2017 年度博士論文

川端康成の代筆問題及び文体問題に関する計量的研究

同志社大学大学院文化情報学研究科
文化情報学専攻博士課程（後期課程）

48141002

孫昊

指導教員　金明哲教授

2017 年 12 月 28 日提出

目次

第 1 章 序論

本章では、まず、川端康成の生い立ちを概観し、本論文の着眼点となる川端康成の代筆問題と文体問題をまとめる。次に、各問題の先行研究をまとめた上で、残された問題点を挙げ、本論文の目的を述べる。最後に、本論文の構成を紹介する。

1.1 はじめに

川端康成は 1899 年 6 月 14 日に川端栄吉の長男として生まれた。川端栄吉は開業医をしていたが、肺を病んで 1901 年 1 月 17 日に 33 歳という若さで亡くなり、妻のゲンもその 1 年後の 1902 年 1 月 10 日に 37 歳でこの世を去った。川端家を襲った悲劇はそれだけではなく、川端康成の祖母・カネは、川端康成が小学校に入学した 1906 年に他界し、姉の芳子も 1909 年に 13 歳で夭折した。川端康成は中学 3 年生になった 1914 年に当時唯一の肉親であった祖父も死去し、この年から 15 歳の川端康成は天涯孤独の孤児となってしまった。川端康成は自身の文学作品に対して孤児の感情を表す「孤児根性」の姿勢を貫いており、「『孤児』は私の全作品、全生涯の底を通って流れる」と語ったこともある[1] (鳥羽, 1969)。家族のあまりにも早い死は幼い川端康成に病気と早死の恐れを与えてしまい、この感情は彼の作品内容だけでなく、精神状態にも影響を及ぼしている。幼い川端康成はすべての肉親を亡くしたショックで精神に異常を来たした。小説『少年』の中に、「私は幼年時代が残した精神の病患ばかりが気になって、自分を憐れむ念と自分を厭う念とに堪へられなかった。」と川端康成は自分の幼少期の精神状態を綴った。このような孤児になった喪失感から生まれた「幼年時代が残した精神病患」は長い間川端康成を苦しめ、のちに精神状態悪化の引き金となってしまった。作家になった川端康成は執筆のため昼夜逆転の生活を送り、不眠症を患って睡眠薬を用いるようになった。川端康成が初めて睡眠薬を用いたのは『東京の人』を執筆した 1954 年とされ、それからどんどんエスカレートし、1960 年頃から大量の睡眠薬を常用するほど精神が不安定になった (木幡, 1992)。1961 年 1 月から『婦人公論』に連載され始めた『美しさと哀しみと』の第一回分「除夜の鐘」の終わりに、「お断り—作者入院のため、少ししか書けませんでした。おゆるしください。」と記してある。川端康成の当時の精神状態は執筆にも影響を及ぼしていることが明らかである (小林, 1982)。

「孤児根性」の他に、川端康成小説のもう一つの主題は「死」である。祖父の病臥中に綴った『十六歳の日記』と最後の肉親である祖父を記念するために書いた『骨拾い』をはじめ、『父母への手紙』、『抒情歌』、『それを見た人達』、『慰霊歌』、『禽獣』、『散りぬるを』などの様々な角度から死を扱った作品が発表されている (木幡, 1992)。これほど「死」をモチーフとした作品を執筆した原因は、川端康成が幼少期に経験した相次ぐ家族の死にあると考えられる。川端康成の戦後の作品は「魔界の文学」と称され、彼は「死」について考えたあげく、その思想も作風も「魔界」に落ちてしまったとされている。川端康成自身もその「魔界」

[1]『川端康成全集』第 1 次全集第 2 巻付記に記されている。

から這い上がることができなかったせいか、彼がノーベル文学賞を受賞してからわずか 4 年後、栄光の絶頂期にガス自殺をしてしまった。

川端康成の研究では、特に注目を浴びているものは代筆問題と文体問題である。代筆問題は川端康成の名義で発表された作品には代筆者が存在する問題を指す。その代表的な作品として、『乙女の港』や『古都』などが挙げられる。文体問題は主に川端康成作品における文体存在問題と文体変化問題を指す。文体存在問題は川端康成が自分の文体を持っているかの問題である。文体変化問題は川端康成の執筆途中に文体の変化が生じたかの問題である。川端康成の代筆問題と文体問題を取り上げた先行研究は多数存在するものの、依然として解決の目処は立っていない。先行研究のレビューをしているうちに、結論を支える客観的な証拠の欠乏が問題解決に至らなかった一因であると気づいた。そこで、本論文では、川端康成の今までの代筆問題と、文体の存在・変化問題をまとめ、それぞれの問題点を示した上で文章から取得したデータに基づいた客観的証拠を提示し、川端康成の代筆問題と文体問題の解明を試みる。

1.2 代筆問題

1.2.1 代筆問題研究の前提

本論文では、研究の一環として川端康成の「代筆問題」を取り上げたが、文体解析によって代筆の客観的な証拠を探ることだけに目的を置き、本論文の結論を用いて川端康成を批判するつもりはない。その理由は川端康成が「好意的」に代筆をさせた行為にある。川端康成は「新人発掘の名人」として知られ、文壇の後輩の指導にも積極的に取り組んでいた。当時の無名な新人作家は自分の名前で作品を発表することが難しく、生活に困っていた者は少なくない。そこで、川端康成は執筆指導を行いながら、新人作家たちに自分の名義を貸して作品を発表させたという逸話が残っている (小谷野, 2013)。本論文では、川端康成はこのようにして好意的に代筆をさせたことを前提として代筆問題を論じる。

1.2.2 代筆問題研究の問題点

川端康成の代筆問題は昔から議論され、その一例として川端康成の代筆説を支持する矢崎(2003)の観点を次に示す。下線は本論文の筆者によるものである。

> 川端康成は若い頃から遅筆だった。したがって原稿の締切りに間に合わないことがしばしばあった。新聞小説を書くなんて、とうてい無理なことだった。それを敢えて引き受けてしまったのは、本人も含めて編集者たちの暗黙の了解が成立していたからである。川端は一日机の前に座っていても、四百字詰めの原稿用紙一枚仕上げるのが至難だった。新聞小説は少なくても二枚。<u>それが毎日なのだから、誰もが代作を認めていたのである。</u>

ここで矢崎 (2003)の言及した「それが毎日なのだから、誰もが代作を認めていたのである。」は、川端康成の背後には代筆者がいたことを示唆している。矢崎氏の父は川端康成の編集者

として勤めていたためある程度内部の事情を知っている可能性がないとは言えないが、具体的な証拠を提示しない限り、この説はあくまで憶測にすぎない。

川端康成の代筆問題についての研究は史料学と文学の分野に集中している。史料学の分野では、主に、川端康成と代筆者の間の書簡を代筆の証拠としている。書簡の内容は、川端康成からの原稿作成依頼や代筆者への執筆指導などである。このような研究は、書簡に記載された代筆についての内容を代筆か否かの判断基準としている。しかし、代筆依頼の書簡があったとしても、川端康成は執筆指導を行った際に代筆者の原稿に手を加えた可能性がある。この場合、川端康成も作品にかかわっているため、書簡だけでの代筆判断は難しい。そこで、原稿の内容か文体に基づいた分析が必要となる。

文学の分野において、研究者が研究対象となる文章を熟読した上に、文書内容に対する理解に基づいて代筆問題を分析する方法が主流である。このような研究は研究者の主観に依存するため、十人十色の結論になりかねない。また、文章の理解に基づいた分析方法には疑いの聲も上がり、ラボック (1957)は次のように述べている。

> 読み進むすぐ後から、作品は記憶の中でとけてゆき、変容しだす。最後のページをめくる瞬間にはもう、その作品の大部分、とくに微妙な点は、あいまいになり、覚束ないものになっている。さらにもう少し経って、数日、また数ヵ月になると、実際そのうちのどれ位が残っているだろうか？一群の印象、漠とした不確かな印象の中から現れてくる二三の明瞭な個所、一般的にいって、これが作品という名で残りそうなすべてである。それを読んだ経験が、後に何か残している。作品名で我われが想い起すのは、こういった名残にすぎない。こうしたものが、作品に判定を下し評価する材料をしかと与えてくれるなどと、どうして考えられようか。

このように、代筆の有無を判断する時に、作品内容はどれほど記憶の中に残っているかがもはや検証できないため、内省の方法が主観的であると疑われている。

1.2.3 代筆問題の研究対象

川端康成の代筆疑惑が持たれた作品は、大きく「代筆認定作品」と「代筆疑惑作品」に分けられる。「代筆認定作品」は 1984 年に完結した川端康成全集の編集の際に削除されたものを指す。削除の理由について川端康成 1984 年に完結した全集の第 1 巻の巻末に解題がある。削除された作品は、「一、他者の協力をあふいでなった著作」である。そのうち『小説の研究』の前半は伊藤整の代筆で、『小説の構成』は瀬沼茂樹の代筆である。「二、少年少女小説のうち、戦後に発表された作品」である。そのうち、『歌劇学校』は平山宮子の代筆であると平山城児が著作『川端康成－余白を埋める』で明かしている。他に削除されていた少年少女小説として、『万葉姉妹』、『花と小鈴』、『親友』と『長い旅』などもあったが、先行研究を調べた限りでは、いずれも削除された理由が記されていない (小谷野, 2013)。

本論文では、代筆問題が解明されていない「代筆疑惑作品」を研究対象とした。この「代筆疑惑作品」はさらに三つのカテゴリに分類できる。一つ目は少女小説である。川端康成名義

で発表した少女小説の『乙女の港』、『花日記』と『コスモスの友』は芥川賞を受賞した女性作家・中里恒子による代筆であると言われている (小谷野, 2013)。二つ目は川端康成の睡眠薬中毒時期小説である。その時期に発表された『古都』は川端康成の弟子の澤野久雄、北條誠、三島由紀夫による代筆と疑われ、同じ時期に発表された『眠れる美女』は三島由紀夫による代筆という説もある (板坂, 1997)。三つ目はその他小説である。『山の音』は三島由紀夫の代筆と言われている (板坂・鈴木, 2010)。川端康成の代筆疑惑作品、研究対象となる小説、代筆疑惑者と代筆の証拠を表 1.1 にまとめる。

表 1.1 川端康成の代筆疑惑作品

カテゴリ	小説名	代筆疑惑者	代筆の証拠
少女小説	乙女の港	中里恒子	書簡、原稿
少女小説	花日記	中里恒子	書簡、論文
少女小説	コスモスの友	中里恒子	論文
睡眠薬中毒時期	古都	北條誠、澤野久雄、三島由紀夫	書簡、証言
睡眠薬中毒時期	眠れる美女	三島由紀夫	証言
その他	山の音	三島由紀夫	証言

1.3 文体問題

川端康成は、詩的や抒情的作品、少女小説といった多彩な文体で執筆活動を行い、その多様な作風で「奇術師」と呼ばれていた。川端康成文体研究では、主に、「文体の存在」問題と「文体の変化」問題が注目されている。

先行研究では、川端康成の作品には文体が存在しない指摘され、いわゆる「文体不在論」である。その代表的なものは、三島 (1956)の論文「永遠の旅人ー川端康成氏の人と作品」で述べた内容である。

> たとえば川端さんが名文家であることは正に世評のとおりだが、川端さんがついに文体を持たぬ小説家であるというのは、私の意見である。なぜなら小説家における文体とは、世界解釈の意志であり鍵なのである。混沌と不安に対処して、世界を整理し、区画し、せまい造型の枠内へ持ち込んで来るためには、作家の道具とては文体しかない。フローベルの文体、スタンダールの文体、プルーストの文体、森鴎外の文体、小林秀雄の文体、……いくらでも挙げられるが、文体とはそういうものである。

三島 (1956)の他に、寺田 (1949)、臼井 (1952)と三田 (1994)も川端康成は「文体不在」の作家であると主張している。1753 年、フランスの博物学者ビュフォンはアカデミーフランセーズの入会演説で「文体は人なり」と述べて以来、文体は文章著者ならではの特徴を反映している説が広く受け入れられるようになった (中村, 2010)。川端康成の「文体不在」説が正しい

とすれば、これは「川端康成の文章には川端康成ならではの特徴が入っていない」というパラドックスになってしまう。このようなパラドックスが生じた原因、すなわち川端康成が文体を持たない作家と言われる原因は、主に、川端康成の代筆問題に関連すると考えられる。もし川端康成の作品は代筆者が書いた説が正しいとすれば、必然的にその文体には複数の代筆者の特徴が混在し、川端康成ならではの特徴が薄れてしまう。川端康成の「文体不在」説と異なり、文体研究者の中村 (2010)、は川端康成の文体特徴を「稲妻の文体」と名付けた。その理由は、作品中の絶えざる改行によって生まれた「閃き」の印象にある。また、川端文学は敗戦を境にして大きく変貌を遂げたと言われている (山中, 1999)。本論文では、川端康成の文体存在問題と文体変化の解明を試みる。また、川端康成は極力平易な表現を用い、限られた語彙で豊かな表現力を生み出していると知られている。本論文では、川端康成の語彙問題として語彙の豊かさの検証も行う。

1.4 コーパスの作成

1.4.1 川端康成コーパス

計量文体学の観点に基づいて川端康成の代筆問題を解明するには、コーパスの作成が必要である。本論文では、川端康成の全集に収録されている小説を用いることにした。川端康成の全集はかつて 4 回出版されている。第 1 回は 1948 年から 1954 年までの 16 巻全集である。第 2 回は 1959 年から 1962 年までの 12 巻全集である。第 3 回は 1969 年から 1974 年までの 19 巻全集である。第 4 回は 1980 年から 1984 年までの 37 巻全集である。この 4 回の全集の情報を表 1.2 に示す。

表 1.2 川端康成の全集リスト

出版回数	出版年度	巻数
第 1 回	1948~1954	16
第 2 回	1959~1962	12
第 3 回	1969~1974	19
第 4 回	1980~1984	37

川端康成の最新の全集は没後に刊行された 1984 年に完結した第 4 回の 37 巻本である。この全集には未刊行・未発表作品、プレオリジナル、新発見の日記などが収められ、代筆疑惑の作品も多く入っている (小林, 1982)。第 4 回の全集と比べ、1974 年に完結した第 3 回の全集には代筆疑惑作品の収録は少ない。第 3 回と第 4 回全集における代筆疑惑作品の収録状況を表 1.3 に示す。

表 1.3 川端康成代筆疑惑作品の収録状況

カテゴリ	小説名	第 3 回全集	第 4 回全集
少女小説	乙女の港	未収録	収録
少女小説	花日記	未収録	収録
少女小説	コスモスの友	未収録	収録
睡眠薬中毒時期	古都	収録	収録
睡眠薬中毒時期	眠れる美女	収録	収録
その他	山の音	収録	収録

表 1.3 から分かるように、川端康成の代筆疑惑作品の中で、『乙女の港』、『花日記』と『コスモスの友』は第 4 回の全集に収録されている。本論文では、川端康成の代筆問題と文体問題を研究するにあたって可能なかぎり多くの作品を用いると同時に、川端康成本人が執筆したものを選ぶ必要もあるため、この二つの条件を満たした第 3 回の全集を用いることにした。また、文体の一致性を保つため、小説のみ用い、第 3 回全集の第 13 巻からの伝記や随筆を対象としない。文体の見分ける基準として、鳥羽・原 (1997)の『川端康成全作品研究事典』を参考にしたが、事典では小説と分類されたものの、筆者の判断で明らかに小説ではない作品(『船遊女』、『古里の音』、『末期の眼』、『文学的自叙伝』、『美しい日本と私』など)を研究対象から除外した。作品中の会話文は意図的に発話者の性格に合わせられ、著者の文体特徴が失われる可能性があるため、本論文では、すべての対象作品から会話文を削除して地の文だけを用いることにした。また、ほかの作品からそのまま取ってきた直接引用文と、作品中の明らかに小説と異なる文体で書いた文 (日記文や手紙など)も削除した。川端康成の掌の小説シリーズをはじめとする作品の会話文削除処理を行うと、残りの文字数はあまりにも短く、安定な統計量を得ることが困難なため、分析の対象から除いた。第 3 回の全集には旧字旧仮名で書かれた作品が多くあり、その影響で文体特徴量を正しく抽出できなくなるので、文化庁の内閣告示・訓令ページに掲載される「常用漢字表」、「現代仮名遣い」と「外来語の表記」などを参考にし、旧字旧仮名を新字新仮名に改めた。また、常体と敬体で書かれた文章の文体分析を行う場合、異なる文末表現の影響で結果は大きく変わる可能性があるため、敬体文が多く入った『青い海黒い海』、『父母への手紙』、『抒情歌』、『寝顔』、『北の海から』、『波千鳥』と『隅田川』も分析対象から外した。このような処理を施した川端康成の小説 90 編を本論文の川端康成コーパスとし、具体的な作品目録を付録 1 に示す。

1.4.2 対照作家コーパス

代筆問題、文体存在問題と語彙問題の研究には対照作家コーパスも必要である。少女小説では代筆者と思われる中里恒子の小説 20 編を選んだ。『古都』では澤野久雄、北條誠と三島由紀夫の小説それぞれ 20 編を選んだ。『眠れる美女』と『山の音』では三島由紀夫の小説 20 編を選んだ。文体存在問題と語彙問題の研究においては、先行研究を踏まえて泉鏡花、徳田秋聲と横光利一の小説をそれぞれ 20 編を選んだ。

1.5 本論文の構成

本論文では、文体計量分析に基づいて川端康成の代筆問題、文体存在問題と変化問題を明らかにする。12 章に分けて議論を進める。

第 1 章では、本論文の位置づけと目的を説明する。

第 2 章では、計量文体研究の先行研究をまとめ、本論文の計量文体学の方法を紹介する。

第 3 章では、少女小説の『乙女の港』の代筆疑惑検証を行う。

第 4 章では、少女小説の『花日記』の代筆疑惑検証を行う。

第 5 章では、少女小説の『コスモスの友』の代筆疑惑検証を行う。

第 6 章では、睡眠薬中毒時期の『古都』の代筆疑惑検証を行う。

第 7 章では、睡眠薬中毒時期の『眠れる美女』の代筆疑惑検証を行う。

第 8 章では、睡眠薬中毒時期の『山の音』の代筆疑惑検証を行う。

第 9 章では、川端康成の文体存在問題を扱う。

第 10 章では、川端康成の文体変化問題を扱う。

第 11 章では、川端康成の語彙問題を扱う。

第 12 章では、本論文の結論を述べる。

第 2 章 日本語計量文体研究の学史と研究方法

計量文体研究は、文章から抽出した文体の特徴を表すデータに対して統計処理を行う一連のプロセスを指す。統計的手法を用いた文体解析の可能性について、哲学者の梅原 (1985)は次のように述べている。

> 文体は思想の表現である。mなる文体をAの人が使うことは、その人間の内的思想がmなる文体によって表されることを意味している。したがって、文体を統計的手法によって研究することにより、その文章mの著者、およびそのできた年代をほぼ決定することが出来る。

