中古和文における個人文体とジャンル文体

小林 雄一郎（日本学術振興会）†
小木曽 智信（国立国語研究所言語資源研究系）

Styles and Genres in Early Middle Japanese

Yuichiro Kobayashi (Japan Society for the Promotion of Science)
Toshinobu Ogiso (National Institute for Japanese Language and Linguistics)

1．はじめに

現在，国立国語研究所では，日本語歴史コーパスが構築中である (近藤 2012)。そして，2012 年 12 月には，平安時代の仮名文学作品 10 作品の短単位解析済みデータが先行公開された。また，言語資源の整備にともない，中古和文の語彙や文体に関する研究も進められている（須永 2011，小木曽 2012）。

本研究の目的は，中古和文コーパスを分析対象とし，個人文体とジャンル文体の関係を明らかにすることである。安本 (1982) は，現代日本の作家の文章における 15 の文体項目を対象に，因子分析を用いて，個人文体とジャンル文体の類型化を行っている。しかしながら，中古和文を対象とする場合，個人文体とジャンル文体の区別はそれほど簡単にはできない。なぜならば，歴史的な資料は現代語の資料と比べて圧倒的に数が限られているからである。その結果，あるテクストと別のテクストの間に見られる言語的差異が，書き手の差によるものなのか，ジャンルの差によるものなのか，はたまた年代の差によるものなのか，を見分けることが難しくなる。そこで本研究では，紫式部の『源氏物語』と『紫式部日記』，そして『更級日記』における助詞・助動詞の使用傾向を調査し，多変量解析などの統計的手法を用いて，書き手による文体差とジャンルによる文体差の関係について検討する。

2．中古文学の計量文体研究

中古文学の文体研究の多くは，この時代を代表する文学作品である『源氏物語』を対象にしている。また，計量文献学の分野においても，『源氏物語』の作者の推定に関わる研究がなされてきた。たとえば，安本 (1958) は，『源氏物語』を宇治十帖 10 巻とそれ以外の 44 巻に分けて統計的検定を行った結果，両者の作者が同一人物であるとは言い難いと結論付けている。これに対して，新井 (1997) は，五十音図の頭子音行列と母音列別の頻度データに基づいて，宇治十帖の作者は他の諸巻の作者と別人であるとは考えられないとする。同様に，土山・村上 (2012) も，名詞，動詞，形容詞，形容動詞，副詞，助詞のそれぞれを変数とする主成分分析とランダムフォレストを行い，宇治十帖他作者説を退けている。『源氏物語』の成立論に関して，村上・今西 (1999) は，高頻度の助動詞を変数とする数量化 III 類を行い，(1) 第 1 部の紫の上系物語，(2) 第 2 部全てと第 3 部の宇治三帖，(3) 第 1 部の玉鬘系物語，(4) 第 3 部の宇治十帖の順で執筆されたという仮説を提唱している。さらに，他の文学作品との比較に関して，土山・村上 (2011) は，名詞，動詞，形容詞，形容動詞，副詞，助詞，助動詞のそれぞれを変数とする主成分分析とクラスター分析を行い，『源氏物語』の使用語彙と『宇津保物語』の使用語彙の間には顕著な差が見られると報告している。

また，物語文学と日記文学を計量的に比較した研究として，坂東 (1990) が挙げられる。この論文では，『枕草子』と『紫式部日記』における名詞率，MVR（動詞数に対する形容詞，形容動詞，副詞，

† kobayashi0721@gmail.com

連体詞の総和数の割合），形容詞，色彩語についての比較が行われている。ただ，その目的は「それぞれの作品の個別的文体」を明らかにすることであり，個人文体とジャンル文体の関係に光を当てるものではない。

3. 調査方法

3.1 資料

本研究で調査対象とする資料は，新編日本古典文学全集の『源氏物語』と『紫式部日記』である。『源氏物語』に関しては，第1部の「桐壺」と「若紫」（いずれも紫の上系物語）と第3部の「橋姫」と「夢浮橋」（宇治十帖の最初の巻と最後の巻）を対象とする。

これらの紫式部による作品に加えて，『更級日記』も調査対象に含める（このデータは，西端ほか 1996 に基づいている）。菅原孝標女による『更級日記』を含めたのは，個人文体とジャンル文体の関係，言い換えれば，書き手による言語的特徴の違いとジャンルによる言語的特徴の違いの関係を明らかにするために，紫式部以外の手によるテクストが必要であるからである。なお，菅原孝標女は『源氏物語』を愛読していたとされ，『更級日記』の文体も『源氏物語』の強い影響を受けていると言われている（上野 1991，上野 1994）。

