中古和文における個人文体とジャンル文体 ： 多変量解析による歴史的資料の文体研究

著者	小林 雄一郎, 小木曽 智信
雑誌名	国立国語研究所論集
号	6
ページ	29-43
発行年	2013-11
URL	http://doi.org/10.15084/00000510

国立国語研究所論集 (NINJAL Research Papers) 6: 29–43 (2013)
ISSN: 2186-134X print/2186-1358 online

中古和文における個人文体とジャンル文体
——多変量解析による歴史的資料の文体研究——

小林雄一郎[a]　　小木曽智信[b]

[a]日本学術振興会 特別研究員PD／国立国語研究所 共同研究員［–2013.10］
[b]国立国語研究所 言語資源研究系

要旨

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。具体的には，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，テクスト間の相互関係，言語項目間の相互関係，テクストと言語項目の結びつきのパターンを定量的に分析する。そして，多変量解析の手法を援用し，中古和文のテクストにおいて，書き手による文体差よりもジャンルによる文体差の影響が大きいことを示す。さらに，個々のテクストにおける語彙使用を詳細に分析するために，対数尤度比による特徴語抽出を行い，多変量解析の結果を補完する *。

キーワード：文体，中古和文，多変量解析，特徴語抽出

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を構築中である（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（『日本語歴史コーパス平安時代編』）が先行公開された（小木曽ほか 2013）[1]。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，冨士池 2013）。

しかしながら，中古和文のような歴史的資料を対象とする文体研究には，大きな問題がある。それは，現代語の資料と比べて圧倒的に数が限られていることである。そして，分析対象であるテクストが極めて少ないために，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い。

文体論の分野では，古くから「文体」とは何かという議論がなされてきた。そして，少なくとも言語学的な立場からの文体研究では，文体が「個人文体」と「ジャンル文体」との 2 つに大別されることは従来広く認められてきた（陳 2012）。ここで言う個人文体とは，「森鷗外の文体」や「川端康成の文体」というものを指し，ジャンル文体とは，「新聞の文体」や「公用文の文体」

* 本稿は，国立国語研究所の萌芽・発掘型共同研究プロジェクト「統計と機械学習による日本語史研究」（プロジェクトリーダー：小木曽智信，2010 年 11 月～ 2013 年 10 月）の研究成果であり，2013 年 2 月 28 日～ 3 月 1 日に国立国語研究所で開催された「第 3 回コーパス日本語学ワークショップ」における発表内容に大幅な加筆・修正をしたものである。また，本稿をまとめるにあたって，有益なコメントをくださった前川喜久雄教授（国立国語研究所）に感謝いたします。
[1] http://www.ninjal.ac.jp/corpus_center/chj/

といったものを指す。このような立場から，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，それぞれの文体を類型化した。ただし，安本による個人文体とジャンル文体の分析はそれぞれ独立したものであり，それら 2 つの文体がどのように関係しているのかという点については，深く述べられていない。従って，このような状況において，個人文体とジャンル文体の関係の調査，さらに，歴史的な資料における文体を研究するための方法論の確立は急務である。

2. 中古和文の計量文体研究

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。そして，『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，『源氏物語』と他の文学作品の比較に関して，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定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，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. 研究方法

3.1 研究目的

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。物語文学と日記文学における助詞・助動詞の使用傾向を調査し，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。それと同時に，多変量解析などの統計的手法を援用し，歴史的な資料における文体を定量的に分析するための方法論を模索する。

3.2 調査資料

中古和文において，複数のジャンルのテクストを残している書き手は，『源氏物語』や『紫式部日記』を書いた紫式部，『古今和歌集仮名序』や『土佐日記』を書いた紀貫之，『俊頼髄脳』や『金葉和歌集』を書いた源俊頼など，非常に限られている[2]。そして，現存するテクストの分量やジャンル，関連研究の量などを考慮した場合，まずは，紫式部のテクストを中心に分析を進めるのが妥当であろう。

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[3]。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（表 1）。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）[4] で採用されている短単位に基づくものである。