梅原 (1985)は、文体の統計分析を通じて文章の著者特定が可能であることを示した。現代では、このような研究分野は著者識別 (authorship attribution)として知られている。本章では、まず、日本語における著者識別の学史を紹介し、次に、本研究の川端康成の文体問題と代筆問題研究に用いた文体特徴量と計量的手法を説明する。

2.1 日本語著者識別研究の学史

著者識別は文章から著者の特徴を推定する科学である (Juola, 2006; Stamatatos, 2009)。候補著者の数によって著者識別の問題は閉集合 (closed-set)問題と開集合 (open-set)問題に大別される。閉集合問題は次の二つの条件を満たす必要がある。一つ目は匿名文章の可能な著者リストが存在し、二つ目は匿名文章の真著者がその著者リストに含まれると仮定できる。開集合問題は可能な候補者のリストが存在せず、もしくは存在してもそのリストに真著者が含まれていない場合を指す。候補者リストには可能な候補者が 1 人しかない場合、著者識別問題は匿名の文章がこの候補者が書いたか否かを判別する問題となり、これは著者検証 (authorship verification)問題として知られている。

日本語における著者識別研究は、主に、古典文と現代文を研究対象としている。古典文では『源氏物語』の著者識別研究が広く知られている。『源氏物語』の伝えられてきたものは写本だけで、昔からその一部は原著者である紫式部以外の著者が書いたと疑われ、特に後半の 10 巻は紫式部の娘の代筆と言われている。『源氏物語』の文体問題を解明するために、安本 (1958)は長さ 1000 字の文章における文の長さ、名詞、助詞、助動詞などを文体特徴量として用い、『源氏物語』の後半の 10 巻の著者が紫式部である可能性が小さいと報告している。Tsuchiyama and Murakami (2013)、土山 (2016)は品詞構成比率、語の頻度と語の長さを用いて『源氏物語』の複数著者説について検証を行い、複数著者の存在は否定できないと結論づけた。代筆問題のほかに、村上・今西 (1999)は助動詞の出現率を用いて『源氏物語』の執筆順番を検討した。小野 (2015)はこの研究を発展させ、分散安定化変換を適用したデータにクラスター分析を行い、『源氏物語』の執筆順番を再確認した。

宗教関連の著作物の計量文体研究も行われていた。仏教思想家日蓮の著作には贋作がある言われ、村上・伊藤 (1991)は日蓮の著作 24 編、贋作と疑われたもの 16 編、日蓮門下の著作 5 編の文献を選んでコーパスを作成し、文の長さ、単語の長さ、品詞の出現率と語彙の豊富さ指標を用いて著者識別を行った。その結果、贋作と疑われた 5 編のうち 2 編は日蓮の著作、3 編は贋作という結論を得ている。

ほかの日本古典文の計量文体研究として、井原西鶴の遺稿集についての研究が挙げられる。江戸中期の俳人と浮世草子作家である井原西鶴の 5 編の遺稿集は弟子の北條団水による代筆という説があった。Uesaka and Murakami (2015)、上阪 (2016)は、品詞の構成比、単語の出現率と bi-gram の出現率を用いて代筆問題の検証を行い、遺稿集の文体は井原西鶴に似ていることから弟子による代筆の可能性は小さいと結論づけた。

小林・小木曽 (2013)は、中古和文コーパスから抽出した『源氏物語』、『紫式部日記』と『更級日記』における動詞と助動詞の使用傾向を調査し、クラスター分析を用いて考察を行った。その結果、中古和文において個人文体よりジャンル文体の文体差が大きいことが分かった。

現代文では、安本 (1959)は、文の長さ、名詞の使用頻度、比喩の使用頻度などの 12 項目を用いて日本現代作家の文章分類を試みた。樺島 (1954; 1955; 1963)は、日常会話、小説中の会話文、哲学書、小説中の地の文、自然科学書、和歌、俳句、新聞記事について分析し、名詞を説明変数とした場合のほかの品詞 (動詞、形容詞類、接続詞類)との関係性を示した。この関係性は「樺島法則」として知られている。

金 (2009)は、芥川龍之介の 309 編作品を用いて執筆時期の推定を行った。その結果、推定値と実際の執筆年度の誤差の標準偏差は 1.4 年で、比較的によい推測結果を得たと言える。尾城 (2016)は、太宰治の精神不安定から生じた文体の変化を探り、第二次世界大戦を境に一文に打つ読点の数が変化していると結論づけた。劉 (2016)は、「文学の鬼」と呼ばれた宇野浩二の文体について計量分析を行い、彼の「脳の大患」から文体が変化したことを明らかにした。

2.2 著者識別のための文体特徴量

文体特徴量 (stylometric feature)は文章に潜んでいる著者の特徴を反映する要素を指す。文章の構成単位は文、文節、単語と文字記号などがあり、このような構成単位から著者の特徴と思われるものを集計すると、文体特徴量が得られる。文体特徴量の抽出は著者識別研究における重要な一環で、抽出した文体特徴量にどれほど著者の情報が含まれるかが著者識別の正確性に直接影響を与える。著者識別のための文体特徴量は初期段階から絶えず研究が行われてきた。最初は取得しやすい語彙レベルの特徴量や文字、単語の n-gram などが多く提案されたが、自然言語処理技術の発達にともなって品詞情報や構文情報、ないし意味情報とまで特徴量の範囲が広がっている。日本語における早期の研究は色彩語や比喩などの文章表現上の特色に注目した (波多野, 1950)。韮沢 (1965)は、「にて」、「へ」などの語彙の比率を用いて『由良物語』の著者問題を論じた。1990 年代に入ってから日本語の自然言語処理技術の発達

で一連の日本語著者識別に有効な特徴量が提案された。本節では、著者識別研究のための文体特徴量を紹介する。

2.2.1 語彙特徴量

語彙特徴量は、主に、語彙の豊富さ特徴量と語彙の頻度特徴量に分けることができる。語彙の豊富さは文章著者の語彙の多様性を測る指標で、最もよく知られているのはタイプ・トークン比 (type-token ratio)である。Type を異なり語数 (同じ単語を 1 語として集計する)、Token を延べ語数 (単語の用いられた度数の総計)とする場合、タイプ・トークン比は Type/Token で計算できる。このような指標から語彙の豊富な著者と貧弱な著者を見分けることができる。しかし、語彙の豊富さ指標は文章の長さに依存し、文章が長くなるにつれて分母の Token は無限に増加し、分子の Type は著者の知っている語彙の範囲に収まるためタイプ・トークン比の値は徐々に小さくなる。文章の長さに依存しないように工夫された語彙の豊富さ指標も複数提案されていたが、完全に文の長さに依存しないものは存在しない。Grieve (2007)の 39 個の文体特徴量を用いた比較研究では、文体特徴量としての語彙の豊富さ指標はそれほど有効ではないことを示した。

語彙の豊富さのほかに、単語の使用頻度も文体特徴量として用いられ、bag-of-words として知られている。Bag-of-words は文中単語の出現頻度を並べてベクトルとして表現したものである。Bag-of-words の使用頻度の多い順からいくつかの単語を集計したものは最頻出単語特徴量で、この特徴量は抽出方法が簡単なため広く応用されている。Bag-of-words 特徴量の次元数が高く、このような高次元データの統計処理は困難であったが、データ解析技術の発達に伴い、特徴量に含まれる最頻出単語の数は 100 個程度から 1000 個になっても処理できるようになった (Burrows, 1992; Stamatatos, 2006)。Bag-of-words 特徴量には文体分析に用いるべきではない内容語も多く含まれている。そこで、内容語を除くために品詞ごとに語彙の頻度を集計し、機能語だけを著者識別に用いるようになった。機能語 (助詞、副詞など)は文章の文法機能を担い、どの著者でも大量に、無意識的に用いているため個人差が現れやすいとされている。文体特徴量としての機能語の数をいくつにするかが一つの問題で、これについての先行研究を表 2.1 にまとめる。

表 2.1 機能語を特徴量とした先行研究

言語	特徴量	語数	先行研究
英語	機能語	150	Abbasi and Chen (2005)
英語	機能語	303	Argamon, Saric, and Stein (2003)
英語	機能語	365	Zhao and Zobel (2005)
英語	機能語	480	Koppel and Schler (2003)
英語	機能語	675	Argamon, Whitelaw, Chase, Hota, Garg, and Levitan (2007)
日本語	助詞	24	金 (1997)
中国語	虚辞	47	李 (1987)
中国語	機能語	35	Yu (2012)

上述の bag-of-words 特徴量を集計する際に失われた文脈の情報は著者識別研究にしばしば必要である。そのために提案された文体特徴量は word n-grams である (Peng and Wang, 2014)。Word n-gram は単語間の文脈情報をある程度保つことができるが、文章の著者識別問題において必ずしも bag-of-words より優れているとは限らない。

2.2.2 文字特徴量

語彙をさらに細かく分けると文字になる。文字の n-gram も著者識別の分野では多く用いられている。この特徴量は自然言語処理のツールを頼らずに簡単に抽出でき、著者特徴の定量化に有効とされている (Grieve, 2007)。文字の n-gram では n の値を決めることが重要である。n の数が大きすぎると著者の特徴情報のほかに文章の内容情報も盛り込まれ、一方 n の値が小さすぎると文字特徴量は単語の一部となり、文脈の情報が失われてしまう。n の値の決定は言語と用いたコーパスに依存し、英語では 4-gram は最も有効とされ、日本語では bi-gram の有効性が示された (Sanderson and Guenter, 2006; 松浦・金田, 2000)。文字特徴量を用いた著者識別研究は数多く行われている。Kjell (1994)は、文字の bi-gram と tri-gram を用いて The Federalist Papers の著者識別を行った。Forsyth and Holems (1996)は、著者識別において文字の n-gram は語彙特徴量より有効であることを示した。Hoornet et al. (1999)は、文字の tri-gram を用いて著者識別を試みた。また、複数文体特徴量の比較研究では文字の n-gram は最も性能がよい特徴量となっている (Grieve, 2007)。

2.2.3 品詞特徴量

品詞は著者識別研究でよく用いられる文体特徴量である。この特徴量は文章の文法機能を担っている。日本語の品詞特徴量を抽出するために文章を形態素ごとに分割することと、各形態素に品詞情報を付与することが必要で、この一連のプロセスは形態素解析という。品詞の比率を用いた早期の研究として、安本 (1958)は、1000 字ごとの名詞、助詞と助動詞の出現回数を用いて『源氏物語』の著者問題を研究した。また、樺島・寿岳 (1965)は、作家 100 人の作品における品詞の比率の分析を行った。Koppel and Schler (2003)は、頻出語彙の品詞情報に着目し、コーパスの中で 3 回以上現れた品詞 bi-gram を文体特徴量として用いた。Gamon (2004)は、819 個の品詞 tri-gram を用いて著者識別を行い、この特徴量は機能語の出現頻度より性能が良いと示した。Zhao and Zobel (2007)は、55 人の著者が書いた 634 作品における著者識別を行い、品詞 bi-gram は unigram よりよい正解率を得ている。金 (2014)は、日本語の小説、作文と日記コーパスにおける品詞 bi-gram の有効性を実証した。また、特定の品詞を用いた著者識別の研究も行われている。金 (2002)は、日本語の中で出現率が最も高い品詞である助詞の n-gram が著者識別に有効であることを示した。

2.2.4 構文特徴量

構文は文の構造を指し、文章の著者によって構文の複雑度も異なる。構文情報を文体特徴量として用いた早期の研究として Bayyen et al. (1996)が挙げられる。彼らは英語コーパスに対し文ごとの構文木を作り、この構文木に基づいて構文特徴量を抽出した。また、Samatatos et

al. (2006)は、自然言語処理ツールを用い、現代ギリシア語の文を幾つかのチャンクに分割して構文情報の抽出を試みた。日本語において、金 (2013)は、文節パターン特徴量を提案し、日本語の小説、作文と日記の著者識別ではこの特徴量の有効性示した。また、韓国語において、Lee et al. (2017)は構文特徴量の語節パターンの有効性を確認した。

2.3 著者識別方法

著者識別モデルの研究は文体特徴量の抽出より少し遅れを取ったが、統計学と機械学習の発達に伴い発展が進んできている。その発展段階をたどると、主に、記述・推測統計学、多変量解析と機械学習の手法が挙げられる。

2.3.1 記述・推測統計学

記述統計学を用いた著者識別の研究の先駆者は Mendehall である。Menenhall (1887)は、単語の長さの単純集計を行い、シェークスピアの文章には長さ 4 文字の単語が最も多いのに対して、ベーコンの文章には長さ 3 文字の単語が最も多いことを示した。Yule (1938)は、文体の計量の研究に文の長さの平均値、中央値と四分位数などの統計量を文体特徴量として用いた。

記述統計学のほかに推測統計学も早期の著者識別研究に適用されていた。アメリカの名作家、Mark Twain が南北戦争に関与していると言われ、その決定的な証拠は 1861 年の New Orleans' Daily Crescent に刊行された 10 通の手紙である。この手紙には、Quintus Curtius Snodgrass の署名があるにも関わらず、実は、Mark Twain が書いたと疑われていた。Bringar (1963)は、平均の差の検定とカイ二乗検定を用いて Mark Twain の作品を分析し、Quintus Curtius Snodgrass は Mark Twain とは異なる人物と結論づけた。

2.3.2 多変量解析

記述・推測統計学の扱っている変数は限られているため、結果に偏りが生じる可能性も大きい。現在著者識別の分野では、数多くの変数を同時に解析する多変量解析の手法を用いるのが一般的である。著者識別に用いられた主な多変量解析の手法として、主成分分析、因子分析、対応分析、多次元尺度法、クラスター分析、ニューラルネットワークなどが挙げられる。特に、主成分分析 (PCA)は著者問題の研究に応用されることが多い。主成分分析は、高次元データをできる限り情報の損失なしに 2 次元平面に射影し、2 次元平面のプロットでデータの関連性を考察する手法である。主成分分析を適用した早期の研究として Burrows (1987)がある。因子分析 (FA)は、変数の間の相関関係から共通因子を求め、その共通因子に基づいてデータを説明する手法である。安本 (2009)は、因子分析を 100 人の作家の作品に適用し、作品は大きく 8 つのグループに分類できることを示した。対応分析 (CA)は分割表の行の項目と列の項目の相関が最大になるように、関連性が強いものが近づくように解析する手法である (金, 2016)。Zaitsu and Jin (2015)は、対応分析を「グリコ事件」の犯罪者が書いた恐喝状の著者識別に適用した。劉 (2016)は、対応分析を用いて、宇野浩二の文体変化を分析した。多

次元尺度法 (MDS)は、データ間の位置関係を保ちながら高次元データを低次元空間 (2, 3 次元が多い)に示す手法で、主成分と同じく次元削減の手法の一つである。Aljumily (2015)は、多次元尺度法を用いてシェークスピアの作品の代筆問題の解明を試みた。クラスター分析は同じ特徴を持つデータをグルーピングする手法である。Eder (2015)は、文体特徴量の可視化研究に階層的クラスター分析を用いた。Uesaka and Murakami (2015)は、井原西鶴の遺作の著者識別問題に階層的クラスター分析を適用し、遺作は弟子の代筆ではないことを明らかにした。Sun and Jin (2017)は、川端康成の名著『山の音』の代筆問題に階層的クラスター分析を導入し、『山の音』の三島由紀夫代筆説を否定した。ニューラルネットワークは人間の脳にある神経細胞の構造をモデル化したものであり、自己組織化マップ (SOM)は教師なしニューラルネットワークの代表例である。自己組織化マップには入力層と出力層があり、入力データに一番近いものを勝者とし、その勝者の周辺にあるニューロンを勝者に近づくように調整しながら分類を行う。金 (2003)は、日本語の著者識別では自己組織化マップの結果が主成分分析、対応分析と階層的クラスター分析より優れていると報告した。

2.3.3 機械学習

多変量解析 (教師なしの手法)のほかに、機械学習 (教師ありの手法)も著者識別に用いられ、特に、分類器 (classifier)の応用は急速に広まっている。著者識別に用いた主な分類器はナイーブベーズ (Naïve Bayes)、k 近隣法 (k-NN)、サポートベクターマシン (SVM)とランダムフォレスト (RF)などがある。機械学習方法の著者識別研究への早期の応用は Mosteller and Wallace (1964)の The Federalise Papers に対する研究である。この研究では、20 種類の単語を特徴量とし、ナイーブベース (Naïve bayes)法を用いて著者識別を行った。Peng et al. (2003)は、tri-gram とナイーブベース (Naïve bayes)分類器を用いてギリシア語文章の著者識別を行った。先行研究の 72%の精度に対し、ナイーブベース分類器では 90%の精度を得ている。Hoorn et al. (1999) は tri-gram とニューラルネットワーク、ナイーブベーズと k 近隣法を用いて詩の分類を試みた。2 群の場合 80%~90%、3 群の場合 70%前後の精度で判別できている。SVM は高次元データ解析に相応しく、著者識別研究に長年用いられてきた分類器である (De Vel et al., 2001; Zheng et al., 2006)。金 (2007)の比較研究では、RF の性能は SVM より優れていることが分かり、RF 法も著者識別に多く用いられるようになった。Tabata (2012)は、RF 法を用いて、Dicken の小説の文体を分析した。孫他 (2015a; 2015b; 2015c)は、RF 法を含めた分類器を用いて、川端康成少女小説の代筆問題解明を試みた。

2.4 本論文で用いた文体特徴量と識別方法

著者識別のためのプロセスは主に二つある。一つは川端康成、代筆者と代筆疑惑作品からの文体特徴量抽出で、もう一つは著者識別の方法適用と結果解釈である。この一連のプロセスを図 2.1 に示す。

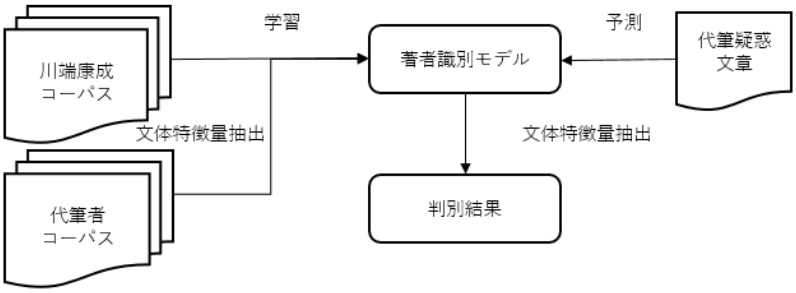

図 2.1 著者識別のプロセス

本論文では、文字記号 bi-gram、タグ付き形態素と文節パターンを文体特徴量として用いることにした。本節では、この文体特徴量及び抽出方法を紹介する。

2.4.1 文字記号 bi-gram

文字 n-gram は自然言語処理の分野で多く用いられているモデルである。日本語では、松浦・金田 (2000)が初めて文字 bi-gram の有効性を実証した。また、文字だけでなく、文章中の記号も有効な文体特徴量とされている (Grieve, 2007)。そこで、金 (2014)の研究では、文字記号 bi-gram を一つの特徴量として用い、その有効性を示した。本論文では、先行研究を踏まえて文字記号 bi-gram を文体特徴量として用いることにした。