また，これらの資料に対して自動形態素解析を行い，解析誤りを手作業で修正した。データに付与されている単語情報は，形態素解析辞書中古和文 UniDic（小木曽ほか 2010）で採用されている短単位に基づくものである。

3.2 変数

本研究では，各テクストにおける助詞と助動詞の語彙素の頻度を変数とする。これらの変数を選んだ理由は，日本語が膠着語であり，助詞や助動詞が表現の論理や情緒を表すにあたって重要な働きを持っているからである（此島 1971）。なお，個々のテクストの総語数が異なるため，分析にあたっては，必要に応じて，観測頻度を出現率に変換した相対頻度を用いる。

3.3 手法

各テクストにおける助詞と助動詞の頻度を分析するにあたって，最初にカテゴリー別の頻度をバープロットで視覚化する。次に，個々の助詞や助動詞の観測頻度に基づく相関係数を算出し，テクストの類似度を求める。さらに，テクストと変数の関係を把握するために，多重因子分析とクラスター分析による次元縮約を行う。そして最後に，対数尤度比を用いて，個々のテクストに特徴的な助詞と助動詞を抽出する。

4. 結果と考察

4.1 変数の分布

図1は，中古和文 UniDic による形態素解析の結果に基づき，格助詞，係助詞，終助詞，副助詞，接続助詞，助動詞の頻度の相対頻度を視覚化したものである。この図を見ると，他のテクストと比べて，『更級日記』における格助詞の頻度が高い。そして，物語文学と比べて，日記文学における助動詞の頻度が低い。これらについては，後段で詳しく見る。

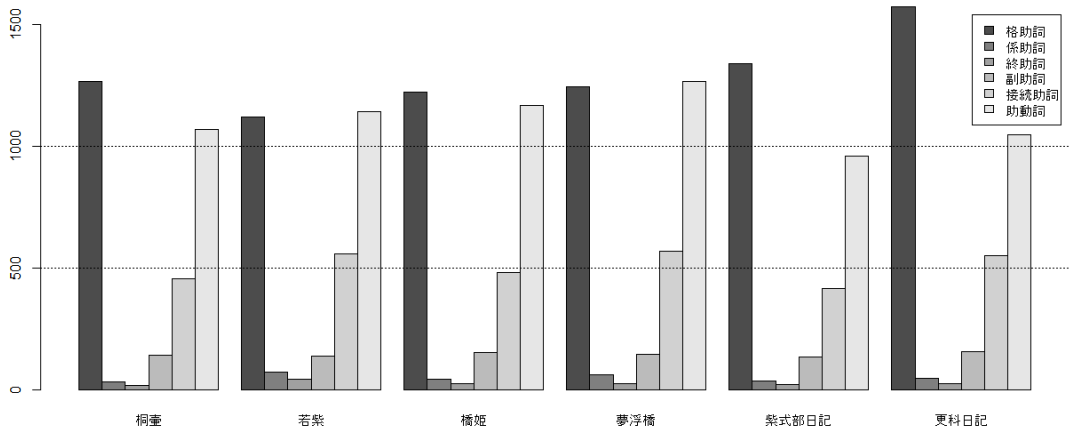

図1　各テクストにおける助詞・助動詞の相対頻度

4.2　テクストの相関関係

図2は，個々の助詞と助動詞の観測頻度に基づき，各テクスト間の相関係数（Pearsonの積率相関係数）を求めた結果である。これを見ると，最も高い値は「桐壺」と「橋姫」の0.988，最も低い値でも「夢浮橋」と『紫式部日記』の0.893であり，全体的に非常に高い値となっている。本研究の調査対象がいずれも11世紀のテクストであることから，大まかな助詞と助動詞の使い方は極めて類似したものになっていると思われる。

4.3　テクストと変数のクラスタリング

前節までの分析手法では，助詞と助動詞の使用傾向に関して，各テクスト間における顕著な差異は認められなかった。しかしながら，これまでの多くの研究において，『源氏物語』と『紫式部日記』，あるいは『源氏物語』における宇治十帖と他の巻との文体差が指摘されてきた（上野 1990, 野村 1970a, 野村 1970b, 安本 1958）。本節では，多重因子分析とクラスター分析を用いて，高次元のデータを低次元に圧縮し，その結果を直感的に解釈しやすい形式での視覚化を試みる。