表 1　形態素解析結果（一部）

作品名	表記	読み	品詞	語彙素	読み	本文種別
源氏物語 - 桐壺	何れ	イズレ	代名詞	いづれ	イズレ	和
源氏物語 - 桐壺	の	ノ	助詞-格助詞	の	ノ	和
源氏物語 - 桐壺	御	オオン	接頭辞	御	オオン	和
源氏物語 - 桐壺	時	トキ	名詞-普通名詞-副詞可能	時	トキ	和
源氏物語 - 桐壺	なり	ナリ	助動詞	に	ニ	和
源氏物語 - 桐壺	か	カ	助詞-係助詞	か	カ	和
源氏物語 - 桐壺	,		補助記号-読点	,		記号

表 2 は，本稿で調査対象とするテクストの総語数をまとめたものである。

表 2　テクストの総語数

桐壺	若紫	橋姫	夢浮橋	紫式部日記	更級日記
6500	12861	9959	4731	20346	16652

[2]『金葉和歌集』には 650 首強の和歌が収められているが，源俊頼によるものは 35 首とされている。
[3] 紫式部以外の日記文学を加えるのではなく，紫式部以外の物語文学を加えるという選択肢もあるが，物語文学の方が日記文学よりもジャンル内での文体の差異が大きいと予想される。
[4] http://www2.ninjal.ac.jp/lrc/index.php?UniDic

3.3 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）。特に，助詞の使い方には，書き手の文体的特徴が顕著に現れるとされる（金 2002）。また，歴史的資料を対象とする著者推定の研究においても，助詞は極めて有効な文体指標となる（上阪・村上 2013）。

3.4 分析手法

各テクストにおける助詞と助動詞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し，各テクストにおける頻度分布を概観する。次に，多重因子分析とクラスター分析による次元縮約を行い，テクストと変数の関係を把握する。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。

3.4.1 多重因子分析

言語研究において，大量のテクストや変数を定量的に分析する場合，因子分析，主成分分析，対応分析といった多変量解析の手法がよく用いられる（Burrows 1987, Biber 1988, Nakamura and Sinclair 1995）。また，テクストにおける書き手の差とジャンルの差を分析する本研究に近い問題意識を持つ研究として，井上・太田（1989）による方言の地域差と個人差の分析が挙げられる。これは，岐阜県徳山村の 8 集落から各 4 人ずつ話者を選び，回答語形の一致率に基づく因子分析とクラスター分析を用いて，地域差と個人差の関係を調査したものである。

このような多変量解析の手法は，高次元のデータから有益な情報を抽出し，その結果を直感的に解釈しやすい形式で視覚化できる点において，非常に有効なアプローチである。しかしながら，あまりにも多くの変数を投入した場合は，分析の結果として得られる散布図の視認性が著しく下がり，結果の解釈が難しくなる。また，言語データ，特に単語の頻度などを扱う場合，全ての変数を同列に扱うのではなく，「名詞」や「動詞」といったカテゴリーごとの分析が必要になることもある。だが，コーパスに現れる全ての単語をカテゴリー（品詞）別に集計したものを分析に用いると，カテゴリー間の差が明確になる一方で，カテゴリー内の頻度分布などの情報が失われてしまう。

上記の問題を踏まえ，本研究では，多重因子分析という手法を用いる。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，変数をカテゴリー（群）に分けて指定することができる（Wong *et al.* 2002, Pagès 2004）。このように変数のカテゴリーを指定することで，ときに「銀河」（Nakamura 2002）などと表現される，大量の単語が布置された図の解釈が容易になる。なお，以下の分析では，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カテゴリーとして指定する。

3.4.2 クラスター分析

クラスター分析は，個々のデータ（書き手，テクストなど）の非類似度を距離として表現し，距離の近いデータ同士をまとめてクラスターを作っていく手法である。そして，言語研究においても，この分析手法は，書き手やテクストの分類などに広く用いられてきた（金・樺島・村上 1993，Hoover 2001，小林 2013）。以下の分析では，データ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（Gordon 1999, 安形・安形 2009）を用いる。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 1973）を用いる[5]。

3.4.3 対数尤度比

本研究では，多重因子分析とクラスター分析を行ったのち，対数尤度比（Dunning 1993）を用いて，各テクストに特徴的な変数（助詞・助動詞）を抽出する。対数尤度比（LLR）は，コーパスから特徴語を抽出するための手法の 1 つであり（Scott and Tribble 2006, Archer 2009），『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）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。

