『計量国語学』アーカイブ

ID	KK290802
種別	研究ノート
タイトル	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—
Title	Statistical Reanalysis about “the Tale of Genji”: On the Data of Murakami and Imanishi (1999)
著者	小野　洋平
Author	ONO Yohei
掲載号	29巻8号
発行日	2015年3月19日
開始ページ	296
終了ページ	312
著作権者	計量国語学会

計量国語学 29 巻 8 号（2015 年 3 月）pp.296-312.

研究ノート

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―

小野　洋平（総合研究大学院大学複合科学研究科統計科学専攻大学院生）

要旨

源氏物語の研究の記念碑である村上らのデータベースに基づき，統計的な解析を行ったのが村上・今西（1999）である．彼らは源氏物語の全巻を 4 グループにわけ，数量化 3 類の 2 次元布置から，それらの成立順序に関して示唆的な結果をもたらした．しかし，本研究では村上・今西（1999）の数量化の解釈をより明瞭にするために，数量化 3 類の 5 次元布置の座標から計算される距離データに関してクラスター分析にかけたが，解釈不能な結果しか得られなかった．そこで，次に村上らのデータに分散安定化変換を適用し，変換後の値をクラスター分析にかけることで，村上・今西（1999）で推測された源氏物語の成立論に関する結果を示唆する樹形図を得た．この結果は，村上・今西（1999）の推論とは矛盾しないが，より統計学的に厳密にいうと，高々2 グループの成立順序を示唆するものであった．今後は，より綿密に，文献学的知見と統計学的手法の融合による，相補的な考察が深められるべきである．

キーワード：源氏物語，成立論，紫の上系，玉鬘系，第二部，匂宮三帖，宇治十帖，分散安定化変換，Box-Cox 変換，クラスター分析

1．はじめに

源氏物語は全 54 巻にわたる長編小説で，成立後千年を過ぎている．紫式部（930? ～ 1014?）が書いたとされ，平安時代の貴族生活を背景に光源氏の恋と栄華，そして平安朝における貴族社会の生態を描き出し，我が国の古典の最高峰と目され，諸外国にも広く翻訳され，古くから研究がなされてきた．

源氏物語 54 巻は全体の構成の観点から以下の 3 部に分けるのが通説になっている．（池田，1951）

第 1 部：（巻 1）「桐壺」～（巻 33）「藤裏葉」
第 2 部：（巻 34）「若菜上」～（巻 41）「幻」
第 3 部：（巻 42）「匂宮」～（巻 54）「夢浮橋」

特に，「源氏物語が如何にして成立したか」については今日まで様々な議論が繰り返され，今西・室伏（2010）によれば，

どうも，『源氏物語』成立論が不毛な水かけ論に堕してしまったのは，各研究者の用語や概念や思考方法が食い違っていたからであるように思えてなりませんでした．多くの

『源氏物語』研究者が，以下の四つの概念をゴチャマゼにしたまま議論をしているのです．

P，初巻の成立順序（各巻は，どういう順序で執筆されたか）
Q，初巻の配列（各巻をどのように並べるのが妥当か）
R，年立（物語内の時間は，どのように進行しているか）
S，作者（諸巻の作者は，一人か二人か，それ以上か）
この四点は，それぞれ別個に考察されるべき問題です．

（今西・室伏，2010:18-19）

たとえば，S に関していえば，後半の 45 巻から 54 巻までの宇治十帖を別作者とするものが有名である．この宇治十帖別作者説については，古くから一条兼良の「花鳥余情」，一条冬良の「世諺問答」などで，紫式部の娘である大弐三位が作者であるという説がある．これに対して，池田（1951）や大野（1984）は，否定的な態度をとっている．この説について計量的な分析を試みたのは，「宇治十帖の作者—文章心理学による作者推定」（安本，1957）や「文体統計による作者推定—源氏物語，宇治十帖作者について」（安本，1958）などが嚆矢であり，安本（1957）や安本（1958）では，源氏物語の 1 頁あたりの，和歌，直喩，声喩，色彩語，心理描写文，句点の数，および源氏物語の各巻から千字選んで，（各巻の最初の五百字と最後の五百字の）品詞分類を行い，各々の名詞，用言，助詞，助動詞，品詞のあわせて 11 個について，宇治十帖の 10 巻と，それ以外の 44 巻との間での統計的検定により両者の作者が同じとは言いがたいと結論づけている．これに対して，新井（1997）は五十音図の頭子音行列と母音列の頻度データに基づいて，宇治十帖別作者説を否定している．近年の研究（Tsuchiyama and Murakami，2013）では，主成分分析を名詞，動詞，形容詞，形容動詞，副詞，助動詞，助詞に対して適用した．またランダムフォレストという手法を用いて特徴量を抽出した上で，主成分分析にかける方法も行っている．しかし，結果は宇治十帖別作者説に対して否定的なものであった．

一方，P，に関して「源氏物語」が現在みられる巻序で成立したものでないという可能性を最初に指摘したのは，1922 年に発表された，和辻哲郎の「源氏物語について」（後に『日本精神史研究』に所収）である．和辻は，「帚木」巻の冒頭部の記述について分析し，以下のように述べている．

（略）... すなわち周知の題材の上にまず短い『源氏物語』が作られ，それに後からさまざまの部分が付加せられたと見るのである．が，いずれであるにしても，とにかく現存の『源氏物語』が桐壺より初めて現在のままに序を追うて書かれたものでないことだけは明らかだと思う．

（和辻，1926:209-210）

その後，「源氏物語執筆の順序」（阿部，1939），「源氏物語成立攷」（玉上，1940）などを経て，「源氏物語の研究」（武田，1954）において，前半部 33 巻のうち，長編的要素からなる紫の上系 17 巻と短編的性格の強い玉鬘系 16 巻との異質性を根拠に，玉鬘系 16 巻は紫の上系 17 巻が成立した後に執筆され，現行の順に挿入されたものであるという，源氏物語成立論が展開されるに至った．「源氏物語の研究」（武田，1954）において，武田は

各巻の登場人物に関して表 1 を示した.

