 巻「藤裏葉」までが第一部，第 34 巻「若菜上」から第 41 巻「幻」まで第二部，そして第 42 巻「匂宮」から第 54 巻「夢浮橋」までが第三部とされる．第一部は主人公である光源氏が栄華を極めるまで，第二部は光源氏が死去するまで，そして第三部は光源氏死後の物語である．

このうち，第一部の成立巻序がたびたび議論の俎上に載る．第一部では，第２巻「帚木」の冒頭に「いひ消たれ給ふとが多かんなるに」や「かかる好きごと」などといった記述があるが，第１巻「桐壺」において，これらの記述が指し示すようなことは記述されていない．このことから，「桐壺」と「帚木」は連続していない[2]と論じられている．

また，[2]に加えて，登場人物の出現傾向のばらつきから第一部は２つの系統に分割される[3][4]と論じられており，一般にこの２つ系統は，「桐壺」が属すると考えられる「紫上系」，「帚木」が属すると考えられる「玉鬘系」とそれぞれ称される．「紫上系」および「玉鬘系」に属する諸巻は表１に示す通りである．

[3]および[4]によれば，「紫上系」の登場人物は「玉鬘系」にも出現する一方で，「玉鬘系」の諸巻において初出の登場人物は「紫上系」に出現することはない．このことにより，[4]においては，「紫上系」が先行して構想および記述され，「玉鬘系」が後に記述挿入されたとされる．くわえて，[4]では「紫上系」と「玉鬘系」との間には文体や技巧などといった点においても相違も見出せると論じている．

表１　「紫上系」および「玉鬘系」

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		

また，「紫上系」は本紀の形式を取り入れていること，一方，「玉鬘系」は列伝の形式を取り入れていることが指摘されている[5]．本紀とは年月を追う記述の形式であり，列伝とは個人の行動などの重要な点の記述に主眼をおいた形式である．このようなことから，「紫上系」と「玉鬘系」の文章の間には質的な相違が認められると言える．

1.2. 目的

本研究では『源氏物語』における「紫上系」，「玉鬘系」について，計量的な分析手法を用いることで，２つの系統の間に文体的相違が認められるか検討を加える．くわえて，文体的相違が認められた場合，顕著に相違する要素を抽出し指摘する．

また，分析に際しては，語の bigram を求め，これに対し主成分分析およびランダムフォレストを行った．主成分分析は各グループ間において，語の使用傾向が相違することを検討するため，ランダムフォレストは両系の間で顕著に相違する要素を抽出するために用いた．語の bigram および分析手法の詳細については後述する．

２．資料

本研究では電子化されたデータベースを用いた．データベースのもととなったテキストは『源氏物語語彙総索引』[6]である．

これは『源氏物語大成』[7]の単語分割に従い，『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，『源氏物語』には様々な写本系統があり，これらは一般に青表紙本系，河内本系，別本系の３系統に大別される．上述の『源氏物語大成』は青表紙本系の１つとされる大島本を主たる底本として『源氏物語』の本文を校訂したものである．ゆえに，本研究も大島本を底本としている．

3. 分析

3.1. 対象データ

『源氏物語』は 54 巻によって構成されるので，本研究においては，それぞれ１巻をひとつの分析対象とする．

ただし，『源氏物語』において，第 27 巻「篝火」の総語数が 653 語であるのに対し，第 35 巻「若菜下」が 20223 語であるというように，各巻の総語数に非常に大きなばらつきが認められるため，総語数が 2000 語未満の巻は分析から除外することにする．

したがって，「篝火」に加えて，『源氏物語』の第 11 巻「花散里」（724 語），第 16 巻「関屋」（934 語）を分析から除いた．ゆえに分析の対象数は 51 となる．