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（10000 語あたり）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。

[5] ウォード法にはユークリッド距離の平方を用いるのが基本であるが (Romesburg 1984)，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（金 2009)。

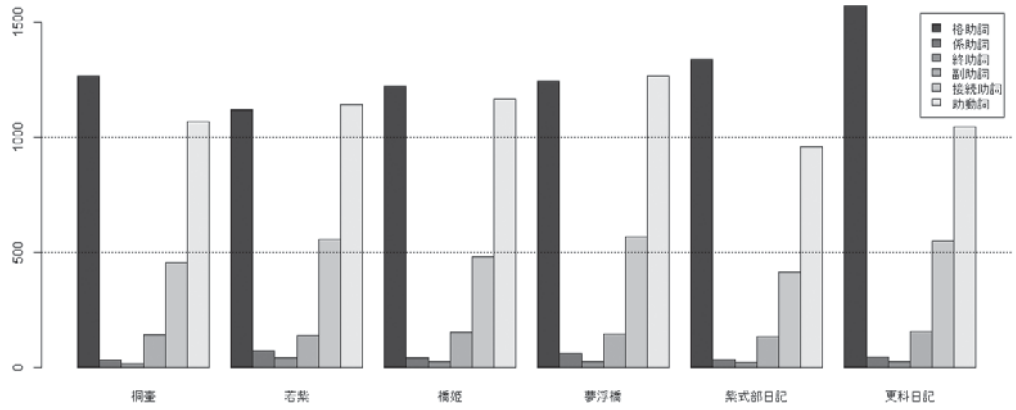

図 1　各テクストにおける助詞・助動詞の相対頻度

4.2 テクストと変数のクラスタリング

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。まず，図 2 は大局的ケース図であり，テクスト間の相互関係を表している。この図においては，近くにプロットされているテクストは類似した性質を持っており，遠くにプロットされているテクストは異なった性質を持っている。これを見ると，『紫式部日記』と『更級日記』という日記文学が第 2 象限（左上）に分布しており，それ以外の物語文学が第 1 象限（右上），第 3 象限（左下），第 4 象限（右下）に分布している。さらに言えば，「桐壺」は，物語文学の中でも若干異なった性質を持っているように見える。

次に，図 3 は大局的負荷図であり，図 2 と対応するものである。従って，格助詞の多くが左上を向いていることから，それらの格助詞が日記文学に顕著であることが読み取れる。

また，図 4 は群表示であり，各カテゴリーと各主成分の関係を表している。この図を見ると，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ことが分かる（ただし，助動詞は，第 2 主成分だけでなく，第 1 主成分にも高く寄与している）。

そして，図 5 が群の大局への寄与を表したものである。この図からは，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。