以下は，個々のテクストをケースとし，助詞と助動詞の頻度を変数とした多重因子分析の結果である。多重因子分析は（因子分析ではなく）主成分分析の一種であり，変数をグループ（群）に分けて指定することができる (Wong *et al.* 2002, Pagès 2004)。この分析では，格助詞，係助詞，終助詞，副助詞，接続助詞，助動詞という6つのカテゴリーを変数のグループとして指定した。まず，図3は大局的ケース図であり，partial pointsとの結線を表示したものである。これを見ると，第2象限に『紫式部日記』と『更級日記』という日記文学がプロットされている。次に，図4は大局的負荷図であり，格助詞の多くが左上（第2象限）を向いていることは注目に値する。また，図5は群表示であり，終助詞以外の5つのカテゴリーが第1主成分に同等に寄与しており，第2主成分には助動詞が最も寄与している。そして，図6が群の大局への寄与であり，接続助詞と係助詞の第1主成分が大局分析の第1主成分に高く相関し，それ以外の4つのカテゴリーの第1主成分が大局分析の第2主成分に高く相関していることが分かる。

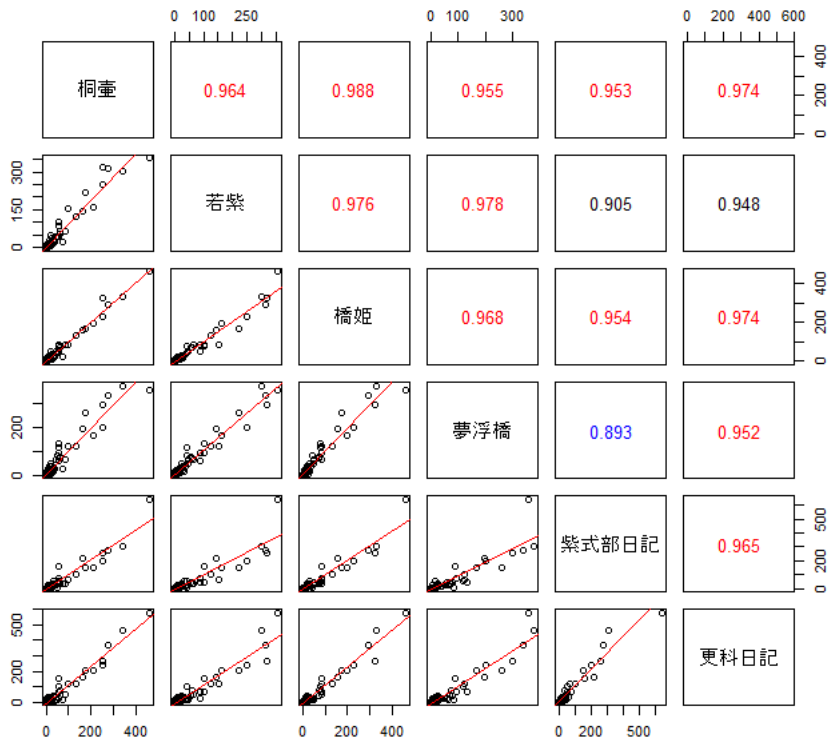

図2　各テクストの相関関係

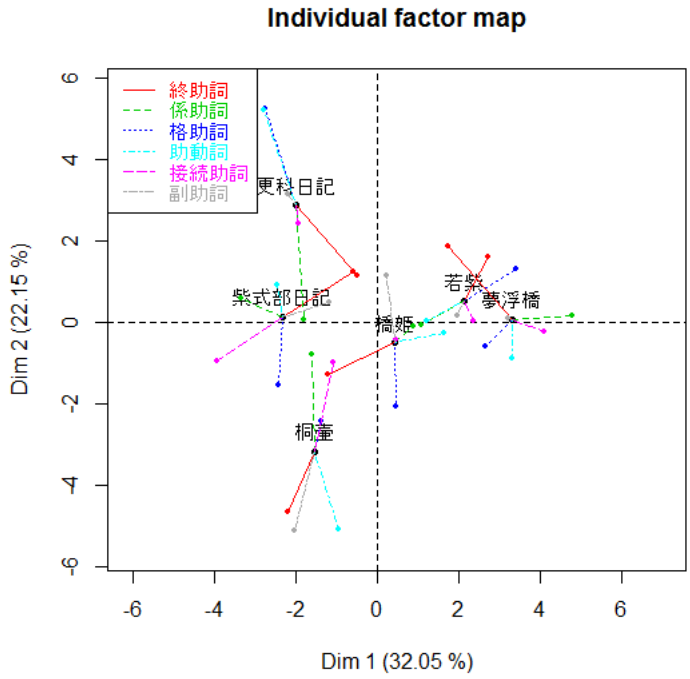

図3　大局的ケース図

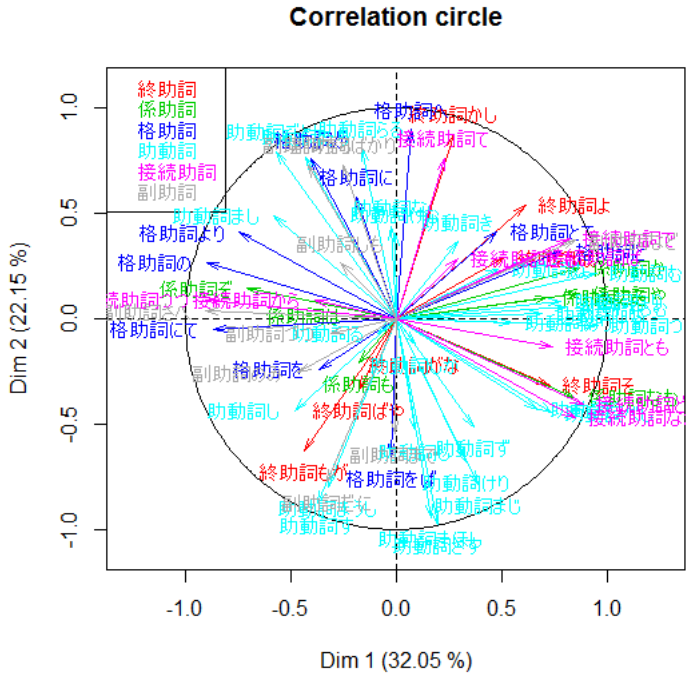

図4　大局的負荷図

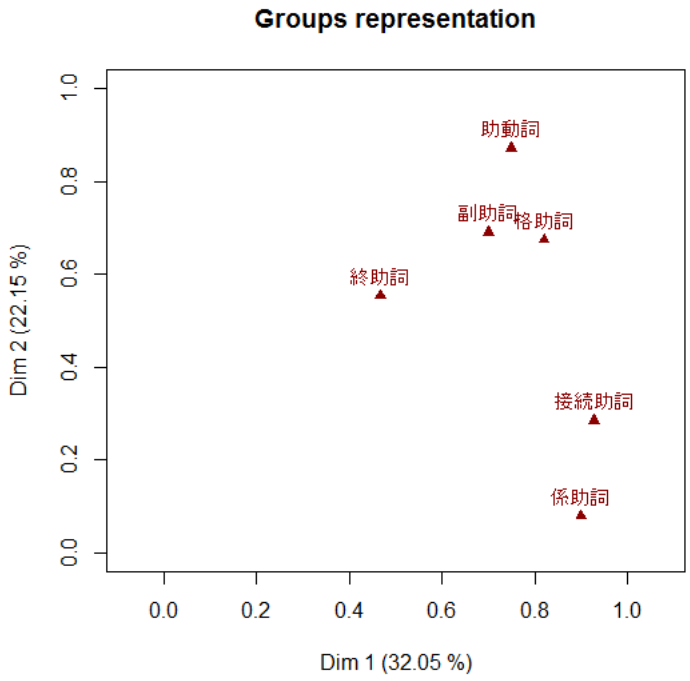