3.2. 分析項目

本研究では語の bigram について分析を行う．語の bigram とは隣接する語の組を意味する．たとえば，「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」という『源氏物語』の冒頭の文における語の bigram は「女御更衣(名詞) ＋ あまた(副詞)」，「あまた(副詞) ＋ さふらひ(動詞)」，「さふらひ(動詞) ＋ 給(補助動詞)」という具合に求められる．

また，本研究では出現率の高い品詞である名詞，動詞，形容詞，形容動詞，副詞，助詞，助動詞を分析の対象とし，その他の品詞については分析から除外した．

分析に際しては、名詞の bigram という特定の品詞の語彙のみによる bigram と、名詞と動詞の bigram という２つの品詞の語彙による bigram を用いた。前掲の例文においては、前者の bigram については、「女御更衣 ＋ なか」といった bigram が該当する。後者は「女御更衣 ＋ さふらひ」が該当する。

また、本研究においては名詞と動詞の bigram と動詞と名詞の bigram を区別する。つまり、bigram において語をペアリングする際に品詞の順序を考慮する。一例を示すと、「女御更衣 ＋ さふらひ」は「名詞 ＋ 動詞」の bigram であり、「さふらひ ＋ なか」は「動詞 ＋ 名詞」の bigram である。以上より、本研究においては 49 通りの品詞の組み合わせによる bigram を分析した。

3.3. 分析手法

先にふれたように，語の bigram に対し，統計的な分析手法を用いる．本研究で使用する分析手法は主成分分析およびランダムフォレスト[8]である．

主成分分析は，多次元データに対する次元縮約の手法であり，もとのデータの変数より新たに合成変数を求めることで，情報の縮約を行う．分析結果は主に 2 次元の散布図に対象を付置し，対象間の関係を可視化する．また，各主成分にデータ全体の情報がどれくらい含まれているかは寄与率によって評価される．

一方，ランダムフォレストはアンサンブル学習と称される手法のひとつで，学習データを用い分類器を構築し，未知のデータの母集団を判別する手法である．母集団を判別する手法としてはバギング，ブースティング，サポートベクターマシン（SVM）などあるが，文章の母集団

の判別，すなわち文章の書き手の識別についてはこれらの手法に比べ，ランダムフォレストの精度が最も高いとされる[9].

くわえて，ランダムフォレストは分類器を構築する際に，分類における変数の重要度の推定を行う．本研究において，重要度の高い変数とは，「紫上系」と「玉鬘系」のどちらかの系統に頻出し，他方の系統にはおよそ現れない語を指す.

また，ランダムフォレストは分類器の構築の際に，乱数を発生させることで精度の向上を図っている．しかし，そのために複数回のランダムフォレストを行うと，同一の結果を得られることはおよそない．したがって，本研究ではランダムフォレストを 100 回繰り返し，変数の重要度の推定値には 100 回推定された値の平均を用いる.

3.4. 分析結果

まず，『源氏物語』から助動詞のみを抜き出し，助動詞の bigram を求め，主成分分析を行った．分析に使用した bigram は出現頻度上位の 50 組である．「紫上系」と「玉鬘系」における助動詞の bigram の主成分分析の結果は図 1 に示す通りである．累積寄与率は 24.7%である。分析の結果，「紫上系」と「玉鬘系」との間に助動詞の bigram の出現傾向に相違が認められる.