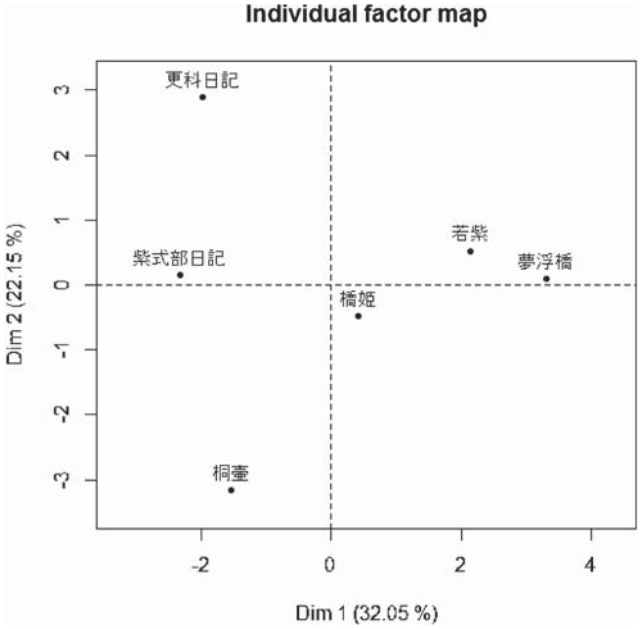

図 2　大局的ケース図

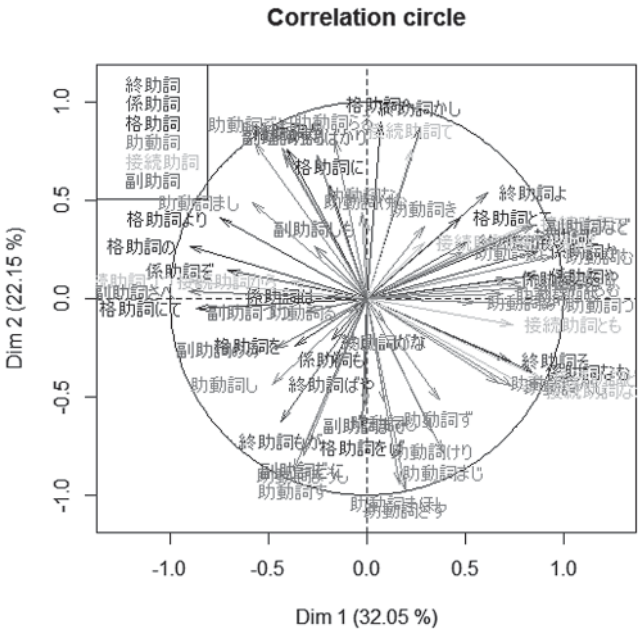

図 3　大局的負荷図

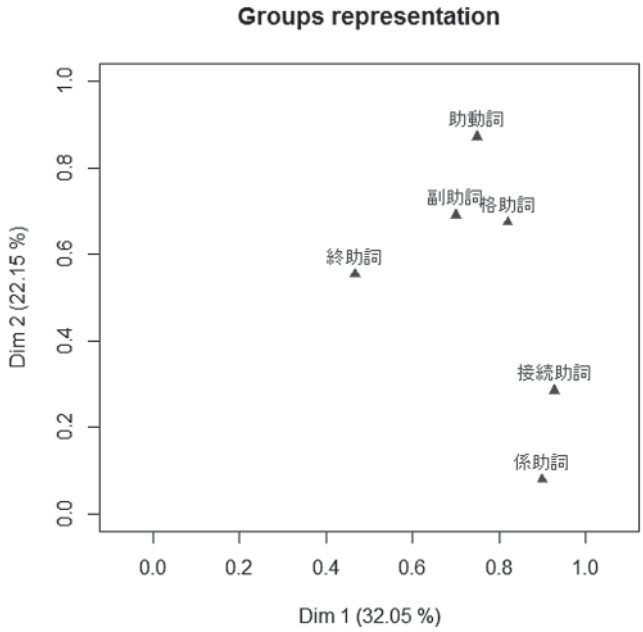

図 4　群表示

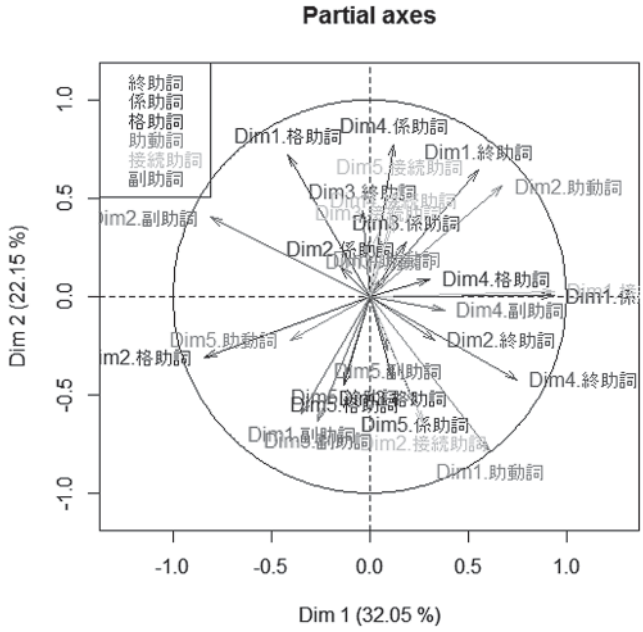

図 5　群の大局への寄与

多重因子分析の結果のうち，特に注目すべき点をまとめると，(1) 助詞と助動詞の使用に関して，日記文学と物語文学の間に一定の差異が見られること，(2) 日記文学には格助詞が顕著で，物語文学には助動詞が顕著であること，(3)「桐壺」は他の物語文学と若干異なった性質を持っていること，などである。このうち，最初の 2 点に関しては，図 1 のバープロットからも読み取ることができる。それに加えて，多重因子分析の結果からは，より詳細なテクスト間の相互関係，変数間の相互関係，カテゴリー間の相互関係，テクストと変数の対応関係，テクストとカテゴリーの対応関係などを読み取ることが可能である。

前述のように，本研究の目的は，中古和文のテクストを定量的に分析し，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。これに関して，図 2 を見ると，日記文学と物語文学がそれぞれ別のクラスターを形成しているように見受けられ，ジャンルの差の方が書き手の差よりも大きいように思われる。だが，第 1 主成分のみに注目した場合は，『紫式部日記』，『更級日記』，「桐壺」が 1 つのクラスターを形成し，残りの 3 つのテクストが別のクラスターを形成しているようにも見える。「桐壺」が他の物語よりも日記に近い特徴的な位置を占めている理由の 1 つは，「桐壺」が『源氏物語』の冒頭の巻であることである。物語の冒頭と日記とは，前段の内容を受けて続きが展開されるものではなく新情報の説明的記述が多くなるという点で，共通性が認められる。以下，分析者による解釈の恣意性を少しでも軽減するために，クラスター分析を用いて，テクスト間の関係をより詳しく検討することにする。