一見してわかるように，玉鬘系に出てくる人物は紫の上系の話には登場せず，紫の上系 17 巻が成立した後に，玉鬘系 16 巻が執筆されたという一つの根拠となっている．さらに，武田は「源氏物語の研究」（武田，1954）におさめられた別稿「源氏物語の最初の形態再論」において，紫の上系 17 巻の成立の後に，玉鬘系 16 巻が執筆されたと考えられる根拠について 11 点をあげている．中心となる 4 点を引用する．（引用文中で使われていた旧字はすべて新字になおした）

一、第一部三十三帖中紫上系十七帖だけで連続統一をもったものとして完結した物語である。
二、玉鬘系十六帖は一見バラバラのように見えるが全体を通じて脈絡があり、紫上系とは別の統一を持って居る。
三、玉鬘系の巻々の事件・人物共に紫の上系の物語上に痕跡を与えず、紫上系は玉鬘系より独立して居り、三十三帖中玉鬘系十六帖を除き去っても何等の支障を来さない。
四、それだけで連続統一を持つ紫上系の物語の巻々の所々に玉鬘系の巻々が入って居る為、紫上系の物語を切断して、無理に割り込ませた形になって居り、紫上系から玉鬘系、玉鬘系から又紫上系へのうつりに、不自然さがある。