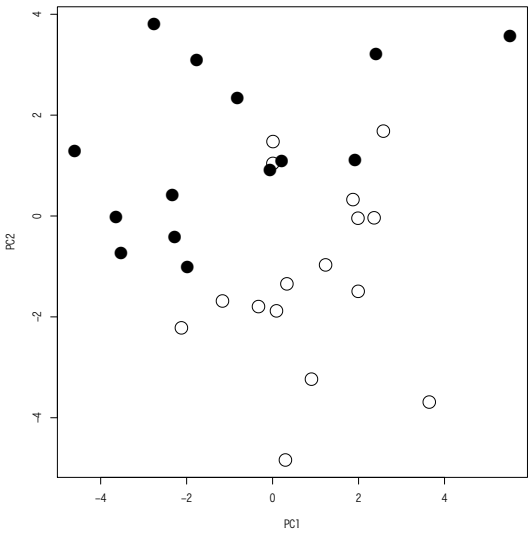

図 1 「紫上系」および「玉鬘系」の諸巻に対する助動詞の bigram の主成分分析の結果

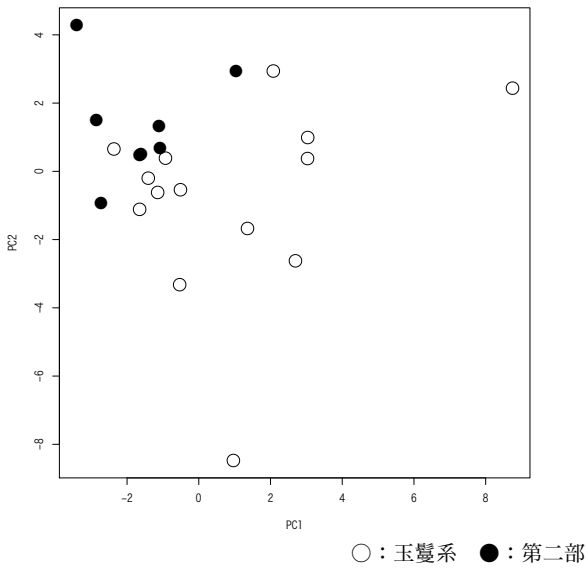

図 2 「玉鬘系」および「第二部」の諸巻に対する助動詞の bigram の主成分分析の結果

なお，両系統が属する「第一部」の直後である「第二部」と「玉鬘系」を対象に出現頻度上位 50 組の bigram に対する主成分分析を行ったところ、分析結果は図 2 のようになった。累積寄与率は 28.7%である。ここでも両グループの間に相違が認められる.

「第二部」においては，「玉鬘系」にのみ出現し，「紫上系」において出現しなかった人物も第 34 巻「若菜上」から合流し，両系の人物が現れる．先にふれたように，「紫上系」は本紀の形式を取り入れており，正確な時間軸に沿って事態が推移していると考えられている．同様に，「第二部」も「年も返りぬ」や「年暮れぬ」という表現が散見され，時間軸を 1 つの指標として有している[5]とされる.

このように、「玉鬘系」の諸巻は「紫上系」および「第二部」と助動詞の bigram の出現傾向が相違していると言える。

したがって，「玉鬘系」の助動詞の用法は『源氏物語』における列伝としての文体的特徴を有している可能性が推測され得る.

ついで，「紫上系」と「玉鬘系」においては，図 3 に示すように助詞と助動詞の bigram の出現傾向にも、助動詞の bigram において認められた相違と同様の相違が認められる．累積寄与率は 25.0%である。なお，先にふれたように、ここにおける bigram は「助詞＋助動詞」のペアのみを意味し，「助動詞＋助詞」のペアは分析の対象としない.