図 6 は，ケースに対して，クラスター分析（キャンベラ距離，ウォード法）を行った結果である。この図を見ると，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。つまり，この結果は，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[6]。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。

[6] 前述のように，宇治十帖に関しては，古くから紫式部以外の作者が書いたものであるという「宇治十帖他作者説」がある。ただ，仮に「橋姫」と「夢浮橋」が紫式部以外の作者によって書かれたものであったとしても，書き手による文体差よりもジャンルによる文体差の方が大きいという本研究の分析結果は変わらない。

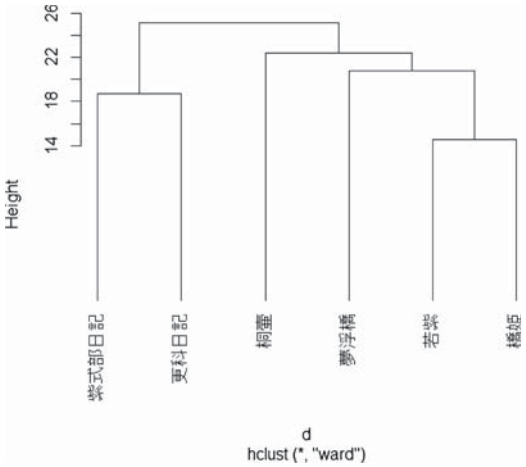

図 6　ケースのクラスター分析

4.3 各テクストに特徴的な変数

前節では，計量文献学的な立場から，多重因子分析とクラスター分析を用いて物語文学と日記文学を分類し，書き手による文体差よりもジャンルによる文体差の方が大きいことを明らかにした。また，物語文学と日記文学では，格助詞と助動詞の使用傾向に一定の差異が見られることも分かった。本節では，より日本語学的な立場から，格助詞や助動詞といったカテゴリーのレベルではなく，テクストで実際に用いられている語のレベルでの分析を行う。具体的には，対数尤度比（LLR）を用いて，各テクストに特徴的な変数を抽出する [7]。表 3 は，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。

表 3　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

[7] 特徴語抽出の打ち切り点は，LLR の平均値である。

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 7）。「桐壺」には，内容的に最も高い敬意を示す必要のある桐壺帝の行為の記述が多く含まれている（表 4 参照）。そのため，帝や皇族に対して用いられる最高敬語として「す」，「さす」が（「たまふ」とともに）多く使用される。帝の行為に関する記述がこれほど多い巻は他になく，これらの敬語は「桐壺」を特徴付けるものである。