本論文で用いた文字記号bi-gramは日本語文の文字、仮名と記号の隣接しているペアを指す。例えば、川端康成の小説『ほくろの手紙』の最初の一文、「あの黒子の、面白い夢を、わたくし昨夜見ました。」から文字記号bi-gramを取る場合、「あの」、「の黒」、「黒子」、「子の」、「の、」、「、面」、「面白」、「白い」、「い夢」、「夢を」、「を、」、「、わ」、「わた」、「たく」、「くし」、「し昨」、「昨夜」、「夜見」、「見ま」、「まし」、「した」、「た。」の22個の文字記号bi-gramが得られる。

2.4.2 タグつき形態素

タグ付き形態素は、形態素とその形態素に付くタグの組合せを指す。形態素は意味を持つ最小単位で、本論文では、形態素の品詞情報をタグと呼ぶ。日本語の文章からタグ付き形態素を抽出するために形態素解析が必要で、本論文では、形態素解析器MeCab (IPA辞書)を用いて形態素解析を行った。「あの黒子の、面白い夢を、わたくし昨夜見ました。」の一文の形態素解析結果を次に示す。解析結果には名詞、動詞と形容詞も含まれているが、このような内容語は著者識別に適切ではないと知られ、本論文では、タグ付き形態素から名詞、動詞、形容詞を含むものを除いた。

最初に現れた「黒子」や「の」などは文章の形態素である。形態素の右側にあるものは形態素の品詞情報を表すもので、すなわち形態素タグである。形態素タグはいくつかの層に分かれ、右側に行くほど形態素の細かい情報が表示されている。本論文では、内容語を除いた第1層の形態素タグの情報を用いる。この一文から「あの_連体詞」、「の_助詞」、「、_記号」

(2回)、「を_助詞」、「まし_助動詞」、「た_助動詞」、「。_記号」の8個のタグつき形態素が得られる。

```
あの　　連体詞、*、*、*、*、*、あの、アノ、アノ
黒子　　名詞、一般、*、*、*、*、黒子、クロコ、クロコ
の　　助詞、連体化、*、*、*、*、の、ノ、ノ
、　　記号、読点、*、*、*、*、、、、、、、
面白い　形容詞、自立、*、*、形容詞・アウオ段、基本形、面白い、オモシロイ、オモシロイ
夢　　名詞、一般、*、*、*、*、夢、ユメ、ユメ
を　　助詞、格助詞、一般、*、*、*、を、ヲ、ヲ
、　　記号、読点、*、*、*、*、、、、、、、
わたくし　　名詞、代名詞、一般、*、*、*、わたくし、ワタクシ、ワタクシ
昨夜　　名詞、副詞可能、*、*、*、*、昨夜、サクヤ、サクヤ
見　　動詞、自立、*、*、一段、連用形、見る、ミ、ミ
まし　助動詞、*、*、*、特殊・マス、連用形、ます、マシ、マシ
た　　助動詞、*、*、*、特殊・タ、基本形、た、タ、タ
。　　記号、句点、*、*、*、*、．、．、．、．
```

2.4.3 文節パターン

大辞林 (第三版)では、文節は「日本語の言語単位の一。文を実際の言語として不自然でない程度に区切ったときに得られる最小単位。」と定義されている (松村, 2006)。日本語を文節に分割するツールとして、係り受け解析器CaboChaがある。上述例文のCaboChaによる解析結果を次に示す。

```
* 0 1D 0/0 0.971456
あの　　連体詞、*、*、*、*、*、あの、アノ、アノ
1 3D 0/1 0.213642
黒子　　名詞、一般、*、*、*、*、黒子、クロコ、クロコ
の　　助詞、連体化、*、*、*、*、の、ノ、ノ
、　　記号、読点、*、*、*、*、、、、、、、
* 2 3D 0/0 2.034163
面白い　形容詞、自立、*、*、形容詞・アウオ段、基本形、面白い、オモシロイ、オモシロイ
* 3 6D 0/1 -2.363218
夢　　名詞、一般、*、*、*、*、夢、ユメ、ユメ
を　　助詞、格助詞、一般、*、*、*、を、ヲ、ヲ
、　　記号、読点、*、*、*、*、、、、、、、
* 4 5D 0/0 0.069551
わたくし　　名詞、代名詞、一般、*、*、*、わたくし、ワタクシ、ワタクシ
* 5 6D 0/0 -2.363218
昨夜　　名詞、副詞可能、*、*、*、*、昨夜、サクヤ、サクヤ
* 6 -1D 0/2 0.000000
見　　動詞、自立、*、*、一段、連用形、見る、ミ、ミ
まし　助動詞、*、*、*、特殊・マス、連用形、ます、マシ、マシ
た　　助動詞、*、*、*、特殊・タ、基本形、た、タ、タ
。　　記号、句点、*、*、*、*、．、．、．、．
```

CaboChaを用いた解析結果では、米印「*」は文節の始まりを示し、直後の数字は文節の番号であり、その次の数字はこの文節がかかる文節の番号である。金 (2013)は、4種類の文節パターンを提案したが、本研究は文節内の助詞・記号を除いた形態素の第1層品詞情報と助詞、記号の原型を組み合せた文節パターンを用いた。「あの黒子の、面白い夢を、わたくし昨夜見ました。」の1文における第2文節は「黒子_の_、」である。この文節には助詞「の」と記号「、」が含まれるため、第2文節の文節パターンは「名詞_の_、」になる。他の文節も同じ考え方に基づいて抽出すると、この1文に含まれた全ての文節パターンは「連体詞」、「名詞_の_、」、「形容詞」、「名詞_を_、」、「名詞」 (2回)、「動詞_助動詞_助動詞_。」の7個である。

2.4.4 本論文で用いた分析手法

本論文では、教師なし学習法の対応分析と階層的クラスター分析、教師あり学習法のエイダブースト (Adaptive Boosting: AdaBoost)、高次元判別分析 (High-Dimensional Discriminant Analysis: HDDA)、ロジスティックモデルツリー (Logistic Model Tree: LMT)、ランダムフォレスト (Random Forest: RF)とサポートベクターマシン (Support Vector Machine: SVM)を分析手法として用いた。

対応分析の分析対象はカテゴリカルデータである。この手法は高次元データを低次元 (2~3次元が多い)に射影し、低次元上の散布図を用いて個体と変数間の関係を考察する手法である(金, 2016)。階層的クラスター分析は、個体間の類似度または非類似度 (距離)に基づいてデータの構造が似ている個体を同じグループにまとめる分類の方法である。

階層的クラスター分析は、まず、元データから距離行列を作り、距離の近い個体またはクラスターから併合してデータのクラスターリングを行い、デンドログラムという樹形図で分類結果を示す。クラスターリングを行うにあたってクラスターの併合方法と距離を事前に決めておく必要がある。本論文では、著者識別の先行研究を踏まえてウォード法 (ward's method)と KLD 距離を用いることにした。ウォード法は、クラスターを結合する際にグループの分散に対するグループ間の分散を最大にする方法である。KLD 距離の式を 2.1 に示す。

$$\mathrm{KLD} = \sqrt{\frac{1}{2}\sum \left(x_i \log \frac{2x_i}{x_i + y_i} + y_i \log \frac{2y_i}{x_i + y_i}\right)} \tag{2.1}$$

本論文で用いた文章の長さはそれぞれ異なり、長い文章から抽出した特徴量の数は短い文章から抽出したものよりあきらかに多いため、同時に統計処理できない。このような文章の長さの影響を除くために、集計した度数f_{ij}を相対度数x_{ij}に置き換えた。変換に用いた式を2.2に示す。

$$x_{ij} = \frac{f_{ij}}{\sum_{j=1}^{n} f_{ij}} \tag{2.2}$$

対応分析とクラスター分析のほか、本論文では、いくつかの機械学習の分類器を用いた。Manuel et al. (2014)は、179個の分類器についてベンチマークUCIデータセットを用いて性能の比較分析を行い、RFとSVMが高性能であることを示した。金・村上 (2007)は、日本語著者識別におけるRFの有効性を実証した。金 (2014)は、日本語の文学作品、作文と日記に対して複数の文体特徴量及びRFとSVMを含む分類器を用いて著者識別を行った。本論文では、Manuel et al. (2014)、金・村上 (2007)、金 (2014)の結果を踏まえ、精度が高く、高次元データ解析に適する次の5つの分類器を用いた。

⑴ エイダブースト

エイダブーストはFreund and Schapire (1996)により提案されたアンサンブル学習法による強分類器である。AdaBoostは前の分類器の誤り情報を用いて次の分類器の精度を上げるように工夫し、分類器を繰り返し作成して強分類器を構築する方法である。

⑵ 高次元判別分析

高次元判別分析は Bouveyron et al. (2007)が提案した高次元判別分析方法で、各クラスにおける高次元を独立に次元縮小するアイディアに基づく高次元データにふさわしい分類器である。文学作品の分類では HDDA は SVM とほぼ同様な性能を示している (金, 2014)。

⑶ ロジスティックモデルツリー

ロジスティックモデルツリーはLandwehr et al. (2005)が提案し、決定木の葉にロジステックモデルを適応した分類器である。著者識別においても高い識別率を得る場合がある。

⑷ ランダムフォレスト

ランダムフォレストはBreiman (2001)が提案し、アンサンブル学習法バギング (bagging)をさらに発展させた分類器である。この手法はブートストラップサンプリングしたデータから作った決定木の結果を統合して分類を行う。分類問題において最良な手法とされている (Manuel et al., 2014)。

⑸ サポートベクターマシン

サポートベクターマシンはVapnik (1998)が提案し、伝統的な線形判別の境界について、マージンを最大化する方法で求める分類器である。分類問題におけるSVMは
RFとほぼ同等の性能を示している (Manuel et al., 2014; 金, 2014)。

本論文で用いた各文体特徴量における分類器の性能を評価するために、一個抜き交差検証(LOOCV)を行い、表2.2に示した混同行列を作る。表2.2のTPの正解は川端康成の文章を正しく判別した回数である。FPは川端康成の文章を間違って代筆者に判別した回数である。FNは代筆者の文章を間違って川端康成に判別した回数である。TNは代筆者の文章を正しく判別した

回数である。分類器の精度を式2.3~2.5に示した適合率 (Precision)、再現率 (Recall)とF-尺度(F-measure)を用いて評価する。

表2.2 混同行列表

	正解	
予測結果	TP	FP
	FN	TN

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.4}$$

$$\text{F} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{2.5}$$

2.5 語彙の豊富さ関連指標と計量方法

本論文の川端康成における語彙課題は川端康成の語彙の豊富さ問題と平仮名多用問題である。本節では、このような問題を解決するための計量的指標を示す。

2.5.1 語彙の豊富さ

語彙の豊富さとして最もよく知られているのはタイプ・トークン比 (TTR)である。タイプ・トークン比は式 2.6 で計算できる。

$$\text{TTR} = \frac{\text{Type}}{\text{Token}} \tag{2.6}$$

TTR の他にも語彙の豊富さを測る指標は複数提案されていたが、このような指標は文の長さに依存することが指摘されている。文長の影響をそれほど受けていない指標はs値であり、本論文ではこの s 値を語彙の豊富さを測る指標として用いる。s 値は式 2.7 で計算できる。

$$\text{s} = \frac{\log(\log(\text{Type}))}{\log(\log(\text{Token}))} \tag{2.7}$$

2.5.2 平仮名使用率

日本語文には漢字 (C)、平仮名 (H)と片仮名 (K)が含まれ、著者によっては漢字の代わりに平仮名を用いる場合もある。本論文では、川端康成の平仮名多用問題を計量するために、文章における平仮名使用率を統計量として用いる。平仮名使用率を求める式を 2.8 に示す。

$$\text{Hirakana Ratio} = \frac{\text{H}}{\text{C} + \text{H} + \text{K}} \tag{2.8}$$

2.5.3 一元配置分散分析

本論文では、川端康成の作品と対照作家の文体指標を比較するために一元配置分散分析を行った。一元配置分散分析は複数群の母平均間に統計的有意差があるかを示す方法で、その手順を次に示す。

(1) 帰無仮説 (H0):「すべての群の母平均が等しい」を真とする。
　対立仮説 (H1):「各群の母平均が等しくない」。
(2) 検定統計量 F の値を 2.9 の式で計算する。

$$\text{F} = \frac{\text{群間平方和/群間の自由度}}{\text{群内平方和/群内の自由度}} \tag{2.9}$$

(3) 有意水準を決め、F 統計量は H0 が真であるときに起きにくい値かを判断する。
(4) 仮説の採択と棄却を決める。

平均の差の検定結果は標本サイズに依存し、その検定統計量の p 値の有効性を示す指標は効果量である。分散分析の効果量として相関比 (η)を用い，η^2 で効果量が表されることが多い (金, 2016)。その計算式を 2.10 に示す。

$$\eta^2 = \frac{\text{ある要因の平方和}}{\text{全体の平方和}} \tag{2.10}$$

一元配置分散分析の帰無仮説 H0 は、「すべての群の母平均が等しい」である。この仮説が棄却されると「すべての群の母平均は等しくない」が言えるだけで、具体的にどの二群間に差があるとまで言えない。そこで各群の間の差を見るために多重比較を行う。多重比較にはいくつかの方法があり、本論文では、最も一般的な Turkey 法を用いることにした。

第 3 章 『乙女の港』の代筆問題研究

川端康成の小説には、若者向けの「少年少女小説」が重要な位置を占め、しかも名作とされたものが多い。川端康成の少女小説は、主に、昭和の初めからの 10 年間に発表されている (小林, 1982)。川端康成の名義で発表された少女小説のうち、代筆と疑われたものは『歌劇学校』、『万葉姉妹』、『花と小鈴』、『親友』、『長い旅』、『乙女の港』、『花日記』と『コスモスの友』などがある。これらの代筆疑惑が持たれた少女小説は、代筆事実の有無で大きく 2 種類に分けることができる。『歌劇学校』、『万葉姉妹』、『花と小鈴』、『親友』と『長い旅』は代筆の事実があると言われ、既に川端康成 1984 年に完結した最新の全集から削除されている (小谷野, 2013)。『乙女の港』、『花日記』と『コスモスの友』は今でも最新の全集に収録されている。

第 3 章では、『乙女の港』を研究対象とする。『乙女の港』は中里恒子の代筆と疑われている(小谷野, 2013)。中里恒子 (1909〜1987)は 1928 年にデビューし、1939 年『乗合馬車』で芥川賞、1974 年『歌枕』で読売文学賞、1975 年『わが庵』で日本芸術院恩賜賞、1979 年『誰袖草』で女流文学賞をそれぞれ受賞した女性作家である。中里恒子は一時的に川端康成に師事し、『乙女の港』を執筆した際に川端康成の指導を受けていた。そのため、この作品は川端康成が中里恒子の書いた草稿に手を加えて完成させたとされている (小谷野, 2013)。

本章では、2.4 節で紹介した著者識別 (authorship attribution)の手法を用いて『乙女の港』の代筆問題を明らかにする。そのため川端康成と中里恒子のコーパスを作成する必要がある。本章では、研究対象の『乙女の港』と同じジャンル、また、なるべく創作期間が近い作品を選んでコーパスを作成した。ジャンルの見分けに関しては、『川端康成全作品研究事典』と論文「中里恒子著作目録－『まりあんぬもの』の可能性－」を参照した (原・羽鳥, 1998; 小関, 2012)。川端康成と中里恒子のコーパスをそれぞれ表 3.1 と 3.2 に示す。

3.1 研究背景

『乙女の港』は 1937 年 6 月から 1938 年 3 月にかけて 10 回に渡って『少女の友』に連載され、横浜のミッション系の女学校における女学生の間のエス (擬似的な姉妹となって交際する行為)を描写している小説である。『乙女の港』は 10 章からなり、1984 年に完結した川端康成全集の第 20 巻に収録されている。乙女の港の各章の詳細情報を表 3.3 に示す。

表 3.1 川端康成作品コーパス

発表時期	作品	文字数
1925 年	白い満月	11889
1926 年	伊豆の踊子	13281
1926 年	文科大学挿話	7577
1927 年	春景色	6360
1928 年	死者の書	4981
1930 年	温泉宿	18578
1931 年	落葉	9260
1933 年	二十歳	15116
1933 年	禽獣	11847
1936 年	雪国	48804
1936 年	夕映少女	6729
1939 年	高原	32728
1939 年	故人の園	6240
1940 年	燕の童女	5297
1940 年	女の夢	5699
1940 年	婦唱夫和	5326
1940 年	夜のさいころ	7577
1946 年	再会	9273
1947 年	夢	3806
1949 年	雨の日	4152

表 3.2 中里恒子作品コーパス

発表時期	作品	文字数
1932 年	泡沫	9614
1932 年	露路	8165
1933 年	ますく	6580
1936 年	自由画	6999
1936 年	祝福	6287
1937 年	ふみむすびと	9928
1937 年	毛皮	5340
1937 年	花火	2145
1937 年	樹下	5360
1938 年	森の中	5885
1939 年	野薔薇	9194
1939 年	乗合馬車	21260
1939 年	日光室	11479
1940 年	天国	4817
1940 年	晩餐会	2286
1940 年	後の月	16943
1940 年	孔雀	9973
1941 年	老嬢	10913
1941 年	競馬場へいく道	6690
1941 年	向日葵	2477

表 3.3 『乙女の港』の各章の詳細情報

『乙女の港』の各章		発行時間	雑誌	文字数
1	花選び	1936 年 6 月	少女の友	4761
2	牧場と赤屋敷	1936 年 7 月	少女の友	5320
3	開かぬ門	1936 年 8 月	少女の友	5047
4	銀色の校門	1936 年 9 月	少女の友	2470
5	高原	1936 年 10 月	少女の友	5280
6	秋風	1936 年 11 月	少女の友	5085
7	新しい家	1936 年 12 月	少女の友	5104
8	浮雲	1937 年 1 月	少女の友	5503
9	赤十字	1937 年 2 月	少女の友	6209
10	船出の春	1937 年 3 月	少女の友	5692

『乙女の港』の中里恒子による代筆疑惑の有力な証拠は川端康成と中里恒子との往復書簡である。この往復書簡は 1984 年に完結した『川端康成全集』の補巻二に収録されている。『乙女の港』の代筆についての内容を次に示す。

1937年 (昭和12年)9月14日付、川端康成から中里恒子へ (川端康成全集補巻二, p. 300)。

乙女の港はだんだん文章が粗くなり、書き直すのがむつかしく、書き直すといふことは、うまく参りませんゆゑ、なるべく初めの調子でやつていただくと助かります。お書きになるのにもし興が薄れてゆくやうでしたら、早く切り上げ、別のものをまた連載するやうにしても、こちらは結構ですが、受けてゐる様子ゆゑ、なるべく続けていただきたいと思つて居ります。三千子は港に帰つて、洋子の心の戻るのに少し曲折あり、この三角関係少しモメタ方が、つなぎやすいかと思ひますがいかがですか。克子の天下あつてもよいかと思ひます。