助詞と助動詞の bigram の分析においては「玉鬘系」と「第二部」といったような他の分析の組み合わせに出現傾向の相違は認められなかった。

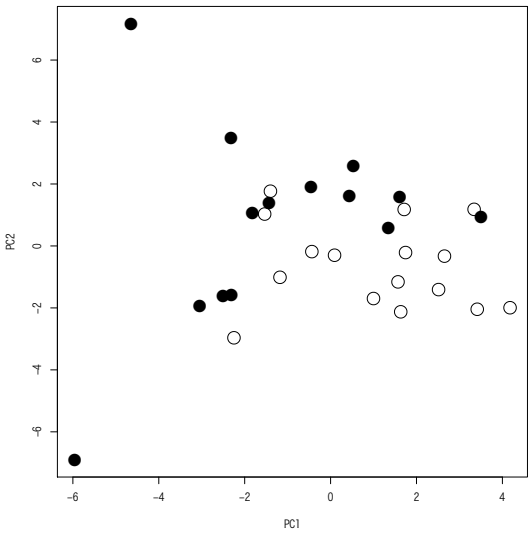

図３ 「紫上系」および「玉鬘系」に対する助詞と助動詞の bigram の主成分分析の結果

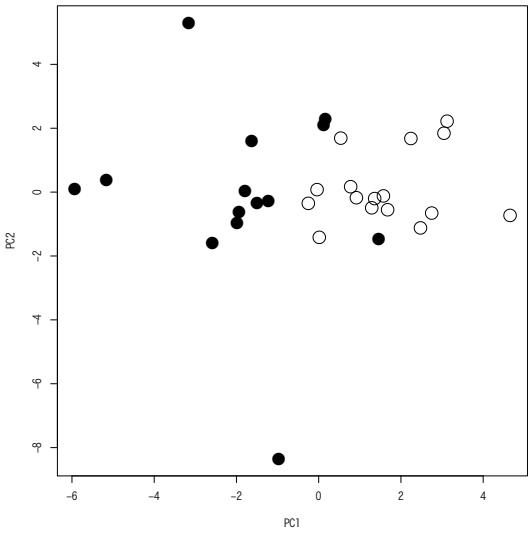

図４ 「紫上系」および「玉鬘系」に対する動詞と助動詞の bigram の主成分分析の結果

動詞と助動詞の bigram に対する主成分分析の結果を図４に示す．累積寄与率は 21.0%である。上述の助詞と助動詞の bigram と同様に，ここにおける bigram は「動詞＋助動詞」のペアのみを意味し，「助動詞＋動詞」のペアは分析の対象としない．分析結果から，動詞と助動詞の bigram については両系の間に出現傾向の相違が認められた．

動詞と助動詞の bigram の分析においても「玉鬘系」と「第二部」といったような他の分析の組み合わせに出現傾向の相違は認められなかった。

助詞と助動詞の bigram および動詞と助動詞の bigram の主成分分析の結果においても「紫上系」と「玉鬘系」との間に bigram の出現傾向に相違が認められた。[4]あるいは[5]において、「紫上系」と「玉鬘系」とでは文章の性質が異なるとされるが、そのような相違が量的な観点による分析によって見出されたと言える。

次に，助詞のみを抜き出し，助詞の bigram を求め，主成分分析を行った．なお，文章を対象とした計量的な分析においては，助詞の bigram に書き手の特徴があらわれるとされている[10][11].

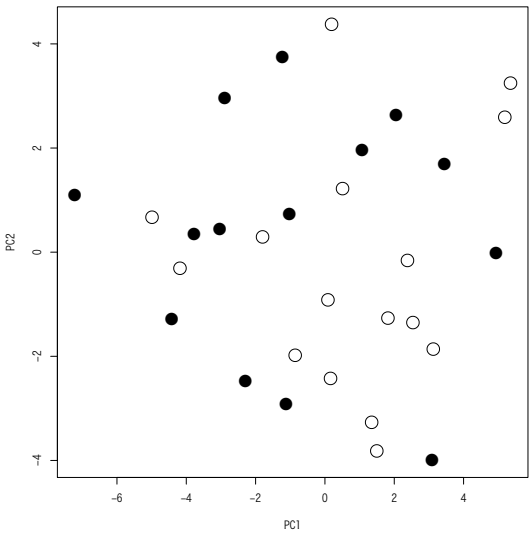

図５ 「紫上系」および「玉鬘系」の諸巻に対する助詞の bigram の主成分分析の結果

「紫上系」と「玉鬘系」における助詞の出現頻度上位 50 組の bigram の主成分分析の結果は図５に示す通りである．累積寄与率は 30.9%である。分析結果から，助詞の bigram については，「紫上系」と「玉鬘系」の間に大きな相違は認められない．なお、同様に「紫上系」と「第二部」，「紫上系」と「第三部」，「玉鬘系」と「第二部」，「玉鬘系」と「第三部」，「第二部」と「第三部」との間においても，出現する語の bigram に大きな相違は認められなかった．