図5　群表示

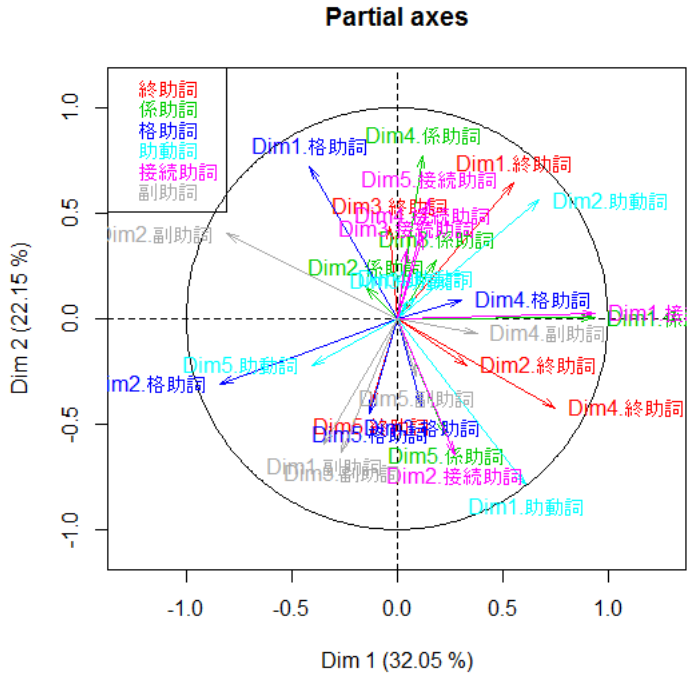

図6　群の大局への寄与

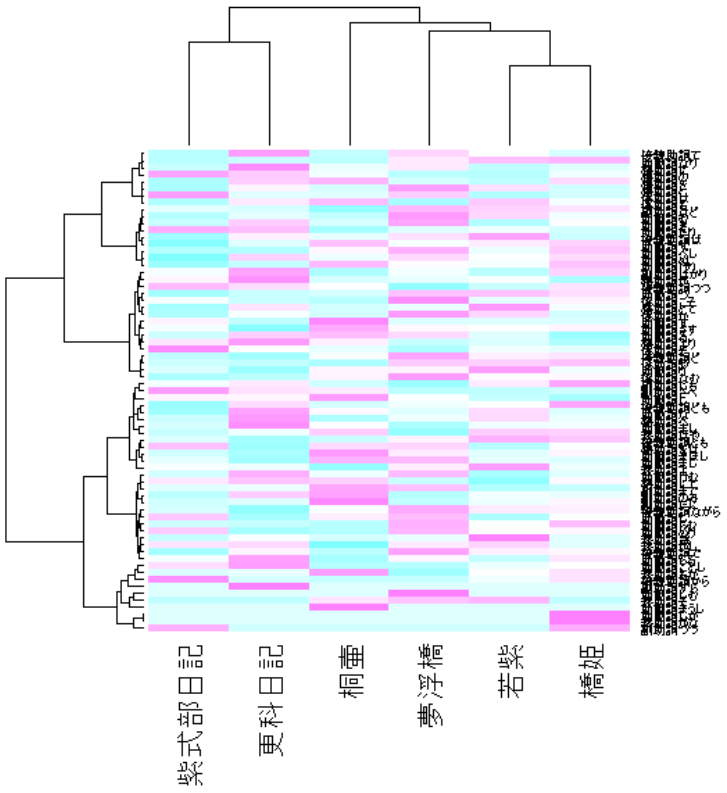

図7　ケースと変数のクラスター分析とヒートマップ

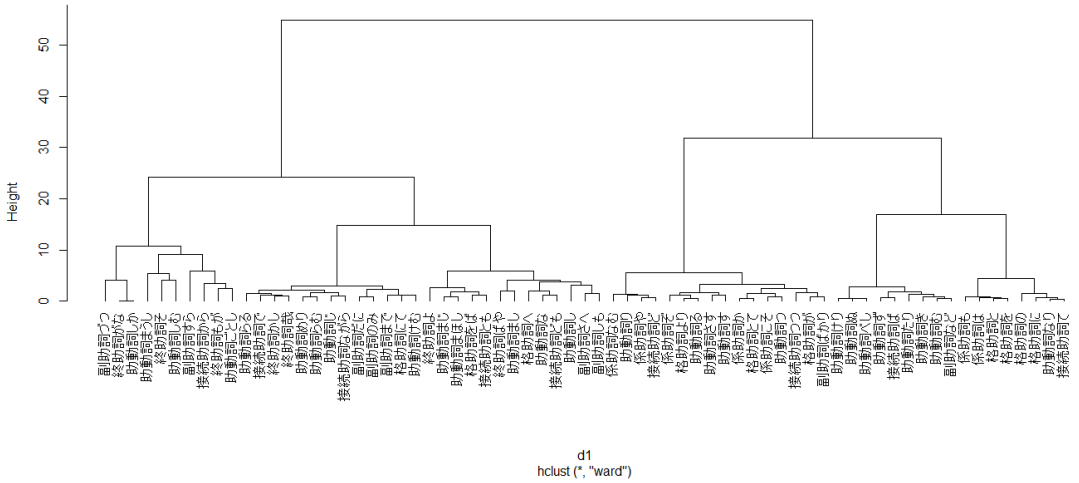

図8　変数のクラスター分析（拡大）

図7は，ケース（テクスト）と変数（助詞・助動詞）の双方に対して階層型クラスター分析を行い，ヒートマップとともに表示した結果である。分析にあたって，ケース間の距離の計算には，値が小さく差が少ないデータ同士に対しても非常に感度が高いとされているキャンベラ距離 (Gorden 1999) を用いた。また，クラスター間の距離の計算には，クラスターの各値からその質量中心までの距離を最小化するため，他の距離関数に比べて分類感度が高いとされているウォード法 (Anderberg 1984) を用いた。[1] そして，ヒートマップとは，クラスター分析に使われている元データ（ここでは，助詞・助動詞の頻度行列）に含まれる値の大小を色で表す視覚化手法である (Chaussabel 2004)。

図7におけるケースクラスタリングの結果を見ると，左側に日記文学（『紫式部日記』，『更級日記』），右側に物語文学（「桐壺」，「夢浮橋」，「若紫」，「橋姫」）がクラスターを形成している。また，『源氏物語』における紫の上系物語（「桐壺」，「若紫」）と宇治十帖（「橋姫」，「夢浮橋」）の差異は認められない。