（今西・室伏，2010:215-216）

武田説は多くの賛同者を集めたが，一方で激しい批判も受け，論争は続いている．

表1：武田（1954）における紫の上系と玉鬘系の登場人物の分布を示した表

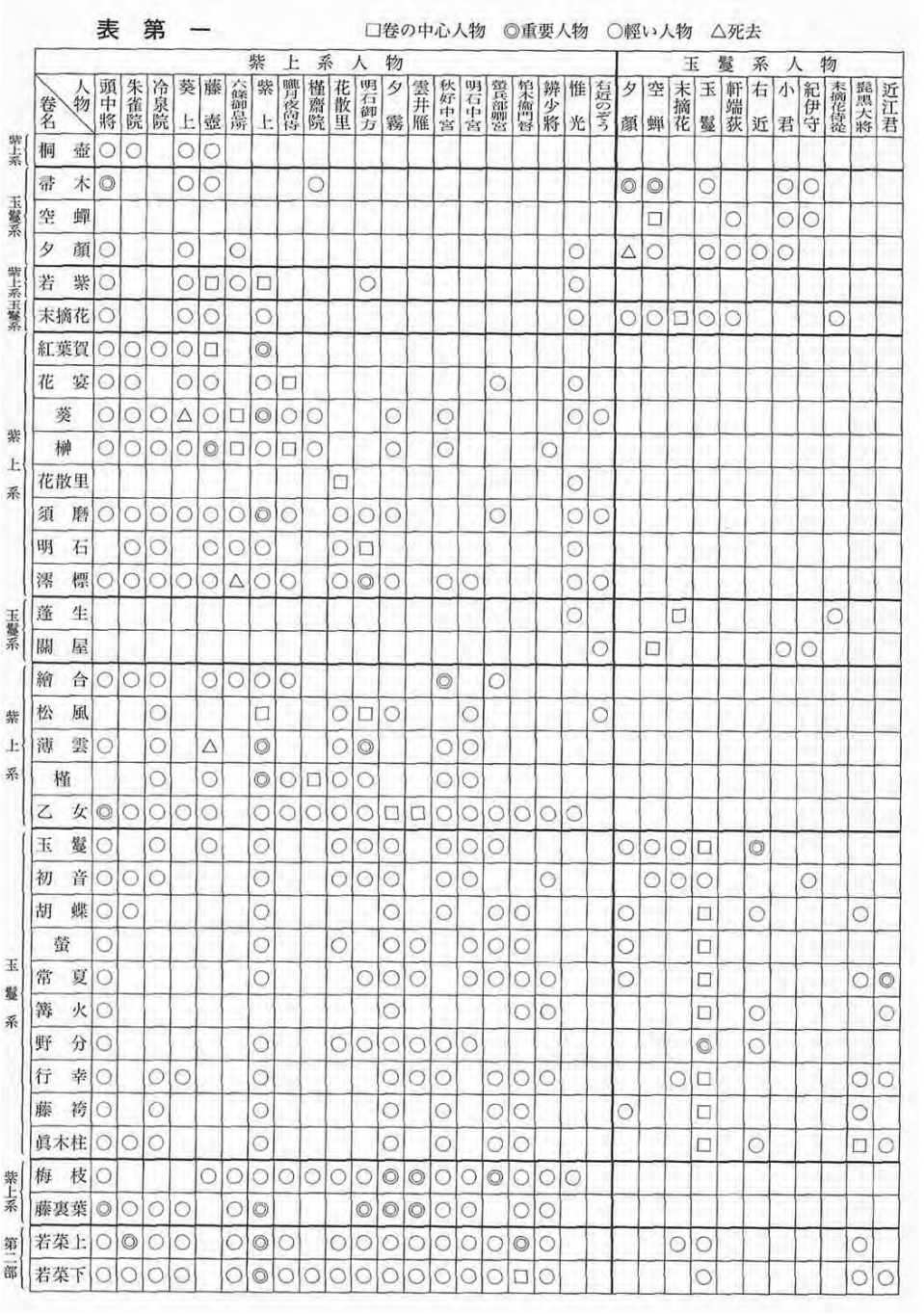

表　第　一　　□巻の中心人物　◎重要人物　○軽い人物　△死去

	巻名＼人物	頭中将	朱雀院	冷泉院	葵上	藤壺	六條御息所	紫上	贈答侍	權齋院	花散里	明石御方	夕霧	雲井雁	秋好中宮	明石中宮	螢兵部卿宮	鉦鼓侍從	辨少將	惟光	右近のぞう	夕顔	空蟬	末摘花	玉鬘	軒端荻	右近	小君	紀伊守	末摘花侍從	髭黒大將	近江君
紫上系	桐壺	○	○		○	○																									
玉鬘系	帚木	◎			○	○				○												◎	◎		○			○	○		
	空蟬																						□			○		○	○		
	夕顔	○			○		○													○		△	○		○	○	○	○			
紫上系玉鬘系	若紫	○			○	□	○	□				○								○											
	末摘花	○			○	○		○												○		○	○	□	○	○			○		
紫上系	紅葉賀	○	○	○	○	□		◎																							
	花宴	○	○		○	○		○	□								○			○											
	葵	○	○	○	△	○	□	◎	○	○			○		○					○	○										
	榊	○	○	○	○	◎	□	○	□	○			○		○				○												
	花散里										□									○											
	須磨	○	○	○	○	○	○	◎	○		○	○	○				○			○	○										
	明石		○	○		○	○	○			○	□								○											
	澪標	○	○	○	○	○	△	○	○		○	◎	○		○	○				○	○										
玉鬘系	蓬生																			○				□					○		
	關屋																				○		□				○	○			
紫上系	繪合	○	○	○		○	○	○	○						◎		○														
	松風			○				□			○	□	○			○					○										
	薄雲	○		○		△		◎			○	◎			○	○															
	槿			○		○		◎	○	□	○	○			○	○															
	乙女	◎	○	○	○	○		○	○	○	○	○	□	□	○	○	○	○	○	○											
玉鬘系	玉鬘	○		○		○		○			○	○	○		○	○	○					○	○	○	□		◎				
	初音	○	○	○				○			○	○	○		○	○			○				○	○	○			○			
	胡蝶	○	○					○					○		○		○	○				○			□		○			○	
	螢	○						○			○		○	○		○	○	○				○			□						
	常夏	○						○				○	○	○		○	○	○	○			○			□					○	◎
	篝火	○											○					○	○						□		○				○
	野分	○						○			○	○	○	○	○	○									◎		○				
	行幸	○			○	○		○					○	○	○		○	○	○					○	□					○	○
	藤袴	○		○				○					○		○		○	○				○			□					○	
	眞木柱	○	○	○				○					○		○		○	○							□		○			□	○
紫上系	梅枝	○				○	○	○	○	○	○	○	◎	◎	○	○	◎	○	○	○											
	藤裏葉	◎	○	○	○		○	◎				◎	◎	◎	○	○		○	○												
第二部	若菜上	○	◎	○	○		○	◎	○		○	○	○	○	○	○	○	◎	○					○	○					○	
	若菜下	○	○	○	○		○	◎	○	○	○	○	○	○	○	○	○	□	○						○					○	○

２．本研究が検証する先行研究 - 村上・今西（1999）

2.1　村上・今西（1999）の概要

上述した研究史の上でも，P（成立順序）に関して，計量的な分析を試みた記念碑的論文と言えるものが，村上・今西（1999）である[1]．村上・今西（1999）では，54 巻各巻の総語数にたいして，「ず」,「む」,「たり」,「けり」,「なり」,「り」,「ぬ」,「き」,「べし」,「つ」,「る」,「す」,「めり」,「さす」,「らむ」,「らる」,「じ」,「けむ」,「まじ」,「まし」,「まほし」の出現率の多い上位 21 語の助動詞の各 54 巻での出現率をもとに，分析を行った．

村上・今西（1999）は，源氏物語 54 帖を以下のように分類し，紫の上系を A グループ，玉鬘系を B グループ，第 2 部及び匂宮三帖を C グループ，宇治十帖を D グループとした．

A　紫の上系 ：1，5，7，8，9，10，11，12，13，14，17，18，19，20，21，32，33
B　玉鬘系　：2，3，4，6，15，16，22，23，24，25，26，27，28，29，30，31
C　第二部　：34，35，36，37，38，39，40，41
C　匂宮三帖 ：42，43，44
D　宇治十帖 ：45，46，47，48，49，50，51，52，53，54

村上・今西（1999）では，54 巻と 21 の助動詞の変数に対して，数量化 3 類を適用し，得られた 2 次元の布置図から，主に以下の 4 つの結論を得ている．

1．A グループ（紫の上系）と B グループ（玉鬘系）はあまり重なっておらず，これは武田宗俊の源氏物語成立論と関係が示唆される．
2．C グループ（第二部と匂宮三帖）と D グループ（宇治十帖）とは重なりがあまりない．
3．全体の布置を見ると，A グループ（紫の上系）と C グループ（第二部と匂宮三帖）とが重なり，B グループ（玉鬘系）と D グループ（宇治十帖）とが重なっている．
4．A グループ（紫の上系）より D グループ（宇治十帖）の方があとに書かれたという前提に立てば，執筆の順序は，A グループ（紫の上系）→ C グループ（第二部と匂宮三帖）→ B グループ（玉鬘系）→ D グループ（宇治十帖）と考えられる．

特に 4 番目の結論は，紫の上系成立後に玉鬘系が挿入されたという，武田宗俊の説と合致している．

2.2　村上・今西（1999）の限界

村上・今西（1999）では，数量化 3 類を用いて各巻の布置から，A グループ，B グループ，C グループ，D グループの大まかな関係を把握していた．ただし，彼らは 2 次元目までの結果に基づいて考察している．筆者はこうした数量化 3 類の布置の大まかなグループ分類が，より客観的な方法によって示されることを期待して，数量化 3 類で得られた布置

1　村上・今西（1999）によれば，この分析を行うために，「源氏物語大成」（池田 , 1985）をもとにし，源氏物語 54 巻の全文を単語に分割した上で，品詞コード等の数量分析に必要な情報をつけた約 37 万 6 千語のデータベースを作成し，このデータベースをもとに分析を行っている．古文データがコンピュータを用いて形態素解析することが可能な今日と異なり，当時はすべて手作業でデータを作成しており，この意味でも村上・今西（1999）は記念碑的論文である．

にクラスター分析を適用した．次元の決定には，たとえば因子分析でみられるように固有値が一定の値以上のものを選ぶ方法（Bartholow, Knott and Moustaki, 2011）や，固有値のスクリープロットから，固有値の減少が緩やかになる点の一つ前の次元までを採用する方法（足立・村上，2011）などがあるが，本稿では現時点の一つの選択肢として後者を選択した．数量化3類の結果得られた固有値のスクリープロット（図1）から，5次元とするのが適切と判断し，5次元までの座標を用いた．以降統計的な分析や図の出力にはR（2014）を使用した．

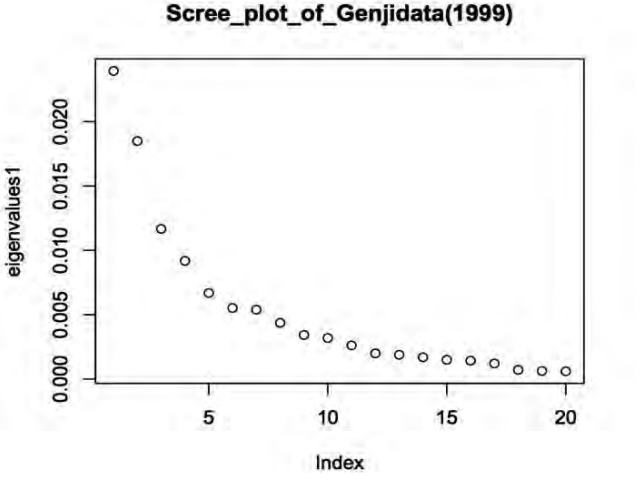

図1：村上・今西（1999）のデータに対して数量化3類を適用し，固有値をプロットした結果．6次元から急に固有値の減少が緩やかになることから，クラスター分析には5次元目までの座標を用いた．