したがって，「紫上系」と「玉鬘系」との間で，語の bigram という観点においては，作者が異なると言えるほどの文体的相違は認められない．ゆえに，助動詞の bigram、助詞と助動詞

の bigram および動詞や助動詞の bigram において認められた相違は，作者の執筆態度の相違に起因すると考えられる．

最後に，「紫上系」と「玉鬘系」の間において，顕著に使用傾向が相違する助動詞の bigram，助詞と助動詞の bigram，動詞と助動詞の bigram を明らかにするためにランダムフォレストを行った．すなわち、「紫上系」と「玉鬘系」との間におけるもっとも相違する要素を指摘する。

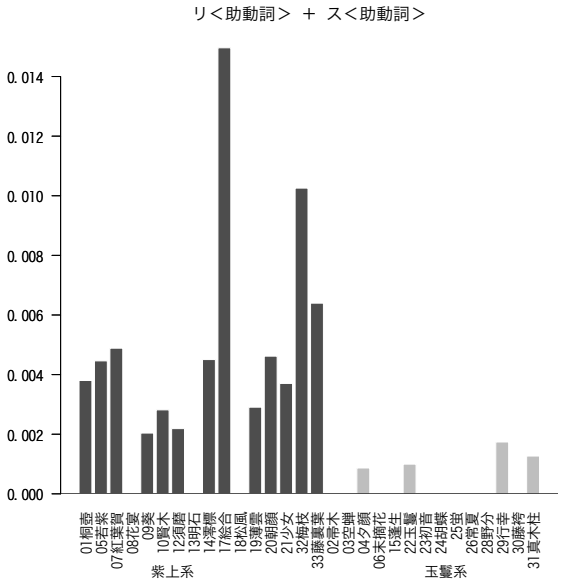

図６　「り ＋ す」の各巻における出現率

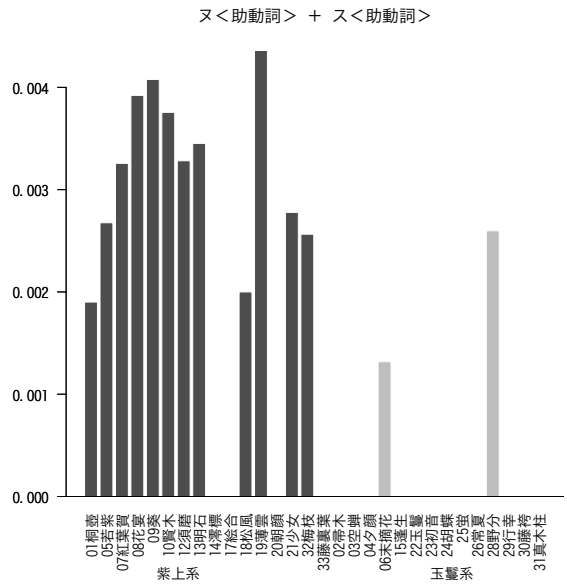

図７　「ぬ ＋ す」の各巻における出現率

ランダムフォレストによって抽出された両系で顕著に出現傾向に相違が認められる助動詞の bigram は図６、図７に示す「り ＋ す」および「ぬ ＋ す」である。助動詞の「り」と「ぬ」はともに完了を意味し、「す」は使役を意味する。すなわち、完了の助動詞と使役の助動詞といった bigram に注目するとき、両系の間に出現傾向の相違が認められ、「り ＋ す」および「ぬ ＋ す」という２つの bigram はどちらも相対的に「紫上系」に特徴的に多用されている。