この書簡は、『乙女の港』の執筆指導を行うために川端康成が中里恒子宛に送ったものである。川端康成は『乙女の港』の添削が難しくなってきたことを示したほか、小説は受けているからなるべく書き続けてほしいとも述べた。この書簡から中里恒子は『乙女の港』の執筆に関わっていることがあきらかである。

1937年 (昭和12年)9月18日付、中里恒子から川端康成へ (川端康成全集補巻二, p. 292)。

乙女の港お言ば通り注意いたしませう。どんな風に書いても、うまくなほして下さる。こんなわがままな考へ方が私にあるからかもしれません。一回分終り、二回めの十枚まですすみましたがお手紙拝見してなほすつもりになりました。廿二日頃までーもし間にあはねば一回分だけお送りいたします。

この書簡は中里恒子から川端康成への返信である。「乙女の港お言ば通り注意いたしませう。どんな風に書いても、うまくなほして下さる。」の一文から、川端康成は既に『乙女の港』に手を加えたと推察される。

1937年 (昭和12年)10月16日付、川端康成から中里恒子へ (川端康成全集補巻二, p. 302)。

軽井沢が二度続き、話の進みもヤマも前と余り変りませんので、少し工夫して、大分書き変えました。戦争は入れないこととし、戦前のつもりにしたいと思ひますがいかがですか。最初のやうな調子でなるべく願ひます。

この書簡は川端康成から中里恒子宛の書簡で、「少し工夫して、大分書き変えました」の記述からも、川端康成が実際に中里恒子の文章に加筆した事実が浮き彫りになる。

以上の書簡の内容を辿っていくと、1通目は『乙女の港』の執筆における大まかな方向性についての執筆指導である。2通目は川端康成の添削に対する感謝である。3通目は川端康成か

らの添削結果の連絡と執筆指導である。この3通の書簡を貫いた主題は『乙女の港』の執筆指導である。

このような書簡を巡って数多くの先行研究がなされていたが、先行研究では、「川端康成作」、「中里恒子作」と「中立」で意見が分かれている。先行研究の諸説を表 3.4 にまとめる。

表 3.4 『乙女の港』の先行研究

先行研究	作品の帰属
内田静枝 (2009)	川端康成
大森郁之助 (1991)	川端康成
馬場重行 (1981)	川端康成
小谷野敦 (2013)	中里恒子
下條正純 (2009)	中立
中嶋展子 (2010)	中立

表 3.4 に示したように、内田 (2009)は、川端康成は中里恒子の下書きに徹底的に手を加えたことから『乙女の港』は川端康成の作品であると主張した。大森 (1991)は、同性愛のアプローチから考察を行い、『乙女の港』の同性愛モチーフは川端康成の発案だったと述べ、その同性愛の完成度から『乙女の港』は川端康成に近い可能性が高いと主張した。馬場 (1981)は、『乙女の港』の第 6 章は川端康成の加筆要素が入っていることから、作品全体は川端康成作と主張した。以上の『乙女の港』は川端康成に近い作品という意見に対し、小谷野 (2013)は、「私見では、川端は文章を直しただけで、筋は中里のものである」と述べ、『乙女の港』は中里恒子の作品であると主張した。その他の先行研究は中立的な立場を示した。下條 (2009)は、中里恒子が横浜市にあるミッションスクールである横浜紅蘭女学校の卒業生であることから、このような女の子の間の擬似的な交際をテーマとして取り上げたと推測していたが、往復書簡だけではどの程度が中里恒子の下書きであるかが不明であると述べた。中嶋 (2010)は、『乙女の港』は川端康成の加筆により文章表現が改善され、「広がり彩り」が添えられたと述べた。

以上の先行研究から、『乙女の港』の代筆問題について専門家の中で意見が分かれている。そこで、本章では、文体計量分析のアプローチから『乙女の港』の代筆問題を明らかにする。そのために用いた文体特徴量は文字記号 bi-gram、タグつき形態素と文節パターンで、分析手法は対応分析、階層的クラスター分析、AdaBoost、HDDA、LMT、RF と SVM である。

3.2 対応分析の結果

本節では、文字記号 bi-gram、タグ付き形態素と文節パターンの対応分析の結果を紹介する。対応分析のグラフには、各作品グループに描いた楕円は多次元 t 分布に基づいた 95%の信頼楕円である。『乙女の港』のコーパスから抽出した文字記号 bi-gram の次元数は 2117 である。

この値は、出現頻度はより少ない変数を一括した結果である。本研究のデータの次元数についての情報は付録 B にまとめた。文字記号 bi-gram 対応分析個体の第 1、2 スコアの散布図を図 3.1 に示す。図 3.1 では、川端康成の作品は第 2 象限にプロットされた。中里恒子作品の大多数は第 1 象限にプロットされた。『乙女の港』の各章は両作家の作品グループから離れ、第 3 象限にプロットされた。文字記号 bi-gram を用いた場合、『乙女の港』の文体は川端康成と中里恒子の文体とは異なることが見て取れた。

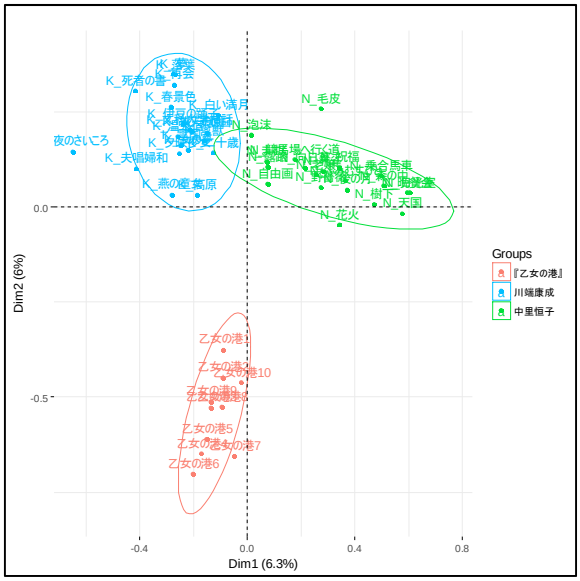

図 3.1 『乙女の港』文字記号 bi-gram 対応分析個体の第 1、2 スコアの散布図

川端康成、中里恒子の小説と『乙女の港』はどのような文体特徴を持つかを明らかするために、各グループに寄与する変数の分析を行った。変数の数が膨大であるため、本研究では、各文体特徴量における対応分析変数の第 1 スコアと第 2 スコアからそれぞれ 10 個を抽出して考察することにした。

『乙女の港』の文字記号 bi-gram 変数の第 1 スコアの棒グラフを図 3.2 に示す。第 1 スコアの大きい順で「園子」、「郁子」、「ヌの」、「ンヌ」、「アン」、「菊代」、「森之」、「之助」、「ヤは」と「デリ」の変数が現れ、中里恒子の作品では、このような文字記号 bi-gram が多く用いられている。第 1 スコアが負の方向には「ち子」、「水田」、「みち」、「。水」、「延子」、「、延」、「牧山」、「。み」、「山は」と「、水」の変数となり、川端康成の作品と『乙女の港』ではこういった文字記号 bi-gram を多く用いられている。特徴分析では、多くの人名についての語彙が現れたため、文字記号 bi-gram を用いた分析を行う場合内容語の影響を受けていることが分かった。

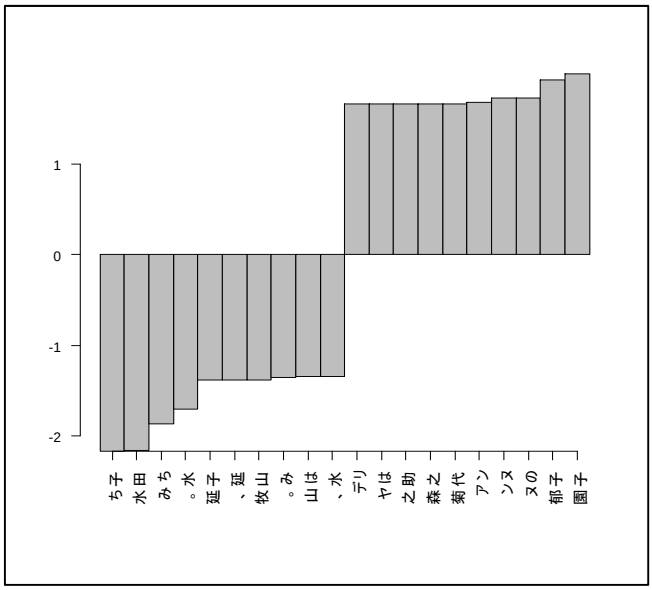

図 3.2 『乙女の港』文字記号 bi-gram 変数の第 1 スコアの棒グラフ

『乙女の港』の文字記号 bi-gram 変数の第 2 スコアの棒グラフを図 3.3 に示す。第 2 スコアが正の方向には「祐三」、「。祐」、「士子」、「踊の」、「富士」、「彼を」、「に彼」、「代子」、「。夫」と「彼に」の変数が現れ、川端康成と中里恒子の作品では、このような文字記号 bi-gram を多く用いられている。第 2 スコアの負の方向には「。克」、「、克」、「克子」、「お姉」、「千子」、「三千」、「…。」、「洋子」、「、三」と「—。」が現れ、『乙女の港』ではこのような文字記号 bi-gram を多く用いられている。

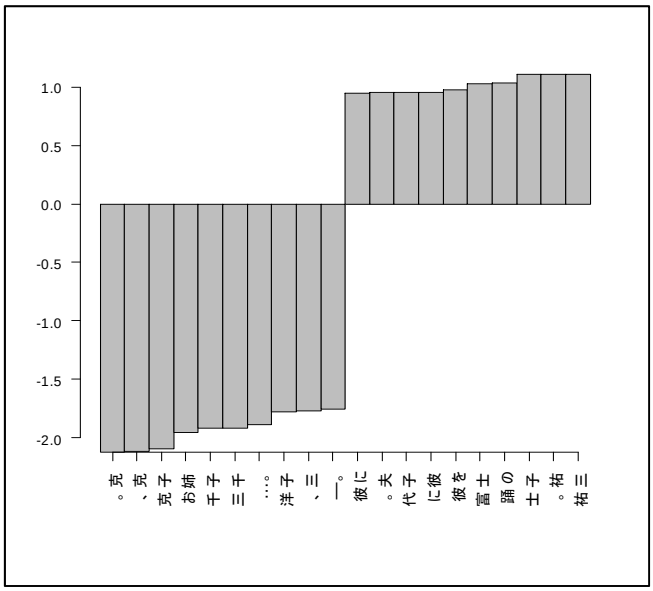

図 3.3 『乙女の港』文字記号 bi-gram 変数の第 2 スコアの棒グラフ

『乙女の港』のコーパスから抽出したタグ付き形態素の次元数は 218 である。タグ付き形態素対応分析個体の第 1、2 スコアの散布図を図 3.4 に示す。図 3.4 に示したように、川端康成の作品は第 1 スコア軸を跨ぎ、第 2、3 象限にプロットされた。中里恒子の文章は第 1 象限にプロットされた。『乙女の港』の各章は両者と離れた第 4 象限にプロットされた。タグ付き形態素を用いた場合、『乙女の港』の文体は川端康成と中里恒子の文体とは異なることが見て取れた。

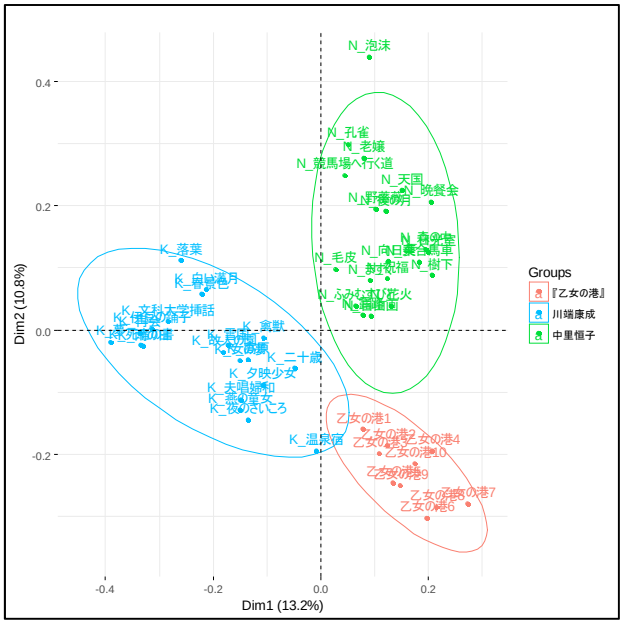

図 3.4 『乙女の港』タグ付き形態素対応分析個体の第 1、2 スコアの散布図

『乙女の港』のタグつき形態素の変数第 1 スコアの棒グラフを図 3.5 に示す。第 1 スコア軸の正の方向には「之_助詞」、「…_記号」、「御_接頭詞」、「小さな_連体詞」、「なんて_助詞」、「丁度_副詞」、「けれど_助詞」、「なんだか_副詞」、「じゃ_助詞」と「こう_副詞」の変数が現れた。中でも、「助詞」と「副詞」が大半を占め、中里恒子の文章と『乙女の港』では助詞と副詞の使用が特徴的であることが分かった。第 1 スコアの負の方向には「と_フィラー」、「らしかっ_助動詞」、「に対する_助詞」、「けれども_接続詞」、「ところが_接続詞」、「あり_助動詞」、「程_助詞」、「だっ_助動詞」「べき_助動詞」と「むしろ_副詞」が現れた。川端康成の作品ではこのようなタグ付き形態素が多く用いられている。

『乙女の港』のタグ付き形態素変数の第 2 スコアの棒グラフを図 3.6 に示す。第 2 スコアの正の方向に「多_接頭詞」、「なんぞ_助詞」、「一層_副詞」、「わ_助詞」、「絶えず_副詞」、「丁度_副詞」、「すぐ_副詞」、「暫く_副詞」、「殆ど_副詞」と「又_接続詞」の変数が現れた。川端康成作品の一部と中里恒子作品では、このような変数が多く用いられている。第 2 スコアの負の方向に「。_記号」、「々_記号」、「ちょうど_副詞」、「こんなに_副詞」、「なにか_副詞」、「けれど_助詞」、「無論_副詞」、「…_記号」、「直ぐ_副詞」と「そうして_接続詞」の変数が現れた。川端康成作品の一部と『乙女の港』ではこのような変数が多く用いられている。

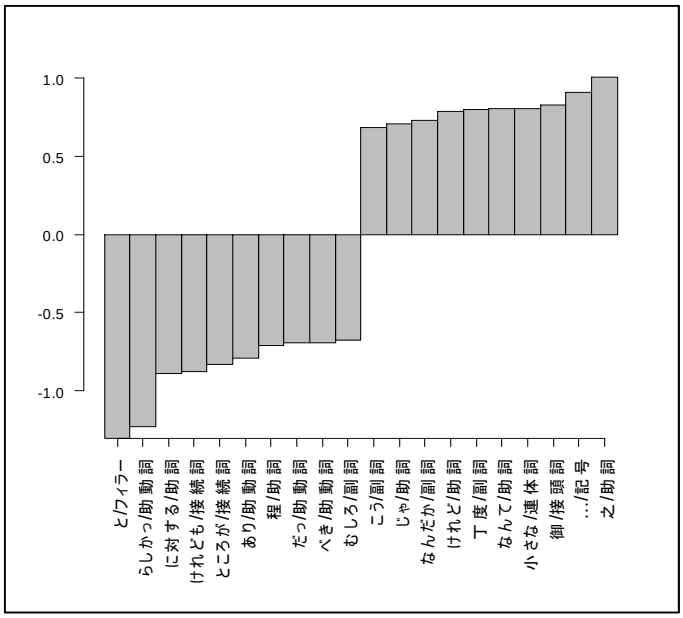

図 3.5 『乙女の港』タグ付き形態素の変数第 1 スコアの棒グラフ

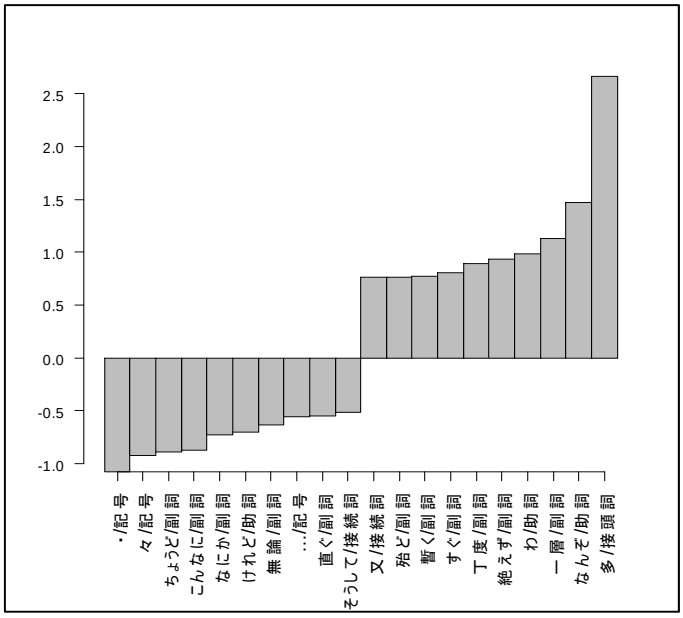

図 3.6 『乙女の港』タグ付き形態素の第 2 スコアの棒グラフ

『乙女の港』のコーパスから抽出した文節パターンの次元数は 351 である。文節パターン対応分析個体の第 1、2 スコアの散布図を図 3.7 に示す。図 3.7 に示したように、川端康成の作品は第 2、3、4 象限にプロットされた。中里恒子の作品は第 2 スコア軸を跨ぎ、第 1 と第 3 象限にプロットされた。『乙女の港』の第 1 スコア軸を跨ぎ、第 1 と第 4 象限にプロットされた。文節パターンを用いた場合、文字記号 bi-gram とタグ付き形態素ほど分類できていなかったことが見て取れた。