クラスター分析（ユークリッド距離，ウォード法（Ward, 1963））の結果得られた樹形図は図2に示すように，A，B，C，Dの区分が全くできていない．クラスター分析の方法のいくつかを試みたがいずれも同様の結果である．5次元目までの累積寄与率は65％であった．また，数量化3類の布置を2次元までとして，クラスター分析（ユークリッド距離，ウォード法）を適用した結果を図3に示す．図2と同様にA，B，C，Dの区別が全くできていない．2次元目までの累積寄与率は40％であった．

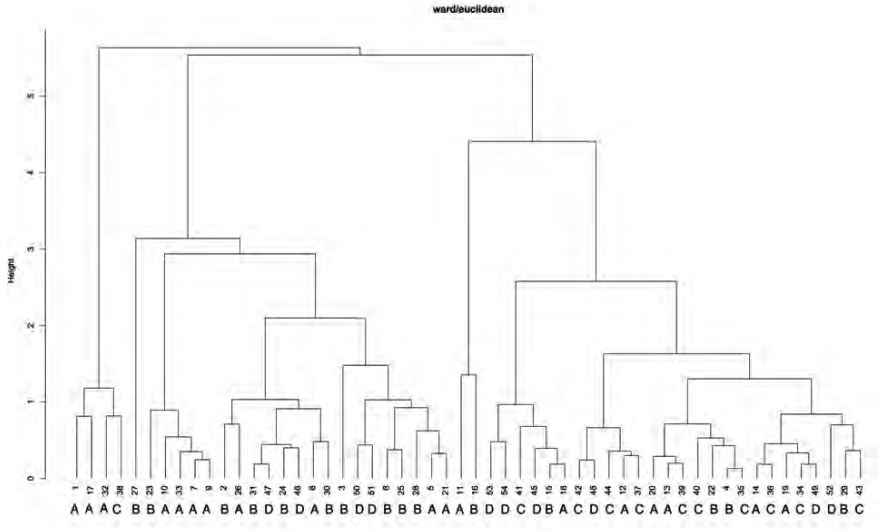

図 2：数量化 3 類を適用し，五次元目までの座標をもとにクラスター分析（ユークリッド距離，ウォード法）を行った結果得られた樹形図．解釈することが難しい．

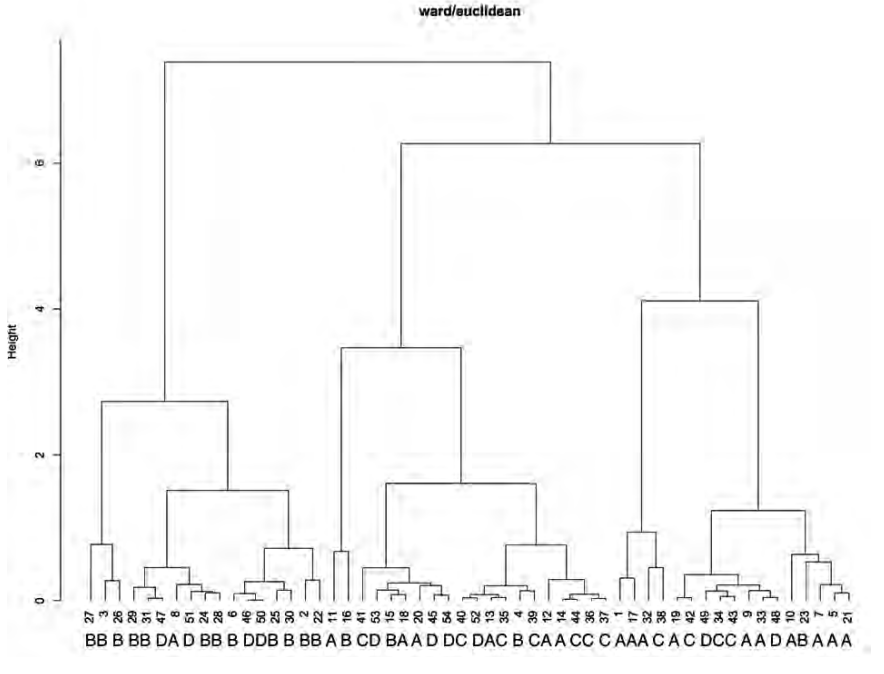

図 3：数量化 3 類を適用し，二次元目までの座標をもとにクラスター分析（ユークリッド距離，ウォード法）を行った結果得られた樹形図．左のクラスターに B がまとまっているが，右の 2 つのクラスターでは A，C，D の判別が難しい．

先に示した，村上・今西（1999）の 4 番目の主張を積極的に支持するためには，数量化 3 類の結果得られた 2 次元の布置から源氏物語の成立順序を推測するだけでは不十分であり，これが村上・今西（1999）の研究の限界であると言える．

3．本研究の目的

村上・今西（1999）のデータを用いて，村上・今西（1999）の 4 つ目の結論，「執筆の順序は，A グループ（紫の上系）→ C グループ（第二部と匂宮三帖）→ B グループ（玉鬘系）→ D グループ（宇治十帖）と考えられる」という主張を検討する．その際，「解釈」が数量化の結果の図を眺めてアドホックになるのではなく，より客観的なものになるように，結果はクラスター分析によって得られた樹形図にて示す．

4．方法

本研究では，村上・今西（1999）のデータ（54 巻× 21 変数）に対して，分散安定化変換（Hocking, 2013）の一種である Box-Cox 変換（Box and Cox, 1964）を適用し，適用した値にクラスター分析を行った．

村上・今西（1999）のデータは，各巻の総語数の大きさを一重中心化（行中心）することによって，総語数の大きさの効果を消している．一重中心化したことによって，データは見かけ上連続値のようになり，分散が過小になり，表面上はデータの性質がかわっているようにも見える．これは，総語数の影響を取り除くためには必要な処理であると考えられる．

よって，村上・今西（1999）のデータを扱うにあたっては，総語数の影響を取り除いたデータであるということを留意して，分析にあたる．クラスター分析を村上・今西（1999）のデータに直接適用した場合は，樹形図を本稿では提示していないが，解釈不能な結果を得た．