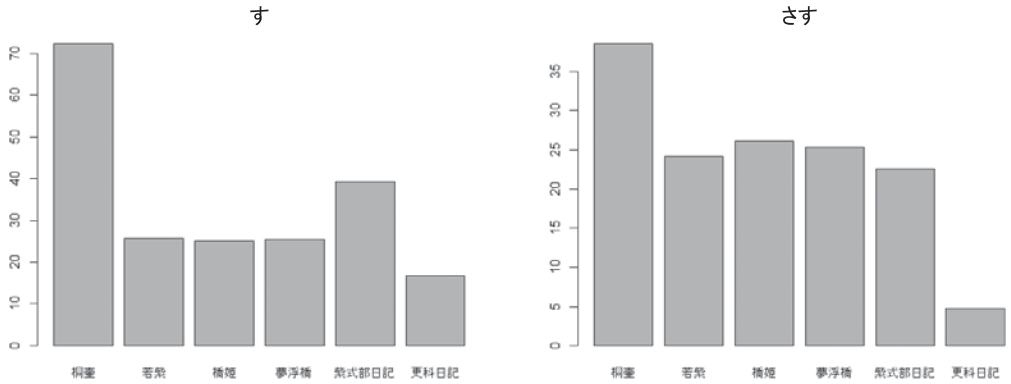

図 7　助動詞「す」，「さす」の分布

表 4 「桐壺」における助動詞「す」，「さす」の用例（一部）

前文脈	キー	後文脈
思さるれば，いま一階の位をだにと贈ら	せ	たまふなりけり。| これにつけても，憎み
て，後のわざなどにもこまかにとぶらは	せ	たまふ。| ほど経るままに，せむ方なう悲
たまはず，ただ涙にひちて明かし暮らさ	せ	たまへば，見たてまつる人さへ露けき秋
うのたまひける。| 一の宮を見たてまつら	せ	たまふにも，若宮の御恋しさのみ思ほし
ばかしうものたまはせやらずむせかへら	せ	たまひつつ，かつは，人も心弱く見たて
まどろませたまふことかたし。| 朝に起き	させ	たまふとても，明くるも知らでと思し出
きこしめさず，朝餉のけしきばかりふれ	させ	たまひて，大床子の御膳などは，いとは
まふ。| 七つになりたまへば読書始などせ	させ	たまひて，世に知らず聡うかしこくおは
と思し定めて，いよいよ道々の才を習は	させ	たまふ。| 際ことにかしこくて，ただ人に
たまへば，宿曜のかしこき道の人に勘へ	させ	たまふにも同じさまに申せば，源氏にな

また，表 3 を見ると，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 8 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。