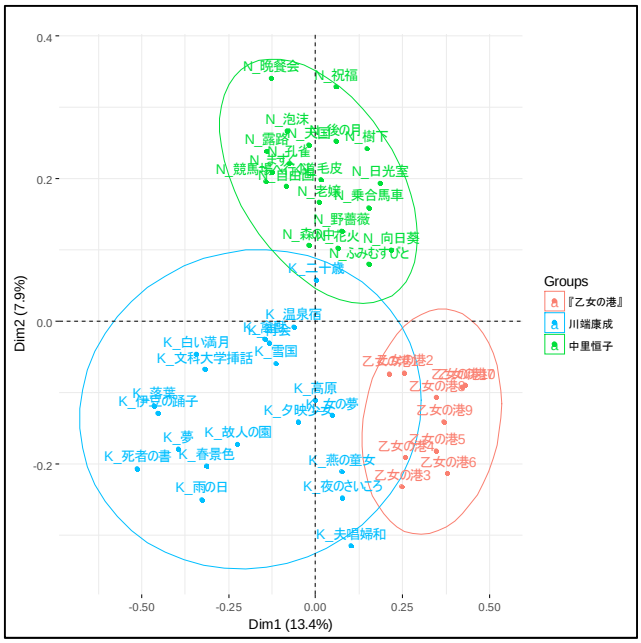

図 3.7 『乙女の港』文節パターン対応分析個体の第 1、2 スコアの散布図

『乙女の港』の文節パターンの変数の第 1 スコアの棒グラフを図 3.8 に示す。第 1 スコア軸の正の方向には「名詞_名詞_名詞_は」、「名詞_。」、「名詞_名詞_。」、「代名詞_、」、「名詞_まで_、」、「名詞_の_、」、「名詞_名詞_が_、」、「動詞_名詞_に_、」、「接頭詞_名詞_名詞_の」と「形容詞_て_、」の変数が現れた。川端康成と中里恒子の一部の作品と『乙女の港』はこのような変数が多く用いられている。第 1 スコア軸の負の方向には「名詞_と_が」、「代名詞_は」、「動詞_助動詞_名詞_助動詞_助動詞_助動詞_。」、「動詞_助動詞_名詞」、「代名詞_に_は」、「代名詞_の」、「代名詞_名詞_は」と「動詞_助動詞_名詞_は」、「動詞_助動詞_名詞_助動詞_。」と「名詞_と_を」が現れ、川端康成作品の一部と中里恒子作品の一部ではこのような文節のパターンが多く用いられている。

『乙女の港』の文節パターンの第 2 スコアの棒グラフを図 3.9 に示す。第 2 スコア正の方向には「動詞_て_動詞_助動詞_名詞_助動詞」、「サ変接続_動詞_動詞_助動詞_。」、「名詞_動詞_を」、「動詞_動詞_て_動詞_助動詞」、「動詞_動詞_名詞_助動詞_助動詞_。」、「名詞_さえ」、「副詞_は」、「サ変接続_動詞_て_動詞_助動詞」、「名詞_動詞_助動詞」と「形容詞_名詞_を」の変数が現れた。川端康成と中里恒子の一部の作品と『乙女の港』はこのような変数を多く用いている。第 2 スコア負の方向には「動詞_て_動詞_名詞_は_、」、「名詞_助動詞_助動詞_が_、」、「副詞_助動詞_助動詞_。」、「動詞_名詞_は_、」、「動詞_助動詞_けれども_、」、「名詞_と_を」、「名詞_名詞_助動詞_助動詞_。」、「名詞_助動詞_ので_、」、「名詞_名詞_。」と「動詞_助動詞_で_、」の変数が現れ、川端康成と中里恒子作品の一部ではこのような文節のパターンが多く用いられている。

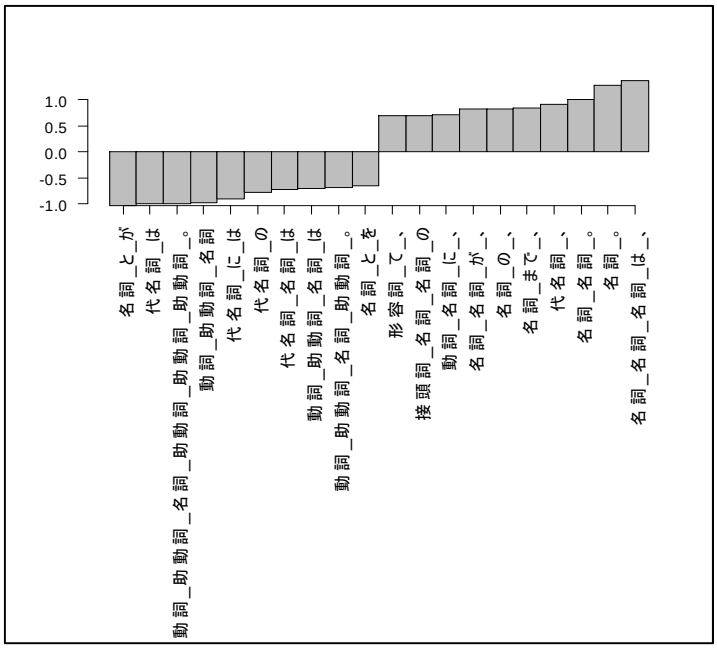

図 3.8 『乙女の港』文節パターン変数の第 1 スコアの棒グラフ

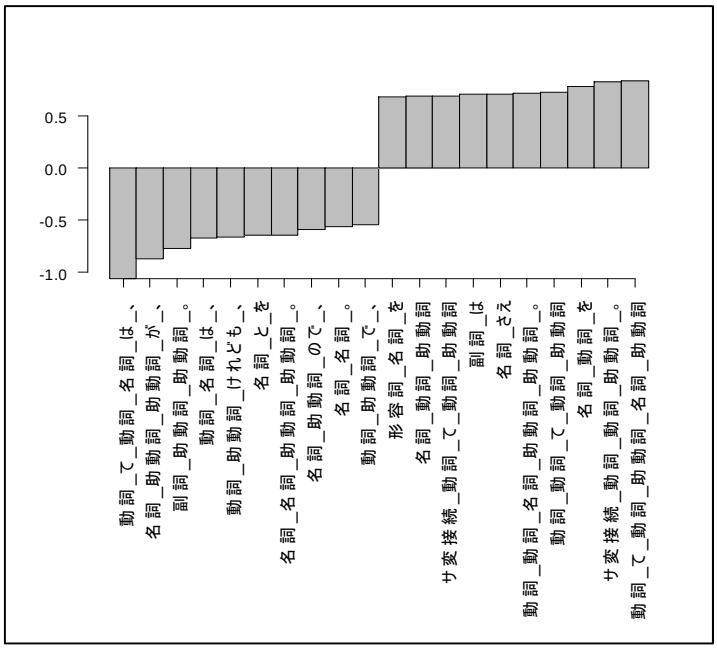

図 3.9 『乙女の港』文節パターン変数の第 2 スコアの棒グラフ

以上の文字記号 bi-gram、タグ付き形態素と文節パターンを用いた川端康成と中里恒子の対応分析の結果では、川端康成と中里恒子の文章は異なる位置に集まり、『乙女の港』の各章は川端康成と中里恒子から離れている位置にプロットされた。この結果から、いずれの文体特徴量においても『乙女の港』の文体は川端康成と中里恒子の文体とは異なることが分かった。

3.3 階層的クラスター分析の結果

本節では、文字記号 bi-gram、タグ付き形態素と文節パターンの階層的クラスター分析を行い、クラスター分析を行う際に ward 法と KLD 距離を用いた。また川端康成の作品、中里恒子の作品と『乙女の港』各章の位置関係の考察を通じて『乙女の港』の帰属を判断した。

図 3.10 に示した文字記号 bi-gram を用いた『乙女の港』のクラスター分析では、樹形図は大きく三つのクラスターに分かれており、左側から順番に『乙女の港』のクラスター、川端康成のクラスター、中里恒子のクラスターになる。

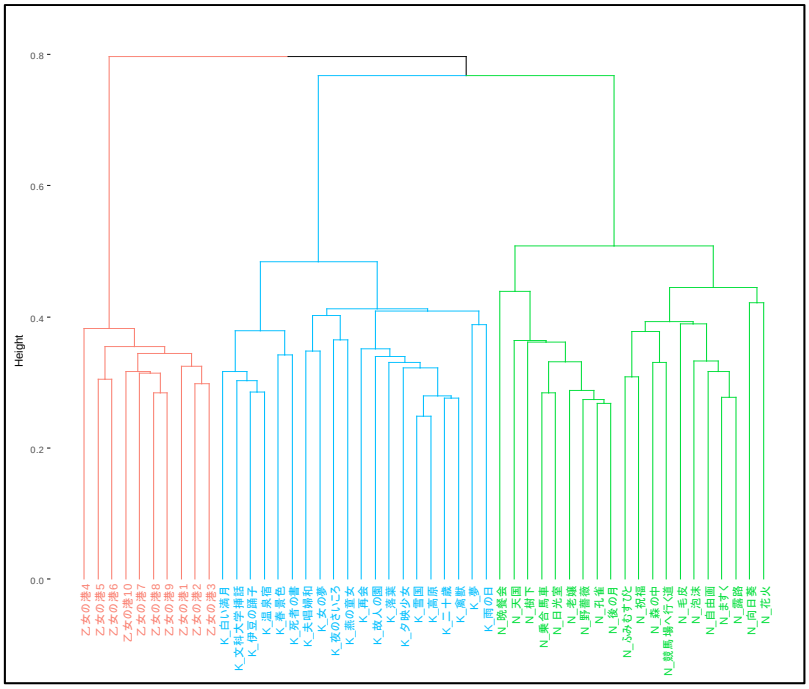

図 3.10 『乙女の港』文字記号 bi-gram の階層的クラスター樹形図

図 3.11 に示したタグ付き形態素を用いた『乙女の港』のクラスター分析でも大きく三つのクラスターに分かれ、左側から順番に川端康成のクラスター、『乙女の港』のクラスターと中里恒子のクラスターになる。

図 3.12 に示したタグ付き形態素を用いた場合、文字記号 bi-gram と同じく、デンドログラムは大きく三つのクラスターに分かれており、左側から順番に『乙女の港』のクラスター、川端康成のクラスターと中里恒子のクラスターになる。

以上の文字記号 bi-gram、タグ付き形態素と文節パターンを用いた川端康成と中里恒子のクラスター分析の結果では、『乙女の港』の各章は川端康成と中里恒子と異なるクラスターに分類されている。タグ付き形態素の場合、『乙女の港』の各章は中里恒子と先に結合したため文体は中里恒子に近いことを示した。文字記号 bi-gram と文節パターンでは、『乙女の港』の文体は川端康成と中里恒子との文体が異なることが明らかになった。

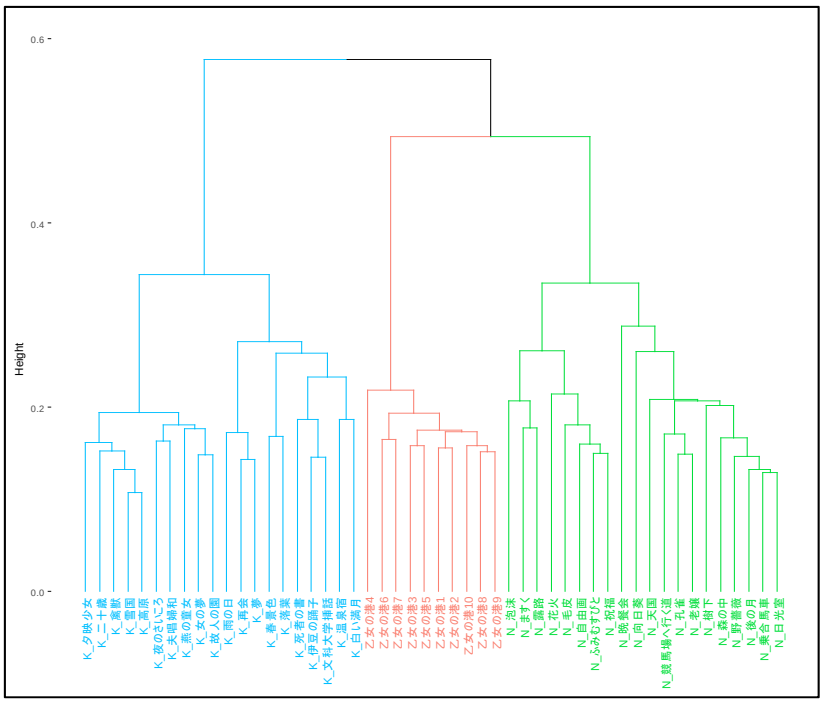

図 3.11 『乙女の港』タグ付き形態素の階層的クラスター樹形図

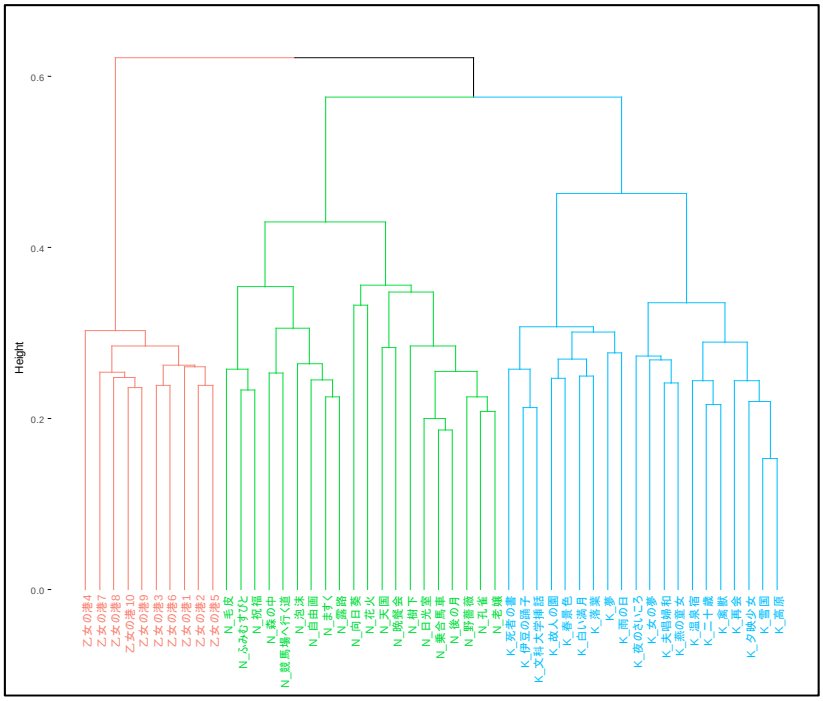

図 3.12 『乙女の港』文節パターンの階層的クラスター樹形図

3.4 分類器による判別結果

本節では『乙女の港』の代筆問題に関して、文字記号 bi-gram、タグ付き形態素と文節パターンの三つの特徴量と、AdaBoost、HDDA、LMT、RF と SVM の 5 つの分類器を用いて判別

分析を行った。各文体特徴量における分類器の性能を示す適合率 (Precision)、再現率 (Recall) と F-尺度 (F-measure)を表 3.5 に示す。

表3.5 各文体特徴量における分類器の性能評価

特徴量	評価指標	AdaBoost	HDDA	LMT	RF	SVM
文字記号 bi-gram	Precision	1	1	1	1	0.87
	Recall	1	0.95	1	1	1
	F-measure	1	0.97	1	1	0.93
タグ付き 形態素	Precision	0.90	0.95	1	1	1
	Recall	0.95	0.9	1	1	0.9
	F-measure	0.93	0.92	1	1	0.95
文節 パターン	Precision	0.88	1	0.83	1	0.95
	Recall	0.75	0.95	0.95	1	1
	F-measure	0.79	0.97	0.88	1	0.98

3.4.1 文字記号 bi-gram

文字記号bi-gramによる『乙女の港』の10章の分類結果と統合結果を表3.6に示す。文字記号bi-gram特徴量において、AdaBoostとRFではすべての章は川端康成に判別された。HDDAでは第1、2、4、5、7、8、9、10章は中里恒子、第3、6章は川端康成に判別された。LMTでは第1、2、3、5、8、9、10章は中里恒子、第4、6、7章は川端康成に判別された。SVMでは第1、3、4、5、6、7、8章は中里恒子、第2、9、10章は川端康成に判別された。5つの分類器の統合結果では、第1、5、8章は中里恒子、第2、3、4、6、7、9、10章は川端康成に判別される結果になった。

表3.6 文字記号bi-gramを用いた5つの分類器による判別結果 (K:川端康成 N:中里恒子)

乙女の港	AdaBoost	HDDA	LMT	RF	SVM	統合結果
第1章	K	N	N	K	N	N
第2章	K	N	N	K	K	K
第3章	K	K	N	K	N	K
第4章	K	N	K	K	N	K
第5章	K	N	N	K	N	N
第6章	K	K	K	K	N	K
第7章	K	N	K	K	N	K
第8章	K	N	N	K	N	N
第9章	K	N	N	K	K	K
第10章	K	N	N	K	K	K

3.4.2 タグ付き形態素

タグ付き形態素による『乙女の港』の10章の分類結果および統合結果を表3.7に示す。タグ付き形態素特徴量において、AdaBoostでは第1、2、3、4、8章は中里恒子、第5、6、7、9、10章は川端康成に判別された。HDDAでは第1、2、7、8、10章は中里恒子、第3、4、5、6、9章は川端康成に判別された。LMTとSVMではすべての章が中里恒子に判別された。RFでは第1、2、3、4、7、8、10章は中里恒子、第5、6、9章は川端康成に判別された。5つの分類器の統合結果では、第1、2、3、4、7、8、10章は中里恒子、第5、6、9章は川端康成に判別される結果になった。

表3.7 タグ付き形態素を用いた5つの分類器による判別結果 (K:川端康成 N:中里恒子)

乙女の港	AdaBoost	HDDA	LMT	RF	SVM	統合結果
第1章	N	N	N	N	N	N
第2章	N	N	N	N	N	N
第3章	N	K	N	N	N	N
第4章	N	K	N	N	N	N
第5章	K	K	N	K	N	K
第6章	K	K	N	K	N	K
第7章	K	N	N	N	N	N
第8章	N	N	N	N	N	N
第9章	K	K	N	K	N	K
第10章	K	N	N	N	N	N

3.4.3 文節パターン

文節パターンを用いた判別結果を表3.8に示す。文節パターン特徴量において、分類器AdaBoostでは第1、2、4、7、8、10章は中里恒子、第3、5、6、9章は川端康成に判別された。HDDAでは第1、2、7、10章は中里恒子、第3、4、5、6、8、9章は川端康成に判別された。LMTでは第1、2、3、6、7、8、9、10章は中里恒子、第4、5章は川端康成と判別された。RFでは第1、2、7、8、10章は中里恒子、第3、4、5、6、9章は川端康成に判別された。SVMでは第1、2、6、7、8、9、10章は中里恒子、第3、4、5章は川端康成に判別された。5つの分類器の統合結果では、第1、2、7、8、10章は中里恒子、第3、4、5、6、9章は川端康成に判別される結果になった。

表3.8 文節パターンを用いた5つの分類器による判別結果 (K:川端康成 N:中里恒子)

乙女の港	AdaBoost	HDDA	LMT	RF	SVM	統合結果
第1章	N	N	N	N	N	N
第2章	N	N	N	N	N	N
第3章	K	K	N	K	K	K
第4章	N	K	K	K	K	K
第5章	K	K	K	K	K	K
第6章	K	K	N	K	N	K
第7章	N	N	N	N	N	N
第8章	N	K	N	N	N	N
第9章	K	K	N	K	N	K
第10章	N	N	N	N	N	N