次に、図８、図９に示すように、「紫上系」と「玉鬘系」との間で顕著に出現傾向に相違が認められる助詞と助動詞の bigram は「なむ ＋ べし」および「こそ ＋ たり」である。どちらの bigram も「玉鬘系」に特徴的に多用されていると言える。助詞の「なむ」および「こそ」はともに係助詞である。つまり、係助詞とそれに対応する助動詞の用法に「紫上系」と「玉鬘系」における文体の相違が現れていると言える。

最後に、動詞と助動詞の bigram に対するランダムフォレストの分析結果は図 10、図 11 に示す通りである。顕著に出現傾向が相違する bigram として「きこゆ ＋ む」および「みゆ ＋ けり」が抽出される。「きこゆ」は「きく」の未然形に上代の受身・可能・自発を意味する助動詞の「ゆ」が結合した語であり、「みゆ」も同様に「みる」に助動詞の「ゆ」が結合した語である。「きこゆ ＋ む」および「みゆ ＋ けり」はともに相対的に「玉鬘系」に多用されている。

以上がランダムフォレストによって指摘しうる「紫上系」と「玉鬘系」との間で顕著に相違する要素である。

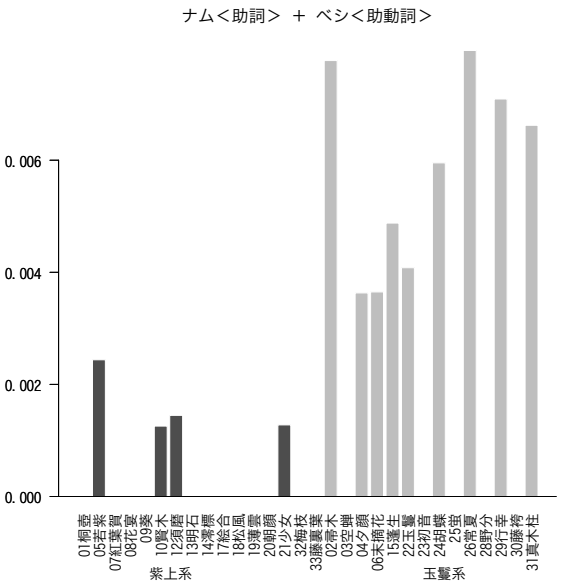

図８　「なむ ＋ べし」の各巻における出現率

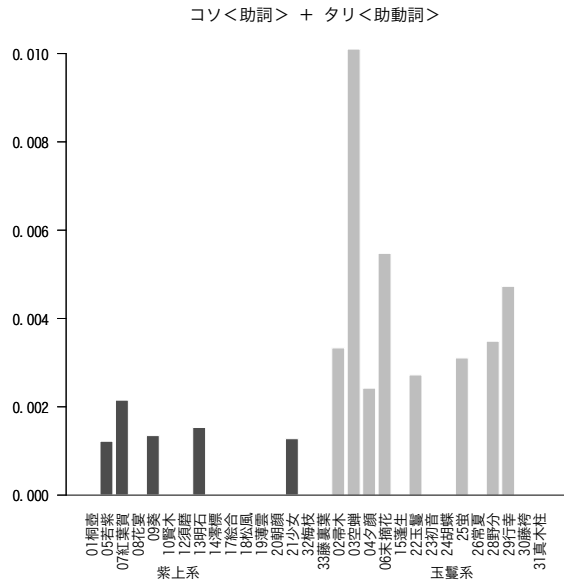

図 9　「こそ ＋ たり」の各巻における出現率

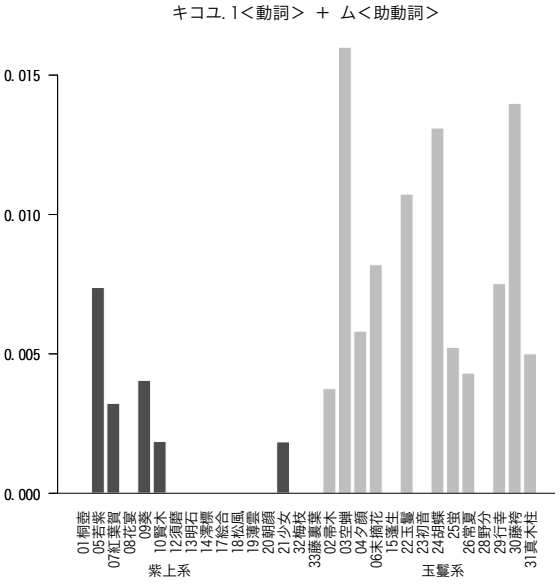

図 10　「きこゆ ＋ む」の各巻における出現率

4. 考察

以上の分析から，「紫上系」と「玉鬘系」との間には，著者が相違するとは考えられないが，文体的相違が認められ得ると言える．とくに助動詞の bigram，「助詞＋助動詞」の bigram，「動詞＋助動詞」の bigram に顕著に相違があらわれる．他方，「紫上系」と「第二部」，「玉鬘系」と「第二部」，「紫上系」と「第三部」，「玉鬘系」と「第三部」，「第二部」と「第三部」との間には本研究において分析した bigram の出現傾向に顕著な相違は認められなかった．