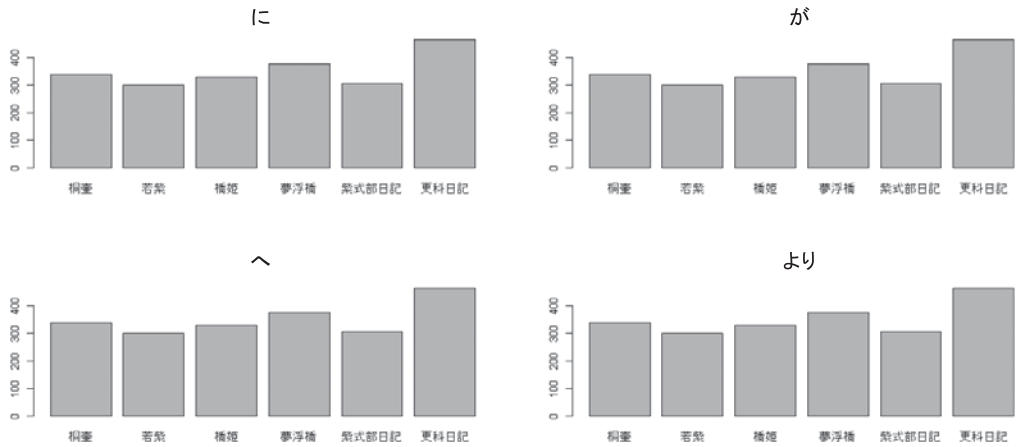

図 8　格助詞「に」,「が」,「へ」,「より」の分布

なお，表 5 に『更級日記』の格助詞の用例の一部を示す。

表 5　『更級日記』における格助詞の用例（一部）

前文脈	キー	後文脈
思ひはじめけることにか，世の中に物語	と	いふもののあんなるを，いかで見ばやと
けることにか，世の中に物語といふもの	の	あんなるを，いかで見ばやと思ひつつ，
といふもののあんなるを，いかで見ばや	と	思ひつつ，つれづれなるひるま，宵居な
ひつつ，つれづれなるひるま，宵居など	に	，姉，継母などやうの人々の，その物語，
ひるま，宵居などに，姉，継母などやう	の	人々の，その物語，かの物語，光源氏の
，宵居などに，姉，継母などやうの人々	の	，その物語，かの物語，光源氏のあるや
などに，姉，継母などやうの人々の，そ	の	物語，かの物語，光源氏のあるやうなど，
，継母などやうの人々の，その物語，か	の	物語，光源氏のあるやうなど，ところど
の人々の，その物語，かの物語，光源氏	の	あるやうなど，ところどころ語るを聞く
源氏のあるやうなど，ところどころ語る	を	聞くに，いとど床しいさ勝れど，わが

『更級日記』に格助詞が多いことは，書き手による文体の差と見てよいと考えられる。格助詞の性格上，テクストの内容や成立年代の違いが影響を与えていることも考えられるが，その可能性は次に見るように否定的である。

作品内容の影響として，たとえば，テクスト中で和歌が占める割合が想定される。そこで，『日本語歴史コーパス　平安時代編』で仮名文学作品全体について文体種別ごとの格助詞率を調査すると，地の文で約 13%，会話で約 10%，和歌で約 17%，全体で約 15% となっており，和歌が多いほど格助詞率は高まる。そして，『更級日記』は『源氏物語』や『紫式部日記』と比べて和歌が占める割合が高く，それによって格助詞率が高くなっていると説明することができそうである。しかし，『更級日記』の格助詞率は地の文で約 17%，会話で約 16%，和歌で約 20%，全体で約 17% と，

文体種別によらず格助詞率が高い。この数値は，男性の手による『土佐日記』とともに日記・物語全体の中でも際立って高い。このように，格助詞率の高さは，単に和歌の多寡によって説明できるものではない。むしろ，文体種別によらず格助詞率が高いことが個人文体としての特徴であることを示唆している。

一方，成立年代の差による影響としては，日本語の歴史上，時代を下るほど格助詞が明示されるようになるという言語変化が想定される。たしかに，『更級日記』は，『源氏物語』に遅れること約 50 年のちの成立である。しかし，より成立年代が下る『讃岐典侍日記』の格助詞率を調査すると，地の文で約 15%，会話で約 12%，和歌で約 19%，全体で約 14% と，『更級日記』ほどに高くはなく，むしろ平安仮名文学作品の平均に近い。従って，仮名文学作品の中だけで，格助詞使用率の歴史的上昇を指摘することは困難である。

このように，『更級日記』は，女流仮名文学作品の中で突出して格助詞の使用率が高く，これは菅原孝標女の個人文体によるものと言える。

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，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差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」といった文体種別情報を活用し，より詳細な分析を行う必要がある。

参照文献

Anderberg, Michael R. (1973) *Cluster analysis for applications.* New York: Academic Press.

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): 397–413.

Archer, Dawn (2009) *What's in a word-list? Investigating word frequency and keyword extraction.* Farnham: Ashgate.

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: 64–69.

Biber, Douglas (1988) *Variation across speech and writing.* Cambridge: Cambridge University Press.

Burrows, J. F. (1987) *Computation into criticism: A study of Jane Austin's novels and an experiment in method.* Oxford: Clarendon Press.

陳志文（2012）『現代日本語の計量文体論』東京：くろしお出版.

Dunning, Ted (1993) Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1): 61–74.

冨士池優美（2013）「『枕草子』長単位データを用いた格の類の分析」『第 3 回コーパス日本語学ワークショップ予稿集』291–298. 東京：国立国語研究所.