3.5 本章のまとめ

『乙女の港』は中里恒子が原稿を書き、執筆指導のため川端康成が原稿に手を加えて発表したとされた小説である。先行研究では、『乙女の港』は川端康成作と中里恒子作で意見が分かれている。本章では、文字記号bi-gram、タグ付き形態素と文節パターンを文体特徴量とし、対応分析、クラスター分析、AdaBoost、HDDA、LMT、RFとSVMを計量的手法として分析を行った。

文字記号bi-gram、タグ付き形態素と文節パターンの対応分析とクラスター分析の結果から、『乙女の港』の文体は川端康成と中里恒子から離れ、川端康成の改稿により両者の文体特徴が融合したものになっている。

この両者の特徴が融合した文体はどちらの文体に近いかを判断するために、AdaBoost、HDDA、LMT、RFとSVMによる『乙女の港』各章の2群判別を行った。その結果、文字記号bi-gramの統合結果では、第1、5、8章は中里恒子、第2、3、4、6、7、9、10章は川端康成に判別された。タグ付き形態素では第1、2、3、4、7、8、10章は中里恒子、第5、6、9章は川端康成に判別された。文節パターンでは第1、2、7、8、10章は中里恒子、第3、4、5、6、9章は川端康成に判別された。以上の2群判別の結果には川端康成と中里恒子の結果が入り交じり、川端康成の改稿により『乙女の港』の文体は川端康成と中里恒子の文体要素を有するものになっていると考えられる。

以上の分析結果より、『乙女の港』には川端康成と中里恒子の文体要素が含まれているため、この小説は川端康成と中里恒子の共同執筆と見なせる。

第 4 章 『花日記』の代筆問題研究

4.1 研究背景

『花日記』は 1938 年 4 月から 1939 年 3 月にかけて、実業之日本社が発行した少女向けの雑誌『少女の友』に連載された少女小説である。この小説は 12 章からなり、現在、1984 年に完結した新潮社発行の川端康成全集の第 20 巻に収録されている。『花日記』の各章の詳細情報を表 4.1 に示す。

表 4.1 『花日記』の各章の詳細情報

『花日記』の各章		発行時間	雑誌	文字数
1	姉嫁ぐ	1937 年 4 月	少女の友	4387
2	菫のなかで	1937 年 5 月	少女の友	2587
3	嘘の妹	1937 年 6 月	少女の友	3095
4	うしろすがた	1937 年 7 月	少女の友	6900
5	うしろすがた (つづき)	1937 年 8 月	少女の友	4048
6	唱歌会	1937 年 9 月	少女の友	4326
7	夏の海	1937 年 10 月	少女の友	3865
8	花日記	1937 年 11 月	少女の友	2874
9	新学期	1937 年 12 月	少女の友	3956
10	級長選挙	1938 年 1 月	少女の友	4778
11	姉病む	1938 年 2 月	少女の友	4339
12	—[2]	1938 年 3 月	少女の友	4640

『花日記』の代筆問題の証拠となったのは川端康成と中里恒子との執筆指導についての往復書簡である。書簡の内容を次に示す。

1938 (昭和13年)9月17日付、中里恒子から川端康成へ (川端康成全集補巻二, p. 295)。

けふ少女の友買ひ、花日記にかかります。これは自分でも書いてゐてたのしみです。勿論虚構の人物ですけれどその人物に私の思ってゐることをみんなさせているせいかもしれません。

1938 (昭和13年)9月25日付、川端康成から中里恒子へ (川端康成全集補巻二, p. 303)。

原稿拝受。朝日の時評書き上げれば参ります。私共は霧ヶ峰から諏訪へ出て、それから伊那を廻って木曾へ参ることになるかと思ひますが、木曾福島あたりで落ち合へればと存じます。木曾のどこで何日に会ふか、電報しますが、御主人は今日と云って明日お休みに

[2]『花日記』の第 12 章には標題がついていない。

なつていらつしゃるといふわけに参りませんかしら。軽井澤を二十七日か八日に出ると思ひます、いづれ速達か電報で打ち合せます。

往復書簡の「けふ少女の友買ひ、花日記にかかります。これは自分でも書いてゐてたのしみです。」と「原稿拝受」の文面から、『花日記』は中里恒子による代筆の疑惑が浮上した。このような往復書簡を根拠に、川端康成が中里恒子の作成した原稿に手を加えて『花日記』を完成させたと言われている (小谷野, 2013)。また、大森 (1991)は、同性愛の完成度の観点から、少女小説の『乙女の港』、『花日記』と『美しい旅』の比較分析を行い、『花日記』の完成度が低いと結論付けた。しかし、大森 (1991)は、「同性愛のモチーフは川端康成の作意に発するものかを俄に推断し難い」とも述べ、同性愛の内容分析から『花日記』の著者帰属問題が解決できなかった。川端康成研究書である『川端康成詳細年譜』には、『花日記』は川端康成と中里恒子の共同執筆とされている (小谷野・深澤, 2016)。

本章では、『花日記』を章ごとに分け、計量文体分析の手法を用いて、『花日記』の著者帰属問題を明らかにする。そのために用いた文体特徴量は文字記号 bi-gram、タグつき形態素と文節パターンである。計量的手法は対応分析、クラスター分析、AdaBoost、HDDA、LMT、RF と SVM である。

4.2 対応分析の結果

『花日記』のコーパスから抽出した文字記号 bi-gram の次元数は 2111 である。文字記号 bi-gram 対応分析個体の第 1、2 スコアの散布図を図 4.1 に示す。図 4.1 では、川端康成の作品と中里恒子の作品は大まかに 2 群に分かれ、それぞれ第 4 象限と第 1 象限にプロットされた。『花日記』の各章は両者の作品グループから離れた第 3 象限にプロットされた。文字記号 bi-gram を用いた場合、『花日記』の文体は川端康成と中里恒子の文体とは異なることが見て取れた。

『花日記』の文字記号 bi-gram 変数の第 1 スコアの棒グラフを図 4.2 に示す。第 1 スコアが正の方向には「祐三」、「士子」、「。祐」、「富士」、「踊の」、「三は」、「代子」、「戦争」、「。千」と「落葉」の変数が現れた。川端康成と中里恒子の作品では、このような文字記号 bi-gram を多く用いている。第 1 スコアの負の方向には「桃子」、「子姉」、「お兄」、「お姉」、「まは」、「兄さ」、「、英」、「英子」、「清子」と「ほみ」が現れた。『花日記』ではこういった文字記号 bi-gram が多く用いられている。

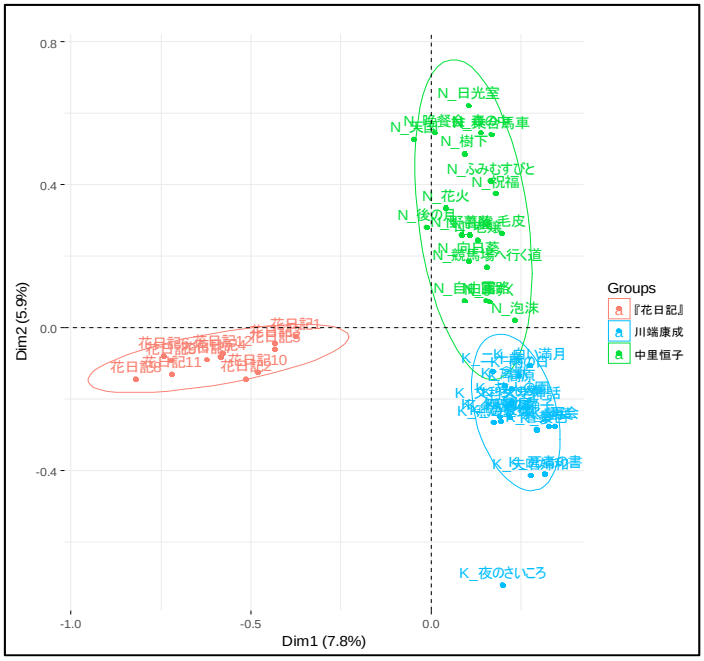

図4.1 『花日記』文字記号bi-gram対応分析個体の第1、2スコアの散布図

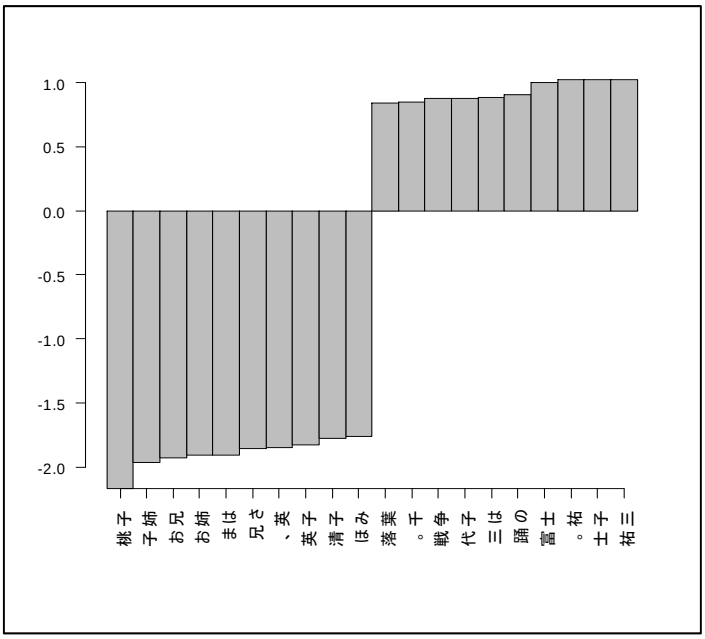

図 4.2 『花日記』文字記号 bi-gram 変数の第 1 スコアの棒グラフ

『花日記』の文字記号 bigram 変数の第 2 スコアの棒グラフを図 4.3 に示す。第 2 スコアの正の方向に「供た」、「アン」、「ヌの」、「ンヌ」、「園子」、「デリ」、「アデ」、「森之」、「之助」と「菊代」の変数が現れた。中里恒子の作品はこのような変数を多く用いている。第 2 スコアの負の方向に「ち子」、「水田」、「みち」、「。水」、「子達」、「、踊」、「、水」、「踊子」、「山は」と「、延」の 10 個が現れた。川端康成作品と『花日記』では、このような変数が多く用いられている。

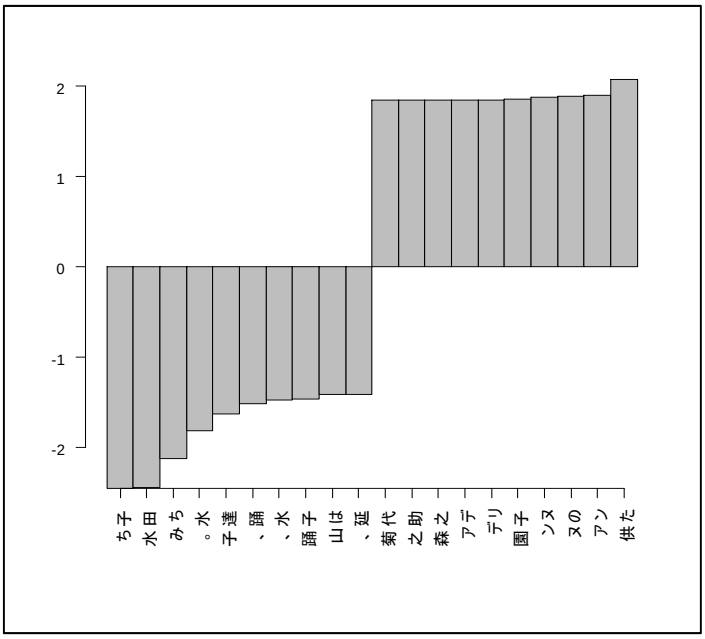

図 4.3 『花日記』文字記号 bi-gram 変数の第 2 スコアの棒グラフ

『花日記』のコーパスから抽出したタグ付き形態素の次元数は218である。タグ付き形態素対応分析個体の第1、2成分のスコアを図4.4に示す。図4.4では、『温泉宿』以外の川端康成作品は第3象限にプロットされた。中里恒子の作品は第2スコア軸を跨ぎ、第1と2象限にプロットされた。『花日記』の各章は両作家の作品グループから離れた第4象限にプロットされ、タグ付き形態素からする場合、『花日記』の文体は川端康成と中里恒子の文体とは異なることが見て取れた。

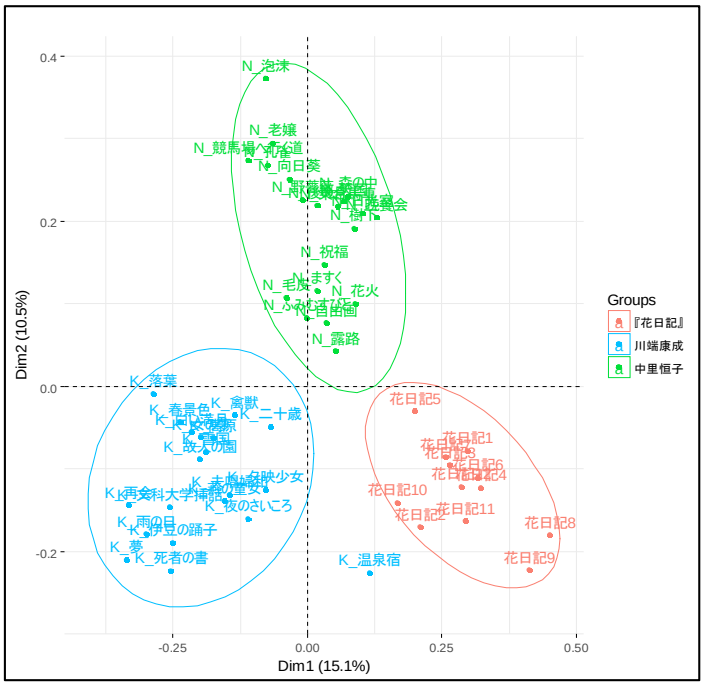

図 4.4 『花日記』タグ付き形態素対応分析個体の第 1、2 スコアの散布図

『花日記』のタグ付き形態素の変数第 1 スコアの棒グラフを図 4.5 に示す。第 1 スコアが正の方向には「お_接頭詞」、「けれど_助詞」、「…_記号」、「けれど_接続詞」、「._記号」、「だり_助詞」、「いっぱい_副詞」、「御_接頭詞」、「わ_助詞」と「かしら_助詞」の変数が現れた。中里恒子作品の一部と『花日記』ではこのようなタグ付き形態素が多く用いられている。第 1 スコアの負の方向には「と_フィラー」、「らしかっ_助動詞」、「けれども_接続詞」、「程_助詞」、「けれども_助詞」、「に対する_助詞」、「無論_副詞」、「とにかく_副詞」、「ところが_接続詞」と「なかろ_助詞」が現れた。川端康成の作品と中里恒子作品の一部では、こういったタグ付き形態素が多く用いられている。

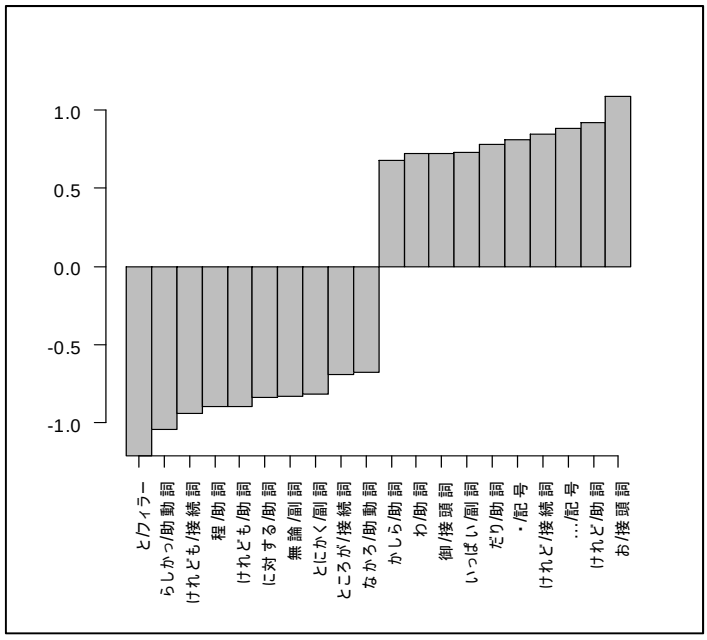

図 4.5 『花日記』タグ付き形態素の変数第 1 スコアの棒グラフ

『花日記』のタグ付き形態素変数の第 2 スコア棒のグラフを図 4.6 に示す。第 2 スコアが正の方向には「多_接頭詞」、「なんぞ_助詞」、「之_助詞」、「丁度_副詞」、「まもなく_副詞」、「暫く_副詞」、「もし_副詞」、「又_接続詞」、「こうして_接続詞」と「絶えず_副詞」の変数が現れた。中里恒子作品ではこのようなタグ付き形態素を多く用いている。第 2 スコアの負の方向には「と_フィラー」、「直ぐ_副詞」、「ちょうど_副詞」、「らしかっ_助動詞」、「._記号」、「なにか_副詞」、「お_接頭詞」、「間もなく_副詞」、「無論_副詞」と「だっ_助動詞」が現れた。『花日記』では、こういったタグ付き形態素が多く用いられている。

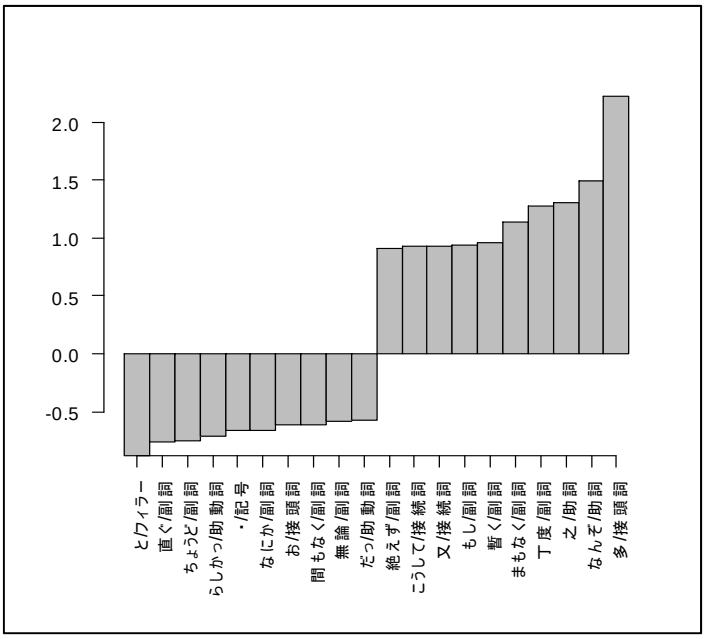

図 4.6 『花日記』タグ付き形態素変数の第 2 スコアの棒グラフ

『花日記』のコーパスから抽出した文節パターンの次元数は352である。文節パターン対応分析個体の第1、2成分のスコアを図4.7に示す。図4.7に示したように、川端康成の作品は第2、3、4象限にプロットされた。中里恒子の作品は第2スコア軸を跨ぎ、第1と第2象限にプロットされた。『花日記』の各章は第1スコア軸を跨ぎ、第1と第4象限にプロットされた。文節パターンからする場合、『花日記』の文体は川端康成と中里恒子の文体とは異なることが見て取れた。