この結果をまとめると，少なくとも本研究で調査対象としたテクストにおける助詞と助動詞の頻度を変数とした分析では，書き手による文体差（個人文体）よりもジャンルによる文体差（ジャンル文体）の方が大きいことを示している。上野 (1990) は，「『源氏物語』と『紫式部日記』の文章が類似し，同一の傾向を潜在的に共有しているにせよ」，「日記と物語とは，やはり別種の作品」であるとする。本研究の結果は，助詞と助動詞の使用傾向から，この説を計量的に裏付けるものと言えるだろう。

なお，図8は，図7における変数クラスタリング結果を拡大し，右に90度回転させて表示したものである。図8を見ると，大きく2つのクラスターが形成されており，左側のクラスターには副助詞と終助詞が，右側のクラスターには格助詞と係助詞が多く含まれている。

4.4　各テクストに特徴的な変数

本節では，対数尤度比 (log-likelihood ratio, LLR) (Dunning 1993) を用いて，各テクストに特徴的な変数を抽出する。この抽出指標は，『古典対照語い表』を用いた古典作品別の特徴語抽出 (宮島・近藤 2011)

[1] ウォード法にはユークリッド距離を用いるのが基本であるが (Romesburg 1973)，計量文献学の分野ではキャンベラ距離とウォード法の組み合わせを用いることもある (石田 2008, 金 2009)。

にも利用されているものである。なお，あるテクストに特徴的な変数を抽出するにあたっては，それ以外の 5 つのテクストを比較対象とする。