本研究では，村上・今西（1999）のデータに対して，分散安定化変換の一種である Box-Cox 変換を適用する[2]．なぜならば，一般に多くの分布においては，平均値と分散の値に同じ変数が入っている．このような場合，クラスター分析などにかけるために距離行列を計算する際に，導かれた距離が，「平均値による影響なのか」，「分散による影響なのか」，わからない．

分散安定化変換を施すことによって，第一に平均値と分散の値が変換後分離することができ，また分散の値は定数や，標本数に依るだけであり，平均値とは関係がなくなる（これを分散の安定化という）．第二に，変換された変数が従う漸近分布は正規分布である（漸近正規性）．正規分布は μ と σ だけできまり，それらは互いに独立している．

よって，分散安定化変換（本研究では Box-Cox 変換）をかけた後にクラスター分析を行うことは，各 54 巻の変数である各助動詞の分布の情報を平均値に集約し，各巻同士の分布の差をより効率よく取り出すことに他ならない[3]．

2　Box-Cox 変換の実用上の利点などについては，Osborne（2010）が詳しい．

3　「効率よく」とは、分散の安定化によって各巻の分散のバラツキがなくなり，分散がほとんど等しくなり，その代わりに平均に各巻のバラツキの情報が集約されていることをさす．各巻の分散がほとんど等しい状況で，ある 2 つの巻での 21 変数の距離をユークリッド距離で求めることは各巻の分布の平均の差をとっていることに他ならないので，平均に情報が集約されればされるほど「効率がよい」

Box-Cox 変換とは以下の式で表す変数変換である．X が元のデータであり，Y が変換後のデータである．

$$Y = \begin{cases} \dfrac{X^d - 1}{d} \ (d \neq 0) \\ \log X \ (d \to 0) \end{cases}$$

上式の d として代表的に使われるものとして，対数変換 d=0 の場合（つまり Y=logX），もしくはいわゆる平方根変換である d=0.5（つまり，Y=$\sqrt{\ }$X）があり，以降の章では，実際のデータの性質も参考にしながら，対数変換と平方根変換のどちらが村上・今西（1999）のデータに適しているかを考える．（実際には X が 0 をとるデータもあるため，対数変換においては Y=log（X+1）とした．）

永田（2011：53-54）によれば，分散安定化変換において，対数変換を採用する場合は，元のデータの標準偏差と平均が一定の比率になっている必要がある．また，平方根変換を採用する場合には元のデータの分散と平均が一定の比率になっている必要がある．以下，表 2 に村上・今西（1999）のデータの基本的な統計量を示す．

表 2：村上・今西（1999）の源氏物語の助動詞に関する統計量一覧．
X1 は各助動詞の 54 巻の平均，X2 は各助動詞の 54 巻の分散，X3 は X2 を X1 で割った値，X4 は各助動詞 54 巻の標準偏差を X1 で割った値．

助動詞	X1	X2	X3	X4
「ず」	0.0145223	0.0000032	0.0002232	0.1239665
「む」	0.0117019	0.0000154	0.0013140	0.3350919
「たり」	0.0116281	0.0000085	0.0007302	0.2505890
「けり」	0.0098042	0.0000069	0.0007006	0.2673210
「なり」	0.0096402	0.0000067	0.0006989	0.2692625
「り」	0.0091697	0.0000074	0.0008028	0.2958800
「ぬ」	0.0083970	0.0000039	0.0004603	0.2341199
「き」	0.0078275	0.0000123	0.0015688	0.4476903
「べし」	0.0075489	0.0000035	0.0004616	0.2472695
「つ」	0.0036193	0.0000025	0.0006966	0.4387062
「る」	0.0040273	0.0000024	0.0005859	0.3814197
「す」	0.0032710	0.0000036	0.0011002	0.5799591
「めり」	0.0025044	0.0000009	0.0003491	0.3733650
「さす」	0.0018256	0.0000009	0.0004797	0.5125997
「らむ」	0.0017003	0.0000005	0.0002892	0.4124435
「らる」	0.0017326	0.0000004	0.0002550	0.3836249
「じ」	0.0012774	0.0000003	0.0002681	0.4581634
「けむ」	0.0012763	0.0000005	0.0003602	0.5312360
「まじ」	0.0013328	0.0000009	0.0006929	0.7210261
「まし」	0.0010765	0.0000007	0.0006874	0.7990600
「まほし」	0.0005415	0.0000001	0.0002391	0.6644674

表 2 の X3 及び X4 の値を比較しても明らかなように，標準偏差と平均の比率と，分散と平均の比率とを比較してもどちらがより安定しているかは，判然としない．

ここで，源氏物語 54 巻の任意の二つの i 巻と j 巻の，21 の助動詞によるユークリッド距離を D*ij* とし，i 巻の k 番目の助動詞の値を *Zik* とすると，以下のような関係が成り立つ．

$$D_{ij} = \sqrt{\sum_{k=1}^{21} \left(Z_{ik} - Z_{jk}\right)^2} \{i, j = 1 \cdots 54\}$$

序で述べたように，仮に源氏物語の巻についていくつかのまとまりが存在していていれば，それぞれの助動詞の 54 巻にわたる値の分布は，いくつかの分布の混合分布となり，多峰性（multimodal）を示すはずである．しかし，村上・今西（1999）のデータを見れば明らかなように，実際の *Zik* の値は極めて小さく，k 番目の助動詞の 54 巻にわたる値の分布は，単峰性（unimodal）と見なせる．多峰性の分布に対する Box-Cox 変換の適用は本論の論旨を超えるものであり，今後の課題としたい．

そこで，本稿では助動詞の 54 巻にわたる値の分布，合計 21 の分布について，対数変換と平方根変換をかけ，正規分布により近い変換を与えたものをそれぞれの助動詞の分布で選び，全体 21 の分布において，対数変換がより多く選ばれれば，対数変換を全体のデータに適用することとする．同様に平方根変換がより多く選ばれれば，平方根変換を全体のデータに適用することとする．