Gordon, A. D. (1999) *Classification.* 2nd ed. Boca Raton: Chapman and Hall.

Hoover, David (2001) Statistical stylistics and authorship attribution: An empirical investigation. *Literary and Linguistic Computing* 16(4): 421–444.

井上史雄・太田有多子（1989）「方言の地域差・個人差の多変量解析」『計量国語学』17(2): 49–63.
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): 225–240.
金明哲（2009）『テキストデータの統計科学入門』東京：岩波書店.
金明哲・樺島忠夫・村上征勝（1993）「読点と書き手の個性」『計量国語学』18(8): 382–391.
小林雄一郎（2013）「教師あり学習と教師なし学習を用いた芥川龍之介と太宰治の計量文体分析」『統計学的マイニング技術を応用したテクスト研究』（統計数理研究所共同研究リポート 298）3–13. 東京：統計数理研究所.
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): 84–92.
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』266–293. 東京：有精堂出版.
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): 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): 774–782.
Nakamura, Junsaku (2002) A galaxy of words: Structure based upon the distribution of verbs, nouns and adjectives in the LOB Corpus. In: Toshio Saito, Junsaku Nakamura and Shunji Yamazaki (eds.) *English corpus studies in Japan*, 19–42. Amsterdam: Rodopi.
Nakamura, Junsaku and John Sinclair (1995) The world of woman in the Bank of English: Internal criteria for the classification of corpora. *Literary and Linguistic Computing* 10(2): 99–110.
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』東京：勉誠社.
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』41–50. 東京：国立国語研究所.
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): 1–8.
小木曽智信・須永哲矢・冨士池優美・中村壮範・田中牧郎・近藤泰弘（2013）「『日本語歴史コーパス　平安時代編』先行公開版について」『第 3 回コーパス日本語学ワークショップ予稿集』269–276. 東京：国立国語研究所.
Pagès, Jérôme (2004) Multiple factor analysis: Main features and application to sensory data. *Revista Colombiana de Estadística* 27(1): 1–26.
Romesburg, H. Charles (1984) *Cluster analysis for researchers.* Belmont: Lifetime Learning Publications.
Scott, Mike and Christopher Tribble (2006) *Textual patterns: Key words and corpus analysis in language education.* Amsterdam: John Benjamins.
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: 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』125–132. 東京：情報処理学会.
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): 1–8.
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: 11–50.
上野英二（1991）「更級日記と文学史」『成城國文學論集』21: 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: 1–27.
上阪彩香・村上征勝（2013）「井原西鶴の『万の文反古』の文体分析」『情報処理学会研究報告』2013-CH-98(4): 1–8.
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: 65–71.
Wong, S., H. Gauvrit, N. Cheaib, F. Carré & G. Carrault (2002) Multiple factor analysis as a tool for studying the effect of physical training on the autonomic nervous system. *Computers in Cardiology* 29: 437–440.
安形輝・安形麻理（2009）「文書クラスタリングによる未解読文書の解読可能性の判定―ヴォイニッチ写本の事例」『三田図書館・情報学会誌』61: 1–23.
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: 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』1–22. 東京：明治書院.

Styles and Genres in Early Middle Japanese: A Multivariate Approach to Historical Corpus Stylistics

KOBAYASHI Yuichiro[a] OGISO Toshinobu[b]

[a]Research Fellow (PD), Japan Society for the Promotion of Science / Project Collaborator, NINJAL [–2013.10]
[b]Department of Corpus Studies, NINJAL

Abstract
The aim of the present study is to investigate styles and genres in Early Middle Japanese. By applying multivariate analysis to historical corpus stylistics, the present paper examines the frequencies of postpositional particles and auxiliary verbs in *The Tale of Genji*, *The Diary of Lady Murasaki*, and *The Diary of Lady Sarashina*, and visualizes in a multi-dimensional space the complex interrelationships among texts, the interrelationships among stylistic features, and the association patterns between texts and stylistic features. By so doing, we demonstrate that genres have more influence than writers on the style of a text. In addition, using log-likelihood ratios, we extract keywords from each text for more detailed analysis of the stylistic differences among texts.

Key words: style, Early Middle Japanese, multivariate analysis, keyword extraction