表 1 は，対数尤度比を用いて，各テクストに特徴的な助詞，助動詞を抽出した結果である。この表を見ると，「桐壺」を最も特徴付ける助動詞として，敬語の「す」と「さす」が抽出されていることである。これらの助動詞は，他のテクストと比べて，「桐壺」に高い頻度で生起している（図 9）。

表 1　各テクストに特徴的な助詞・助動詞

桐壷		若紫		橋姫		夢浮橋		紫式部日記		更科日記	
語	LLR	語	LLR	語	LLR	語	LLR	語	LLR	語	LLR
す	30.183	り	42.378	なむ	25.314	なむ	27.367	の	134.304	に	42.650
さす	9.548	ば	37.110	む	12.629	む	14.841	たり	44.588	き	16.289
けり	9.104	む	16.905	き	9.274	き	10.283	は	36.804	まし	11.119
なむ	6.522	なり	9.333	ど	8.368	と	8.466	ぞ	34.139	けむ	10.481
まで	6.101	と	9.208	と	6.415	ど	8.232			て	9.714
のみ	6.059	なむ	9.039	こそ	4.974	こそ	5.404			ばかり	6.732
を	5.496	哉	7.732	べし	4.107	しむ	5.279			が	6.632
だに	5.377	とて	7.493	らむ	3.381	か	4.218			らる	6.613
まうし	4.846	つ	7.012	か	3.224	べし	3.447			たり	6.036
		や	6.254							へ	5.712
		ど	5.368							より	5.435