変換した助動詞の分布の正規分布への当てはまりの良さに関しては，R の mclust のパッケージの中にある，データを正規分布にフィッティングさせる関数 densityMclust を利用し，指標としてベイズ情報量基準（Bayesian Information Criterion）（Schwarz, 1978）を用いた[4]．BIC が小さい値を与えた変換が統計学的には望ましい変換といえる[5]．BIC とは，モデルの良さを表す指標の一つであり，以下の式で求めることができる．

$$BIC = -2 \cdot (\text{最大対数尤度}) + (\text{自由度}) \cdot \log(\text{サンプルサイズ})$$

最大対数尤度，自由度，サンプルサイズといった用語に関しては，竹内編（1989）を参照のこと．さらに，ある助動詞の分布を対数変換し，R の mclust の densityMclust 関数を適用することで得られた BIC と，同じ助動詞の分布を平方根変換し，densityMclust 関数を適用することで得られた BIC は直接比較することはできない[6]．以下，表 3 にそれぞれの助動詞の分布に対して，対数変換と平方根変換を適用し，得られた BIC を適切に補正

4　以下 BIC と略す．

5　R の mclust のパッケージの中では，BIC の符号の向きが逆である．そのため，通常は BIC が最小のものを選べば良いが，mclust を使った計算では最大のものを選ばなければならない．本稿では煩雑さをさけるため，mclust の BIC の符号の向きを逆転させ，通常通り，解釈ができるようにした．R の mclust で BIC を求める際には留意する必要がある．

6　なぜならば，BIC の中の項をなす尤度の尺度が対数変換と平方根変換では変わってしまっているからである．詳細については，竹内編（1989）などを参照．本稿の表 3 では，対数変換の尺度を平方根変換の尺度に合わせるように補正を行った BIC を記す．

した値を示す.

表 3: 源氏物語の各助動詞の分布に対数変換と平方根変換を適用した際の BIC の値の違い. 2 列目は，対数変換の BIC が平方根変換の BIC に対して小さい場合に○，大きい場合に×を記した．3 列目は平方根変換の BIC が対数変換の BIC に対して，小さい場合に○，大きい場合に×を記した.

助動詞	対数変換	平方根変換	対数変換の BIC	平方根変換の BIC
「ず」	○	×	-368.23	-367.51
「む」	×	○	-275.39	-278.08
「たり」	×	○	-303.10	-306.17
「けり」	×	○	-293.17	-308.68
「なり」	○	×	-327.22	-327.00
「り」	×	○	-297.04	-297.33
「ぬ」	×	○	-326.49	-327.89
「き」	×	○	-259.15	-265.30
「べし」	○	×	-327.38	-322.58
「つ」	○	×	-301.19	-300.67
「る」	×	○	-312.89	-319.76
「す」	×	○	-294.66	-297.14
「めり」	○	×	-339.21	-333.95
「さす」	×	○	-328.45	-330.95
「らむ」	×	○	-351.19	-351.26
「らる」	×	○	-356.85	-356.93
「じ」	×	○	-351.42	-354.67
「けむ」	○	×	-342.94	-342.69
「まじ」	○	×	-332.18	-331.54
「まし」	×	○	-292.47	-297.14
「まほし」	×	○	-349.28	-352.86

表 3 の結果から，21 の助動詞のうち，14 の助動詞で平方根変換が対数変換よりも適しているという結果を得た．よって，本稿においては全体のデータに平方根変換を適用することとした．ただし，BIC は原理的には候補となっている条件の BIC のうち最も小さいものを選べばよいが，表 3 の結果を見ると，対数変換の BIC と平方根変換の BIC の差はわずかなものが多い．このような場合に，どのような判断をするべきかについては今後の検討課題としたい．また，本研究では，21 の助動詞のうち，多数の助動詞で採用された平方根変換を採用したが，このように変数ごとに変換の候補が分かれたときに，多数のものを採用するべきかについても今後の研究が必要である.

さらに，今回は対数変換と平方根変換のどちらを選ぶべきか，ということに絞って論じてきたが，Box-Cox 変換の d には 0（対数変換），0.5（平方根変換）以外にもさまざまな値が入りうるので，どのような d の値が適切かも課題である.

5．結果

村上・今西（1999）のデータ（54 巻× 21 変数）に対して平方根変換を適用し，ユークリッド距離で距離行列を求め，ウォード法によってクラスタリングを行って得られた樹形図が図 4，マンハッタン距離で距離行列を求め，ウォード法によってクラスタリングを行って得られた樹形図が図 5，ユークリッド距離で距離行列を求め，最長距離法（Sørensen, 1948）によってクラスタリングを行って得られた樹形図が図 6 である．

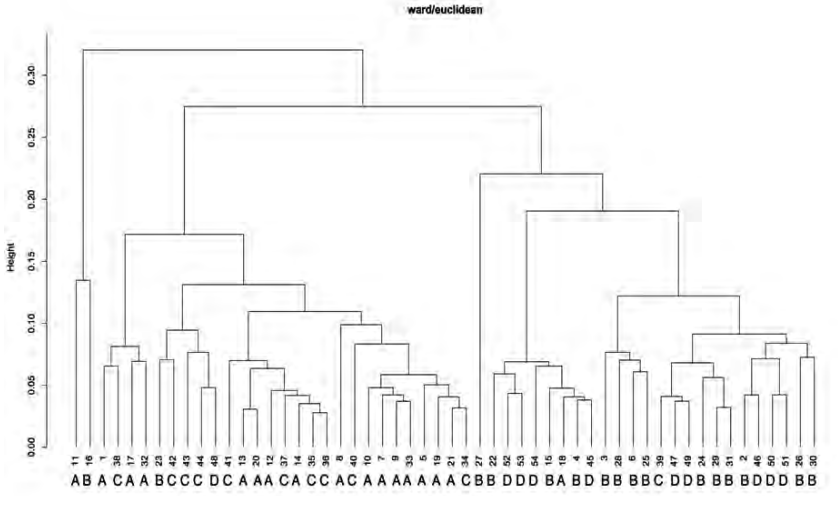

図 4：村上・今西（1999）のデータを平方根展開し，変換後の値にクラスター分析（ユークリッド距離，ウォード法）を適用し，得られた樹形図．左のクラスターに A と C が，右のクラスターに B と D が集まっている．