[5]によれば，「紫上系」は本紀の形成を取り入れ，一方「玉鬘系」は列伝の形式を取り入れているとされるように，「紫上系」と「玉鬘系」では作者の執筆に際する姿勢が異なることが指摘されている．このような相違が上述の bigram の出現傾向にあらわれていると考えられる．

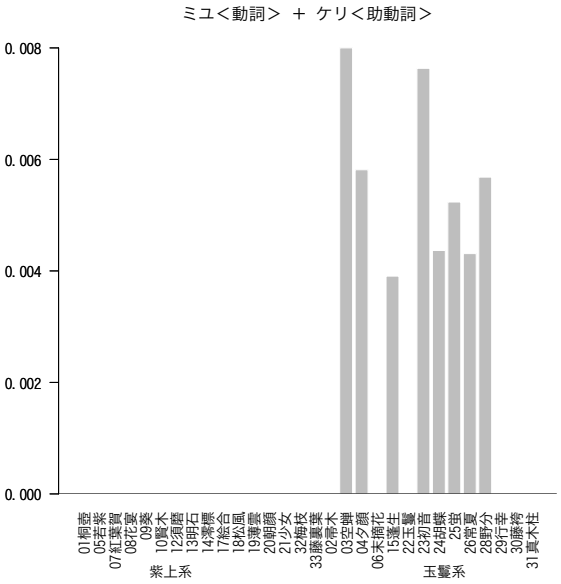

図 11　「みゆ ＋ けり」の各巻における出現率

参考文献

[1] 池田亀鑑. 1951. 源氏物語の構成. 至文堂（『新講源氏物語（上）』所収）.
[2] 和辻哲郎. 1992. 日本精神史研究. 岩波書店.
[3] 青柳秋生. 1939. 源氏物語執筆の順序—若紫の巻前後の諸帖に就いて—. 国語と国文学, 16(8-9)
[4] 武田宗俊. 1952. 源氏物語の最初の形態. 文学, 18(6-7).
[5] 大野晋. 1984. 源氏物語. 岩波書店.
[6] 上田英代・今西祐一郎・藤田真理・村上征勝・樺島忠夫・上田裕一. 1994. 源氏物語語彙総索引. 勉誠社.
[7] 池田亀鑑. 1984. 源氏物語大成. 中央公論社.
[8] Breiman, L. 2001. Random forests. Machine Learning, 24, 123-140
[9] 金明哲・村上征勝. 2007. ランダムフォレスト法による文章の書き手の同定. 統計数理, 55(2), 255-268.
[10] 金明哲. 2002. 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, 11(2), 15-23.
[11] 金明哲. 2002. 助詞の n-gram モデルに基づいた書き手の識別. 計量国語学, 23(5), 225-140.