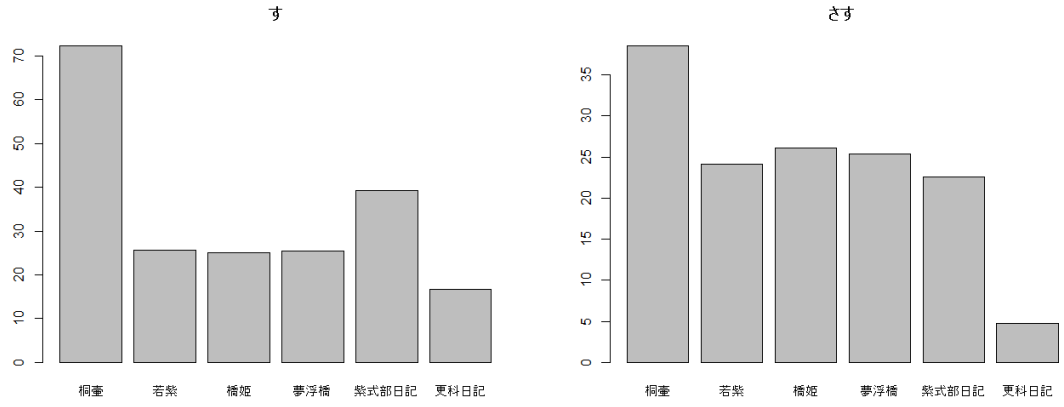

図 9　助動詞「す」，「さす」の分布

また，宇治十帖の「橋姫」と「夢浮橋」に特徴的な助詞と助動詞が極めて類似している。そして，『更級日記』から「に」，「が」，「へ」，「より」という格助詞が抽出されている。『更級日記』に格助詞が多く生起することは前述のとおりである（図 1）。図 10 は，格助詞「に」，「が」，「へ」，「より」の頻度分布を視覚化したものである。

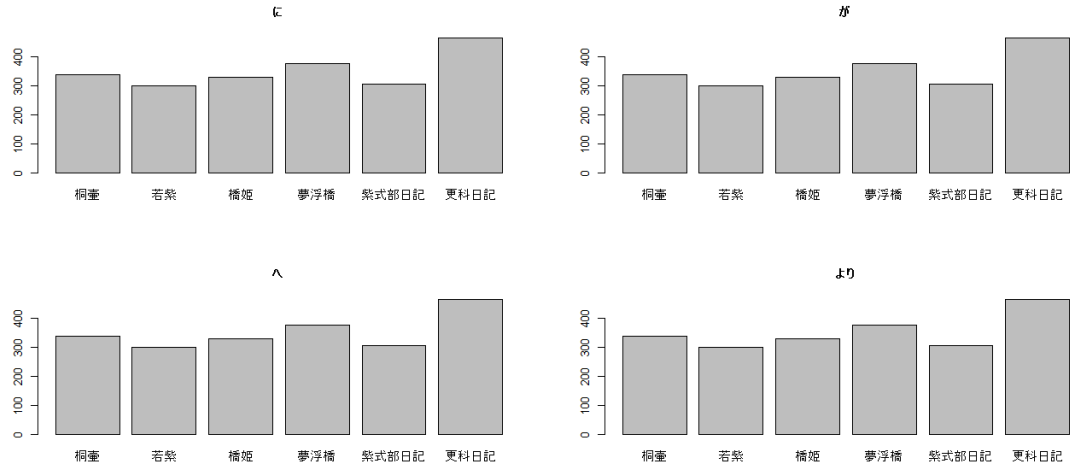

図 10　格助詞「に」,「が」,「へ」,「より」の分布

5. おわりに

本研究では，紫式部の『源氏物語』と『紫式部日記』，そして『更級日記』における助詞・助動詞の使用傾向を調査し，多変量解析などの統計的手法を用いて，書き手による文体差（個人文体）とジャンルによる文体差（ジャンル文体）の関係について検討してきた。その結果，少なくとも今回分析したデータにおいては，個人文体よりもジャンル文体の方が大きいことが明らかにされた。

今後の課題としては，まず，同時代の他のテクストや他の言語項目の分析を積み重ねていかなければならない。また，助詞と助動詞を変数とする場合でも，「なら-じ」(2 語連結）や「なら-ぬ-を」(3 語連結）のような「助詞・助動詞相互の連結関係」(宇都宮 1966) を扱うことも考えられる。さらに，テクスト全体を 1 つのケースとするだけでなく，「会話」,「歌」,「手紙」,「地の文」といった本文種別情報 (小木曽 2012) を活用し，より詳細な文体分析を行う必要がある。

文　献

新井皓士 (1997).「源氏物語・宇治十帖の作者問題：一つの計量言語学的アプローチ」『一橋論叢』117(3), pp. 397-413.