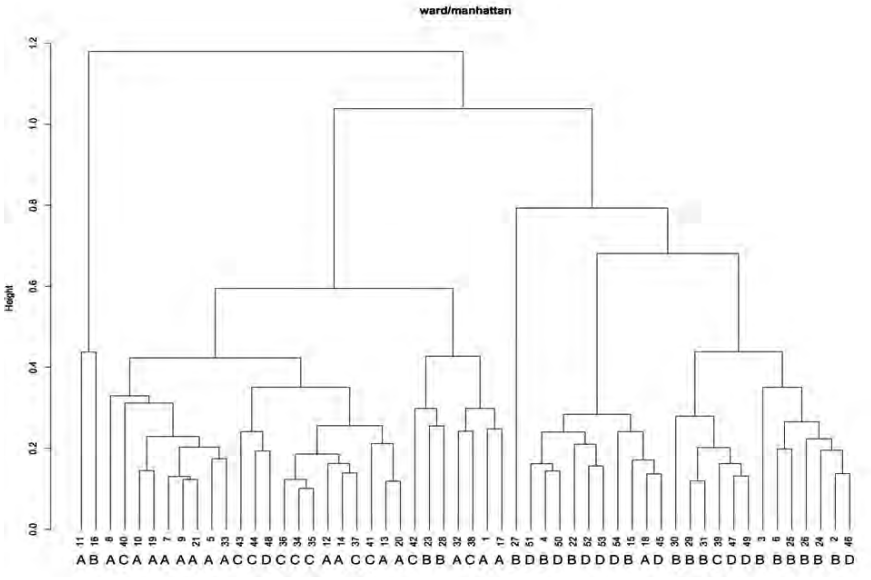

図 5: 村上・今西（1999）のデータを平方根展開し，変換後の値にクラスター分析（マンハッタン距離，ウォード法）を適用し，得られた樹形図．左のクラスターに A と C が，右のクラスターに B と D が集まっている．

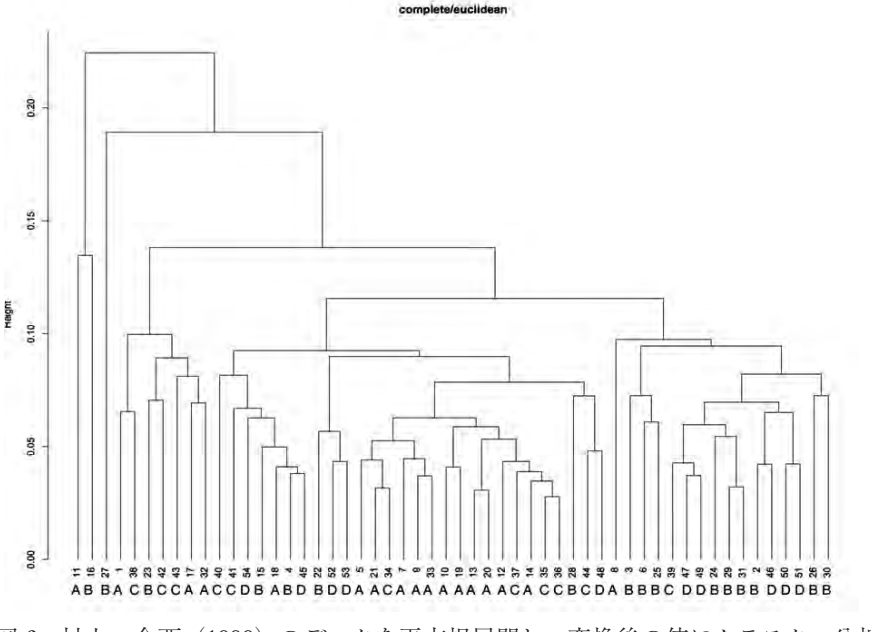

図 6: 村上・今西（1999）のデータを平方根展開し，変換後の値にクラスター分析（ユークリッド距離，最長距離法）を適用し，得られた樹形図．総語数がすくない最も左のクラスターを除けば，最も右のクラスターに B と D が集まり，それ以外に A と C が集まっている．

村上・今西（1999）のデータに対して平方根変換を適用した値を，クラスター分析（ユークリッド距離，ウォード法）にかけた図 4 に関しては，極端に総語数が少ない最も左のクラスター（11）花散里，（16）関屋を除いて考えると，左のクラスターに A と C が，右のクラスターに B と D が集まっている．その中で，左のクラスターに B の（23）初音と D の（48）早蕨が，逆に，右側のクラスターに A の（18）松風，C の（39）夕霧が混在している．

また，平方根変換を適用した値をマンハッタン距離とウォード法でクラスター分析を行った図 5 に関しては，極端に総語数が少ない最も左のクラスター（11）花散里，（16）関屋を除いて考えると，左のクラスターに A と C が，右のクラスターに B と D が集まっている．その中で，左のクラスターに B の（23）初音，（28）野分と D の（48）早蕨が，逆に，右側のクラスターに A の（18）松風，C の（39）夕霧が混在している．

最後に，平方根変換を適用した値をユークリッド距離と最長距離法でクラスター分析を行った図 6 に関しては，極端に総語数が少ない最も左のクラスター（11）花散里，（16）関屋を除いて考えると，最も右のクラスターに B と D が，それ以外のクラスターに A と C が集まっている．その中で，左のクラスターに, B の（4）夕顔，（15）蓬生，（22）玉鬘，（23）初音，（27）篝火 , (28）野分と D の（45）橋姫，（48）早蕨，（52）蜻蛉，（53）手習，（54）夢浮橋が，逆に右側のクラスターに A の（8）花宴，C の（39）夕霧が混在している．

通常 , クラスター分析の結果は，分析に用いる距離，および手法によって大きな影響をうけるが，図 4，図 5，図 6 の結果から，源氏物語 54 巻は一貫して（A，C）と（B，D）とに分かれ，分類の良さは図 4 では 92％，図 5 では 90％，図 6 では 75％ であることから，クラスター分析の結果得られた（A，C）と（B，D）という分類は比較的安定していると考えられる．

6．考察

図 5 から，まず A，C と B，D とが 90％ を超える高い精度で分類されたことがわかる．このことから，A と B，C と D も自ずと高い精度で分類されたことがわかる．また，A よりも D の方が後に書かれたことを仮定すれば，（A，C）が書かれ，後に（B，D）が書かれたということが導かれ，これは，紫の上 17 帖の成立の後に独立して玉鬘 16 帖が書かれたとする武田説を支持するものである．

ただし，村上・今西（1999）で示唆されたような「A → C → B → D」の成立順序を主張するものではない . なぜなら，図 4 の左側のクラスターでは A と C とを分けることは難しく，同様に，右側のクラスターでは B と D とを分けることは難しいからである．