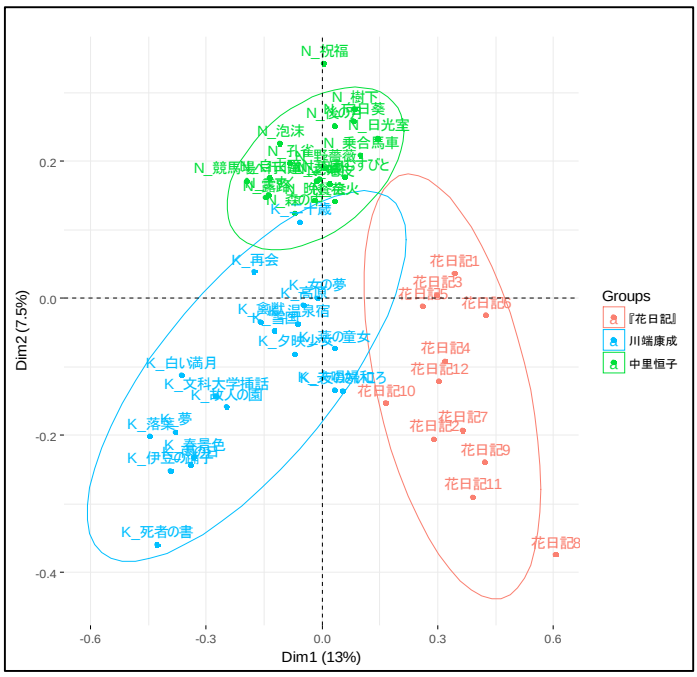

図 4.7 『花日記』文節パターン対応分析個体の第 1、2 スコアの散布図

『花日記』の文節パターン変数の第 1 スコアの棒グラフを図 4.8 に示す。第 1 スコアが正の方向には「名詞_名詞_名詞_は_、」、「接頭詞_名詞_名詞_は」、「接頭詞_名詞_名詞_が」、「接頭詞_名詞_名詞_の」、「名詞_。」、「名詞_名詞_名詞_は」、「動詞_助動詞_に_、」、「形容詞_名詞_に_、」、「形容詞_て_、」と「名詞_名詞_は_、」の変数が現れた。川端康成と中里恒子作品の一部と『花日記』では、このようなタグ付き形態素が多く用いられている。第 1 スコアの負の方向には「動詞_助動詞_名詞_助動詞_助動詞_助動詞_。」、「代名詞_は」、「代名詞_に_は」、「代名詞_の」、「動詞_助動詞_名詞_は」、「代名詞_名詞_は」、「動詞_助動詞_名詞_助動詞_。」、「動詞_て_から」、「動詞_が_、」と「形容詞_動詞_て」が現れた。川端康成と中里恒子作品の一部では、こういったタグ付き形態素が多く用いられている。

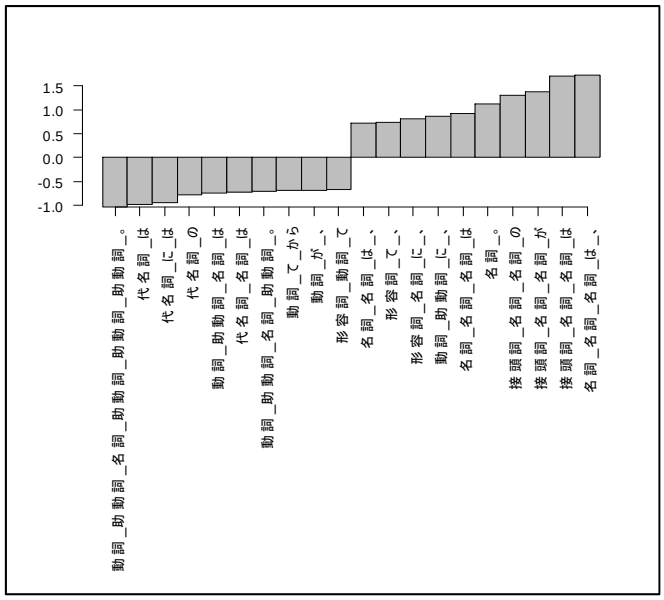

図 4.8 『花日記』文節パターン変数の第 1 スコアの棒グラフ

『花日記』文節パターン変数の第 2 スコアの棒グラフを図 4.9 に示す。第 2 スコアが正の方向には「名詞_さえ」、「連体詞_名詞_助動詞」、「名詞_動詞_を」、「動詞_動詞_て_動詞_助動詞」、「動詞_動詞_名詞_助動詞_助動詞_。」、「サ変接続_が_、」、「動詞_も」、「動詞_て_動詞_助動詞_名詞_助動詞」、「代名詞_でも」と「名詞_動詞_、」の変数が現れた。川端康成の作品、『花日記』の一部と中里恒子の作品ではこのようなタグ付き形態素が多く用いられている。第 2 スコアの負の方向には「接頭詞_名詞_名詞_は」、「名詞_名詞_名詞_は_、」、「名詞_名詞_とも」、「接頭詞_名詞_名詞_の」、「接頭詞_名詞_名詞_が」、「動詞_て_動詞_助動詞_助動詞_。」、「名詞_と_を」、「サ変接続_名詞_は」、「動詞_助動詞_名詞」と「名詞_名詞_名詞_は」が現れた。川端康成と中里恒子作品の一部では、こういったタグ付き形態素が多く用いられている。

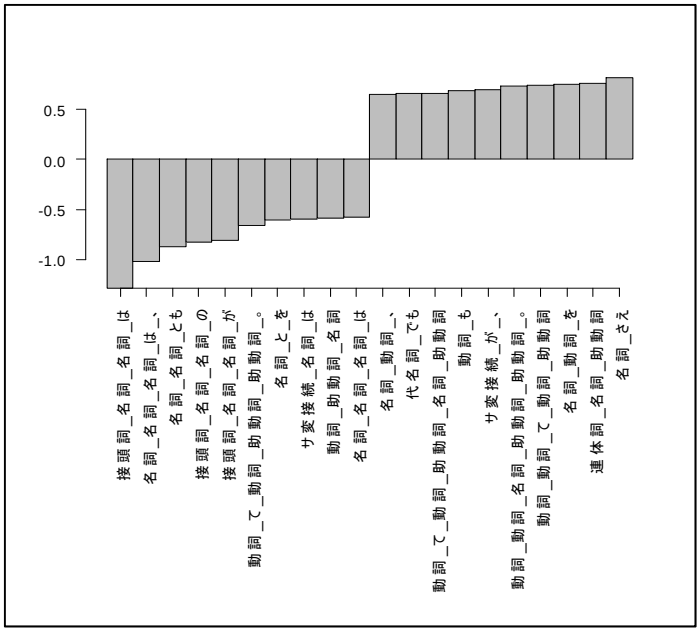

図 4.9 『花日記』文節パターン変数の第 2 スコアの棒グラフ

4.3 クラスター分析の結果

『花日記』の文字記号 bi-gram、タグ付き形態素と文節パターンを用いた階層的クラスター分析の結果を図 4.10~4.12 に示す。

図 4.10 の文字記号 bi-gram を用いた場合、樹形図は大きく三つのクラスターに分かれており、左側から順番に『花日記』のクラスター、川端康成のクラスターと中里恒子のクラスターになる。『花日記』の各章は単独で一つのクラスターを形成している。

図 4.11 のタグ付き形態素を用いた場合、樹形図は大きく三つのクラスターに分かれ、左側から順番に川端康成のクラスター、『花日記』のクラスターと中里恒子のクラスターになる。『花日記』の各章は単独で一つのクラスターを形成し、川端康成の作品と比べて、中里恒子の作品に近い。

図 4.12 ではの文節パターンを用いた場合、樹形図は大きく三つのクラスターに分かれている。左側から順番に『花日記』のクラスター、川端康成のクラスター、中里恒子のクラスターになる。中里恒子の小説『花火』は『花日記』のクラスターに入ったが、全体的な傾向では『花日記』の各章は単独で一つのクラスターを形成している。

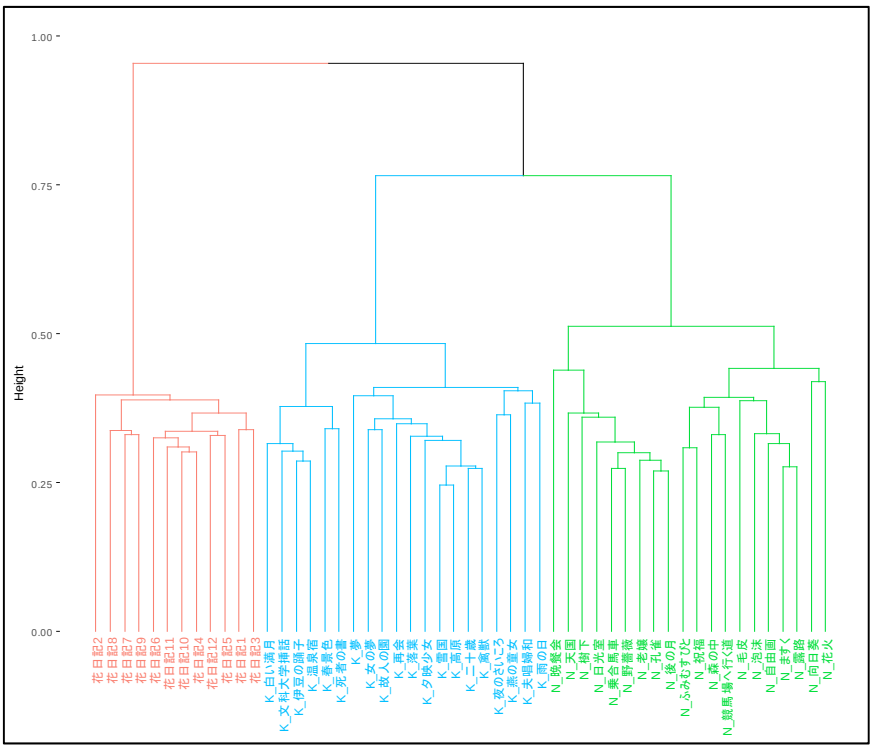

図 4.10 『花日記』の文字記号 bi-gram の階層的クラスター樹形図

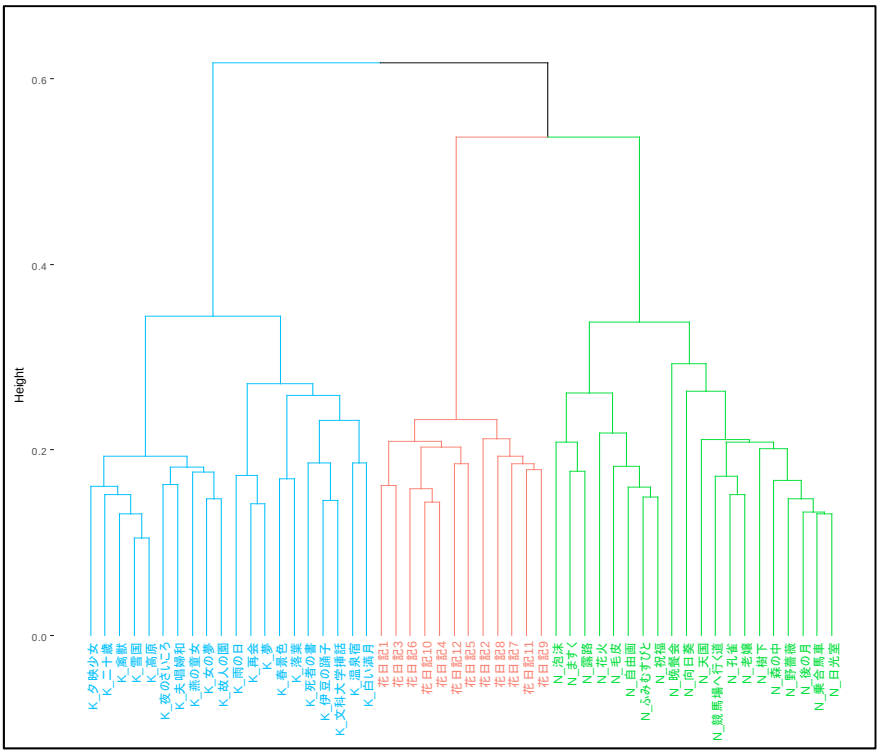

図 4.11 『花日記』のタグ付き形態素の階層的クラスター樹形図

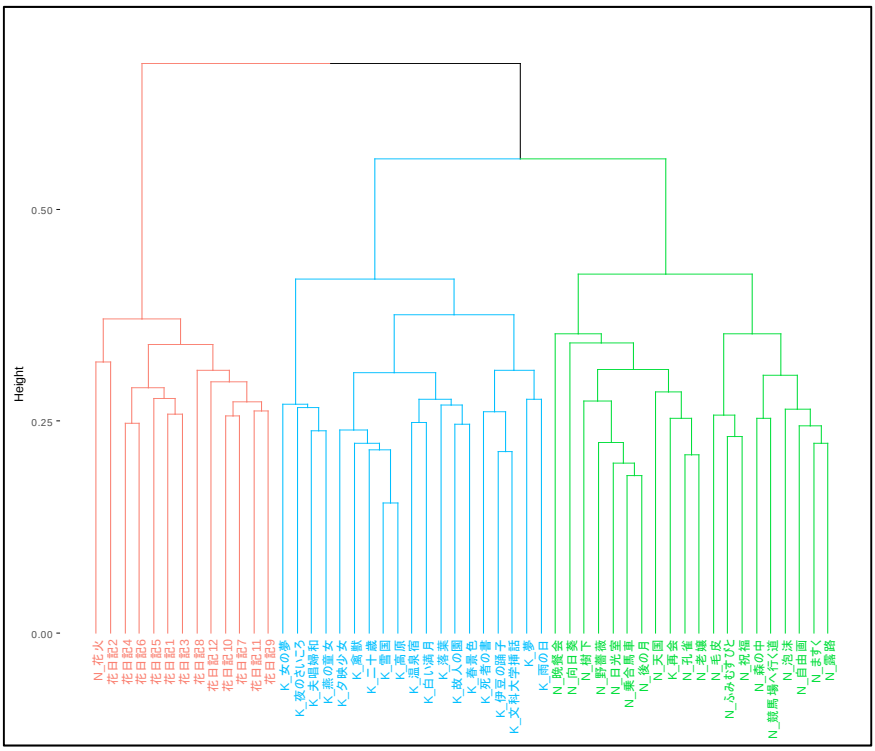

図 4.12 『花日記』の文節パターンの階層的クラスター樹形図

4.4 分類器による判別結果

4.4.1 文字記号 bi-gram

文字記号bi-gramによる『花日記』の12章の分類結果と統合結果を表4.2に示す。表4.2の文字記号bi-gram特徴量においては、AdaBoostでは第3章は中里恒子、第1、2、4、5、6、7、8、9、10、11、12章は川端康成に判別された。HDDAでは第1、3、5、7、8、10章は中里恒子、第2、4、6、7、8、9、11、12章は川端康成に判別された。LMTでは第1、2、3、6、7、8、9、10、11、12章は中里恒子、第4、5章は川端康成に判別された。RFでは、第1、3、7、8、10章は中里恒子、第2、4、5、6、9、11、12章は川端康成に判別された。SVMではすべての章は中里恒子に判別された。5つの分類器の統合結果では、第1、3、7、8、10章は中里恒子、第2、4、5、6、9、11、12章は川端康成に判別される結果になった。

4.4.2 タグ付き形態素

タグ付き形態素を用いた判別結果を表4.3に示す。タグ付き形態素特徴量においては、AdaBoostでは第1、3、4、5、6、7、8、9、10、12章は中里恒子、第2、11章は川端康成に判別された。HDDAでは第10章は中里恒子、第1、2、3、4、5、6、7、8、9、11、12章は川端康成に判別された。LMTではすべての章は中里恒子に判別された。RFでは、第1、3、4、5、6、7、8、9、10、11章は中里恒子、第2、12章は川端康成に判別された。SVMでは第1、3、5、7、8、10章は中里恒子、第2、4、6、9、11、12章は川端康成に判別された。5つの分類器の統合結果

では、第1、3、4、5、6、7、8、9、10章は中里恒子、第2、11、12章は川端康成に判別される結果になった。

表4.2 文字記号bi-gramを用いた5つの分類器による判別結果 (K:川端康成 N:中里恒子)

花日記	AdaBoost	HDDA	LMT	RF	SVM	統合結果
第1章	K	N	N	N	N	N
第2章	K	K	N	K	N	K
第3章	N	N	N	N	N	N
第4章	K	K	K	K	N	K
第5章	K	N	K	K	N	K
第6章	K	K	N	K	N	K
第7章	K	N	N	N	N	N
第8章	K	N	N	N	N	N
第9章	K	K	N	K	N	K
第10章	K	N	N	N	N	N
第11章	K	K	N	K	N	K
第12章	K	K	N	K	N	K

表4.3 タグ付き形態素を用いた5つの分類器による判別結果 (K:川端康成 N:中里恒子)

花日記	AdaBoost	HDDA	LMT	RF	SVM	統合結果
第1章	N	K	N	N	N	N
第2章	K	K	N	K	K	K
第3章	N	K	N	N	N	N
第4章	N	K	N	N	K	N
第5章	N	K	N	N	N	N
第6章	N	K	N	N	K	N
第7章	N	K	N	N	N	N
第8章	N	K	N	N	N	N
第9章	N	K	N	N	K	N
第10章	N	N	N	N	N	N
第11章	K	K	N	N	K	K
第12章	N	K	N	K	K	K

4.4.3 文節パターン

タグ付き形態素を用いた判別結果を表4.4に示す。文節パターン特徴量においては、分類器AdaBoostでは第2、3、4、8、9章は中里恒子、第1、5、6、7、10、11、12章は川端康成に判別された。HDDAでは第1、2、3、4、5、6、7、8、9、12章は中里恒子、第10、11章は川端康成

に判別された。LMTでは第1、2、3、4、8、9、10、11章は中里恒子、第5、6、7、12章は川端康成に判別された。RFでは第2、3、5、6、7、8、9、10、11章は中里恒子、第1、4、12章は川端康成に判別された。SVMでは第1、2、3、4、5、6、7、8、9、10、12章は中里恒子、第11章は川端康成に判別された。5つの分類器の統合結果では、第1、2、3、4、5、6、7、8、9、10章は中里恒子、第11、12章は川端康成に判別される結果になった。

表4.4 文節パターンを用いた5つの分類器による判別結果 (K:川端康成 N:中里恒子)