石田基広 (2008).『R によるテキストマイニング入門』森北出版.

上野英二 (1990).「紫式部における日記と物語」『成城國文學論集』20, pp. 11-50.

上野英二 (1991).「更級日記と文学史」『成城國文學論集』21, pp. 1-36.

上野英二 (1994).「菅原孝標女と源氏物語」『成城國文學論集』22, pp. 1-27.

宇都宮睦男 (1966).「紫式部日記の文体―助動詞・助詞の連結から見た」『国語教育研究』11, pp.65-71.

小木曽智信 (2012).「中古和文における語彙の文体差」『NINJAL「通時コーパス」プロジェクト・Oxford VSARPJ プロジェクト合同シンポジウム「通時コーパスと日本語史研究」予稿集』国立国語研究所, pp.41-50.

小木曽智信・小椋秀樹・田中牧郎・近藤明日子・伝康晴 (2010).「中古和文を対象とした形態素解析辞書の開発」『情報処理学会研究報告』2010-CH-85(4), pp. 1-8.

金明哲 (2009).『テキストデータの統計科学入門』岩波書店.

此島正年 (1971).「源氏物語の助詞」山岸徳平・岡一男（編）『源氏物語講座 第 7 巻 表現・文体・

語法』 有精堂出版, pp. 266-293.
近藤泰弘 (2012). 「日本語通時コーパスの設計について」 『国語研プロジェクトレビュー』3(2), pp. 84-92.
須永哲矢 (2011).「コロケーション強度を用いた中古語の語認定」『国立国語研究所論集』2, pp. 91-106.
土山玄・村上征勝 (2011). 「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シンポジウム論文集—「デジタル・アーカイブ」再考』 情報処理学会, pp. 125-132.
土山玄・村上征勝 (2012). 「語の使用頻度の計量分析による宇治十帖他作者説の検討」 『情報処理学会研究報告』2012-CH-94(5), pp. 1-8.
西端幸雄・木村雅則・志甫由紀恵 (1996). 『平安日記文学総合語彙索引：土佐日記・蜻蛉日記・和泉式部日記・紫式部日記・更級日記』 勉誠社.
野村精一 (1970a). 「宇治十帖の言語と文体」 『源氏物語文体論序説』 有精堂出版, pp. 228-262.
野村精一 (1970b). 「紫式部の文体—紫式部日記について」 『源氏物語文体論序説』 有精堂出版, pp. 263-305.
坂東久美 (1990). 「『枕草子』と『紫式部日記』における文体の比較研究」 『徳島大学国語国文學』 3, pp. 64-69.
宮島達夫・近藤明日子 (2011). 「古典作品の特徴語」 『計量国語学』28(3), pp. 94-105.
村上征勝・今西祐一郎 (1999). 「源氏物語の助動詞の計量分析」 『情報処理学会論文誌』40(3), pp. 774-782.
安本美典 (1958). 「宇治十帖の作者—文章心理学による作者推定」 『心理学評論』2, pp. 147-156.
安本美典 (1982). 「文章様式論」 宮地裕・樺島忠夫・安本美典 (編) 『講座日本語学 8 文体史 II』 明治書院, pp. 1-22.
Anderberg, M. R. (1973). *Cluster analysis for applications*. New York: Academic Press.
Chaussabel, D. (2004). Biomedical literature mining: Challenges and solutions in the 'omics' era. *American Journal of Pharmacogenomics,* 4(6), pp. 383-393.
Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics,* 19(1), pp. 61-74.
Gordon, A. D. (1999). *Classification.* 2nd ed. Boca Raton: Chapman and Hall.
Pagès, J. (2004). Multiple factor analysis: Main features and application to sensory data. *Revista Colombiana de Estadística*, 27(1), pp. 1-26.
Romesburg, H. C. (1984). *Cluster analysis for researchers*. Belmont: Lifetime Learning Publications.
Wong, S., Gauvrit, H., Cheaib, N., Carré, F., & Carrault, G. (2002). Multiple factor analysis as a tool for studying the effect of physical training on the autonomic nervous system. *Computers in Cardiology,* 29, pp. 437-440.

関連 URL

統計と機械学習による日本語史研究

`http://www.ninjal.ac.jp/research/project/c/statisticsja/`

日本語歴史コーパス

`http://www.ninjal.ac.jp/corpus_center/chj/`

中古和文 UniDic

`http://www2.ninjal.ac.jp/lrc/index.php?UniDic`