すなわち，P（成立順序）の問題として図 4 をとらえると，執筆時代によって，助動詞の使い方が変化すると仮定するならば，源氏物語は「（A，C）→（B，D）」の順にかかれたと結論することが可能である．

また，今までは 2 章の今西・室伏（2010）の引用で紹介したように，P（成立順序）の問題としてこのクラスターを考察したが，実際には S（作者）の問題としても，図 4 をとらえ直すことができる．

S（作者）の問題としてとらえると，助動詞の使い方が古典の作者の特徴をあらわして

いると仮定するならば，源氏物語には（A，C）を書いた作者と（B，D）を書いた作者の少なくとも二人を想定することができる．もっとも，これは一人の作者の助動詞の使い方が経年効果で変化した結果，二人の作者がいるようにみえるということも想定しなければならない．これについては，現代の作家であるが村上（2002）における川端康成の文体の変化に関する分析が参考になる．川端の作品では，戦前の作品は読点の前の文字が「と」である比率が高く，逆に戦後の作品は読点の前の文字が「し」である比率が高く，経年変化による文体の変化がみられる（村上 2002：56）．

さらに，今後の課題としては，図 4 において，（18）松風，（23）初音，（39）夕霧，（48）早蕨が，なぜ混在したのか検討が必要である，助動詞だけのデータでは情報量が足りないと考えるべきか，それとも今までの成立論では見落とされていた何かをこれらは物語っているのか．今後の学術的研究を待ちたい．

7．結語と今後の課題

本研究では，村上・今西（1999）の源氏物語の各巻の各助動詞の頻度比率のデータに対して，分散安定化変換の一種である平方根変換を適用し，さらに変換された値に対してクラスター分析を適用した．これにより，村上・今西（1999）では，数量化 3 類の結果得られた布置の「解釈」という形でしか見いだせなかった源氏物語の成立論に関する考察を，樹形図による明確な形で提示することができた．この結果は，村上・今西（1999）において「A → C → B → D」という成立順序の推察とは矛盾しないが，本稿の平方根変換の結果をクラスター分析にかけた結果では，より統計学的に厳密ではあるものの，「（A，C）→（B，D）」という成立順序が示された．

また，分散安定化変換は，これまでは実用上はむしろ回帰分析における不均一分散を解消するための手段として用いられることが多く，クラスター分析に用いられるのは例えば DNA の解析（Zwiener, Frisch and Binder, 2014）の際などのようであり，人文科学においては筆者の知る限りほとんど見られない．今回の分析から，何らかの変数変換がデータの持つかくれた特性を引き出すことがあることがわかったので，その適用範囲と限界についても可能性を探っていきたい[7]．今後は，より綿密に，文献学的知見と統計学的手法の融合による，相補的な考察が深められるべきである．

文献

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011) *Latent variable models and factor analysis: a unified approach* (Vol. 899). John Wiley & Sons.

Box, G. E., and Cox, D. R. (1964) An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 211-252.

Hocking, R. R. (2013) *Methods and applications of linear models: regression and the analysis of variance*. John Wiley & Sons.

7　もっとも，安易なデータ変換はデータの扱いをより困難にすることも我々は肝に銘じておかなければならないだろう．たとえば，正規分布にしたがうあるデータ X と別の正規分布にしたがうあるデータ Y の比の値（X/Y）は，コーシー分布に従う．コーシー分布は平均と分散および高次のモーメントが存在せず，扱いに注意が必要である．

Osborne, J. W. (2010) Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research & Evaluation*, 15(12): 1-9.

R Core Team (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
URL http://www.R-project.org/.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461-464.

Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biol. skr.*, 5: 1-34.

Tsuchiyama, Gen, and Murakami Masakatsu. (2013) Authorship Identification of Classical Japanese Literature Using Quantitative Analysis, *Journal of Mathematics and System Science*. 3:631-640.

Ward Jr, J. H. (1963) Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301): 236-244.

Zwiener, I., Frisch, B., and Binder, H. (2014) Transforming RNA-Seq data to improve the performance of prognostic gene signatures. *PloS one*, 9(1): e85150.

足立浩平，村上隆（2011）「非計量多変量解析法」朝倉書店.

阿部秋生（1939）「源氏物語の執筆順序」『国語と国文学』16（8~9）.

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.

池田亀鑑（1951）「新講源氏物語」至文堂.

池田亀鑑（1985）「源氏物語大成」中央公論社.

今西祐一郎，室伏信助（2010）「紫上系と玉鬘系―成立論のゆくえ」勉誠出版.

大野晋（1984）「源氏物語」岩波書店.

武田宗俊（1954）「源氏物語の研究」岩波書店.

玉上琢彌（1940）「源氏物語成立攷」『国語・国文』10（4）.

永田靖（2013）「統計学のための数学入門 30 講」朝倉書店.

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.

村上征勝（2002）「文化を計る：文化計量序説」朝倉書店.

竹内啓編（1989）「統計学辞典」東洋経済新報社.

安本美典（1957）「宇治十帖の作者―文章心理学による作者推定」『文学・語学』第 4 号 三省堂.

安本美典（1958）「文体統計による筆者推定―源氏物語・宇治十帖の作者について」『心理学評論』2（1）：147-156.

和辻哲郎（1926）「日本精神史研究」岩波書店.

（2014 年 9 月 28 日受付，2014 年 11 月 21 日再受付）

Mathematical Linguistics, Vol.29 No.8（March 2015）pp.296-312.

Note

Statistical Reanalysis about "*the Tale of Genji*": On the Data of Murakami and Imanishi（1999）

ONO Yohei（The Graduate University for Advanced Studies）

Abstract:
The main objective of this paper is to reconsider the work of Murakami and Imanishi（1999）about "*the Tale of Genji*". In Murakami and Imanishi（1999）, they divided "*the Tale of Genji*" into four groups and led the suggestive results on how "*the Tale of Genji*" was written, based on the plot of Hayashi's Quantification Method Ⅲ. However, applying cluster analysis to their plot, the present author cannot obtain the same result as dendrogram. Therefore, utilizing variance stabilizing transformation to their data and applying cluster analysis to those, the obtained dendrogram indicates that this result is consistent with the previous studies and more careful with respect to statistics. For future research, combining philological knowledge and statistical approach, those complementary researches will matter.

Keywords: stylometry, Murasaki Shikibu, *the Tale of Genji, Uji Chapter*,
variance stabilizing transformation, Box Cox transformation, cluster analysis