花日記	AdaBoost	HDDA	LMT	RF	SVM	統合結果
第1章	K	N	N	K	N	N
第2章	N	N	N	N	N	N
第3章	N	N	N	N	N	N
第4章	N	N	N	K	N	N
第5章	K	N	K	N	N	N
第6章	K	N	K	N	N	N
第7章	K	N	K	N	N	N
第8章	N	N	N	N	N	N
第9章	N	N	N	N	N	N
第10章	K	K	N	N	N	N
第11章	K	K	N	N	K	K
第12章	K	N	K	K	N	K

4.5 本章のまとめ

『花日記』は中里恒子が下書きを作り、川端康成がその原稿に加筆して発表したとされた小説である。小谷野・深澤 (2016)は、『花日記』を川端康成と中里恒子の共同執筆としている。本章では文字記号bi-gram、タグ付き形態素と文節パターンを文体特徴量とし、対応分析、クラスター分析、AdaBoost、HDDA、LMT、RFとSVMを計量的手法として分析を行った。

文字記号bi-gram、タグ付き形態素と文節パターンの対応分析とクラスター分析の結果から『花日記』の文体は川端康成と中里恒子から離れ、川端康成の加筆により両者の特徴が融合した文体になっている。

両者の特徴が融合した文体はどちらの文体に近いかを判断するために、AdaBoost、HDDA、LMT、RFとSVMによる『花日記』各章の2群分類を行った。その結果、文字記号bi-gramの統合結果では第1、3、7、8、10章は中里恒子、第2、4、5、6、9、11、12章は川端康成に判別される結果になった。タグ付き形態素では第1、3、4、5、6、7、8、9、10章は中里恒子、第2、11、12章は川端康成に判別される結果になった。文節パターンでは第1、2、3、4、5、6、7、8、9、10章は中里恒子、第11、12章は川端康成に判別される結果になった。

以上の分析より、『花日記』には川端康成と中里恒子の文体要素が含まれるため、小谷野・深澤 (2016)の『花日記』は共同執筆であるという説はデータ分析によって確認された。また、分類器による判別分析では川端康成になる結果が多いため、川端康成は中里恒子の原稿にしっかり手を加えたと考えられる。

第 5 章　『コスモスの友』の代筆問題研究

5.1 研究背景

『コスモスの友』は、1936年『少女倶楽部』の第10号に発表され、1984年完結した川端康成全集の第19巻に収録されている少女小説である。川勝 (2009)は、この作品も中里恒子によるものだと指摘している。この小説は『乙女の港』と『花日記』の1章ぐらいの長さ (4997文字)である。『コスモスの友』は敬体で書かれたもので、本研究では敬体表現を常体表現に改めた上で分析を行った。

5.2 対応分析の結果

『コスモスの友』のコーパスから抽出した文字記号bi-gramの次元数は1915である。文字記号bi-gram対応分析個体の第1、2スコアの散布図を図5.1に示す。図5.1では、川端康成と中里恒子の作品は第2スコア軸を境として分かれた。川端康成の作品は第2スコア軸の負の方向にプロットされ、中里恒子の作品は第2スコア軸の正の方向にプロットされた。『コスモスの友』は川端康成の作品グループに入り、その文体は川端康成に近い。

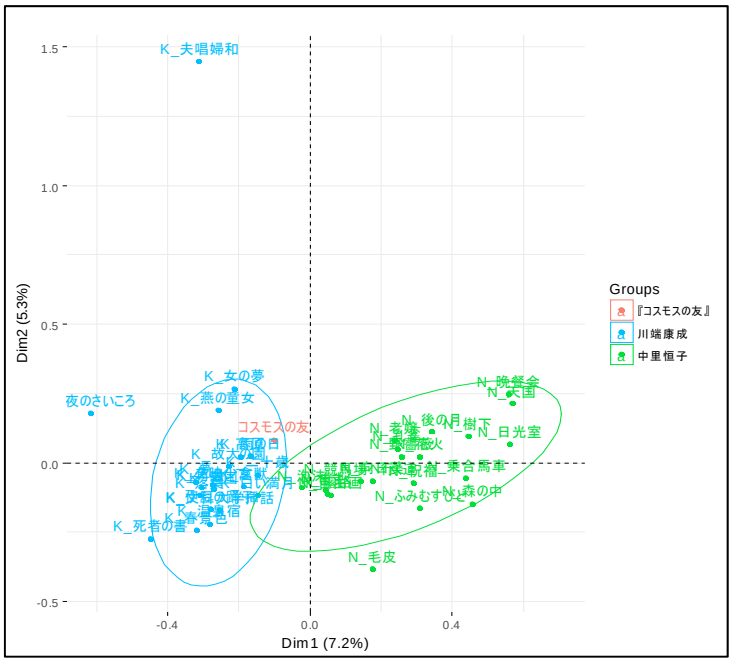

図 5.1　『コスモスの友』文字記号 bi-gram 対応分析個体の第 1、2 スコアの散布図

『コスモスの友』の文字記号 bigram 変数の第 1 スコアの棒グラフを図 5.2 に示す。第 1 スコア軸の正の方向には「郁子」、「園子」、「供た」、「アン」、「ヌの」、「ンヌ」、「一雄」、「お母」、「母さ」と「ルベ」の変数が現れた。中里恒子作品では、このようなタグ付き形態素が多く用いられている。第 1 スコア軸の負の方向には「ち子」、「水田」、「みち」、

「。水」、「。み」、「、踊」、「踊子」、「、水」、「。踊」、と「代子」が現れた。川端康成と『コスモスの友』作品ではこういったタグ付き形態素が多く用いられている。

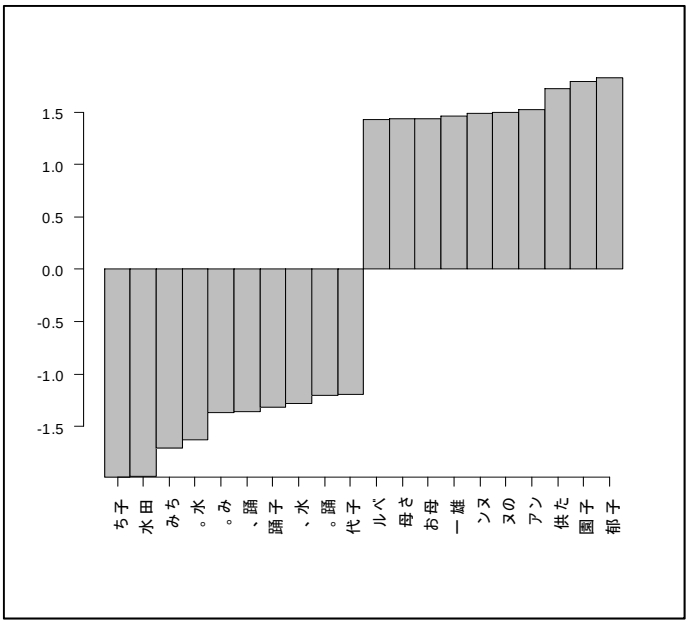

図 5.2 『コスモスの友』文字記号 bigram 変数の第 1 スコアの棒グラフ

『コスモスの友』のコーパスから抽出したタグ付き形態素の次元数は202である。タグ付き形態素対応分析個体の第1、2スコアの散布図を図5.3に示す。図5.3では、川端康成と中里恒子の作品は第2スコア軸を境として分かれた。川端康成の作品は第2スコア軸の負の方向にプロットされ、中里恒子の作品は第2スコア軸の正の方向にプロットされた。『コスモスの友』は川端康成の作品グループに入り、その文体は川端康成に近い。

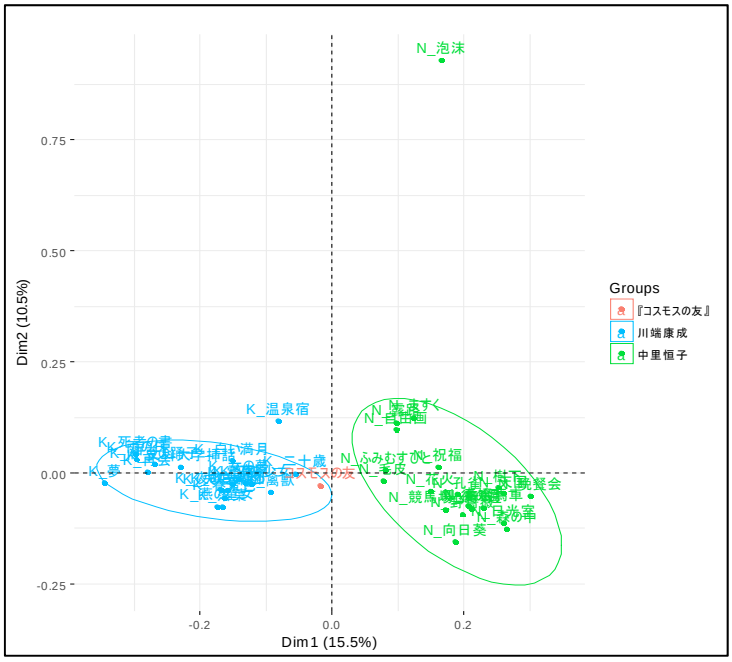

図 5.3 『コスモスの友』タグ付き形態素対応分析個体の第 1、2 スコアの散布図

『コスモスの友』のタグ付き形態素の変数第 1 スコアの棒グラフを図 5.4 に示す。第 1 スコア軸が正の方向には「之_助詞」、「丁度_副詞」、「…_記号」、「小さな_連体詞」、「こう_副詞」、「まもなく_副詞」、「わ_助詞」、「まるで_副詞」、「なんだか_副詞」と「多_接頭詞」の変数が現れた。中里恒子作品ではこのようなタグ付き形態素を多く用いられている。第 1 スコア軸の負の方向には「と_フィラー」、「らしかっ_助動詞」、「直ぐ_副詞」、「に対する_助詞」、「ちょうど_副詞」、「無論_副詞」、「なにか_副詞」、「だっ_助動詞」、「程_助詞」、と「。_記号」が現れた。川端康成と『コスモスの友』ではこういったタグ付き形態素を多く用いられている。

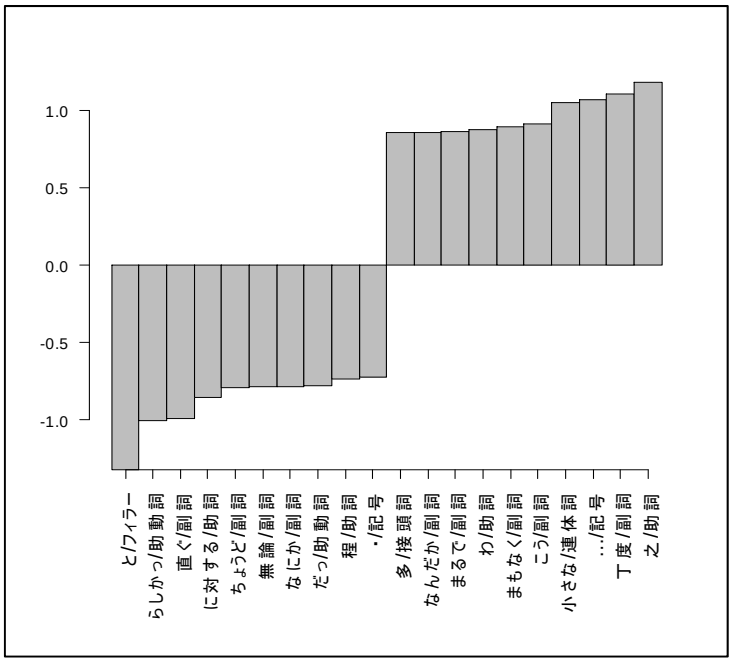

図 5.4 『コスモスの友』タグ付き形態素の変数第 1 スコアの棒グラフ

『コスモスの友』のコーパスから抽出した文節パターンの次元数は324である。文節パターン対応分析個体の第1、2スコアの散布図を図5.5に示す。図5.5では、川端康成の多くの作品は第1スコア軸の負の方向にプロットされた。中里恒子の作品は第1スコア軸の正の方向にプロットされた。『コスモスの友』は川端康成の作品グループに入り、文体は川端康成に近い。

『コスモスの友』の文節パターン変数の第 1 スコアの棒グラフを図 5.6 に示す。第 1 スコア軸の正の方向には「名詞_の_、」、「名詞_名詞_は_、」、「名詞_さえ」、「動詞_動詞_名詞_助動詞_助動詞_。」、「ー_副詞」、「形容詞_名詞_に_、」、「動詞_動詞_て_動詞_。」、「名詞_。」、「名詞_や_、」と「名詞_動詞_、」の変数が現れた。中里恒子作品ではこのようなタグ付き形態素が多く用いられている。第 1 スコア軸の負の方向には「動詞_助動詞_名詞」、「代名詞_は」、「代名詞_の」、「動詞_て_動詞_助動詞_助動詞_。」、「名詞_と_が」、「代名詞_に_は」、「動詞_で_助詞_助動詞_。」、「動詞_て_動詞_名詞_は_、」、「名詞_助動詞_助動詞_が_、」と「動詞_助動詞_名詞_助動詞_助動詞_助動詞_。」が現れた。川端康成と『コスモスの友』作品では、こういったタグ付き形態素が多く用いられている。

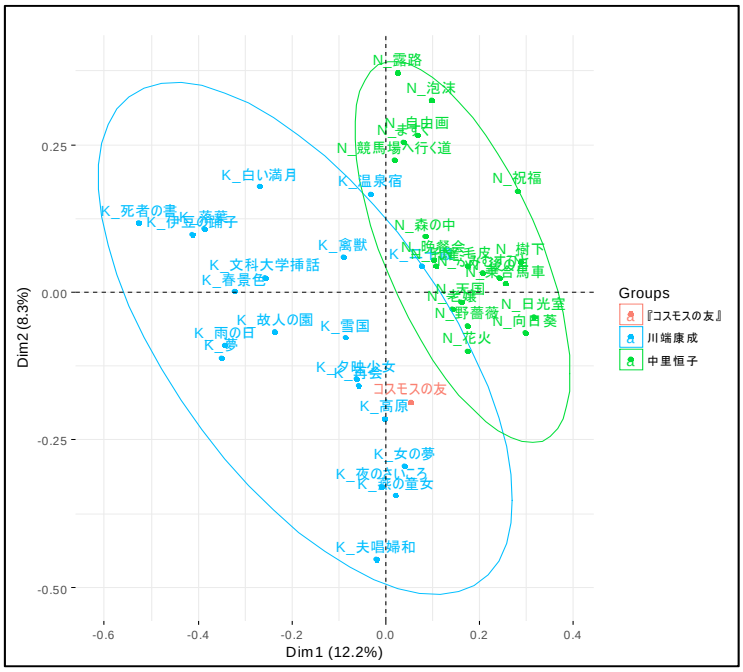

図 5.5 『コスモスの友』文節パターン対応分析個体の第 1、2 スコアの散布図

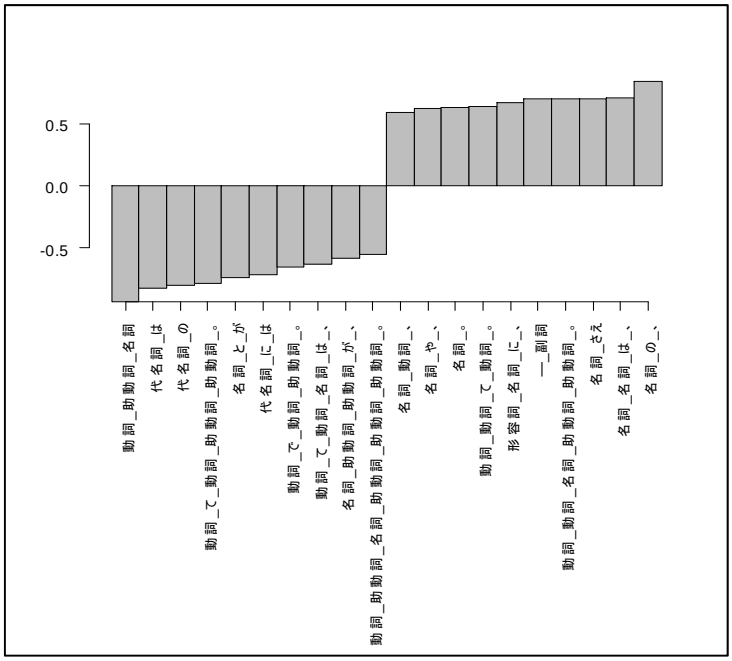

図 5.6 『コスモスの友』文字記号 bigram 変数の第 1 スコアの棒グラフ

5.3 クラスター分析の結果

『コスモスの友』の文字記号bi-gramのクラスター分析を図5.7~5.9に示す。図5.7~5.9では、川端康成と中里恒子の文章はそれぞれ一つのクラスターを形成し、『コスモスの友』は川端康成のクラスターに入っている。この結果は、『コスモスの友』の文体は川端康成に近いことを示した。

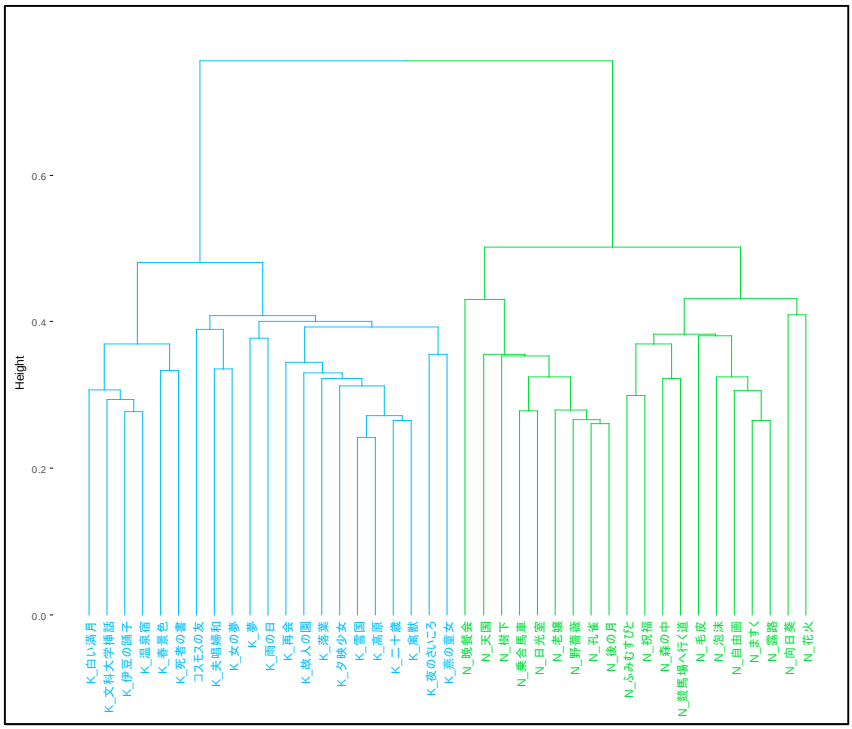

図 5.7 『コスモスの友』文字記号 bi-gram の階層的クラスター樹形図

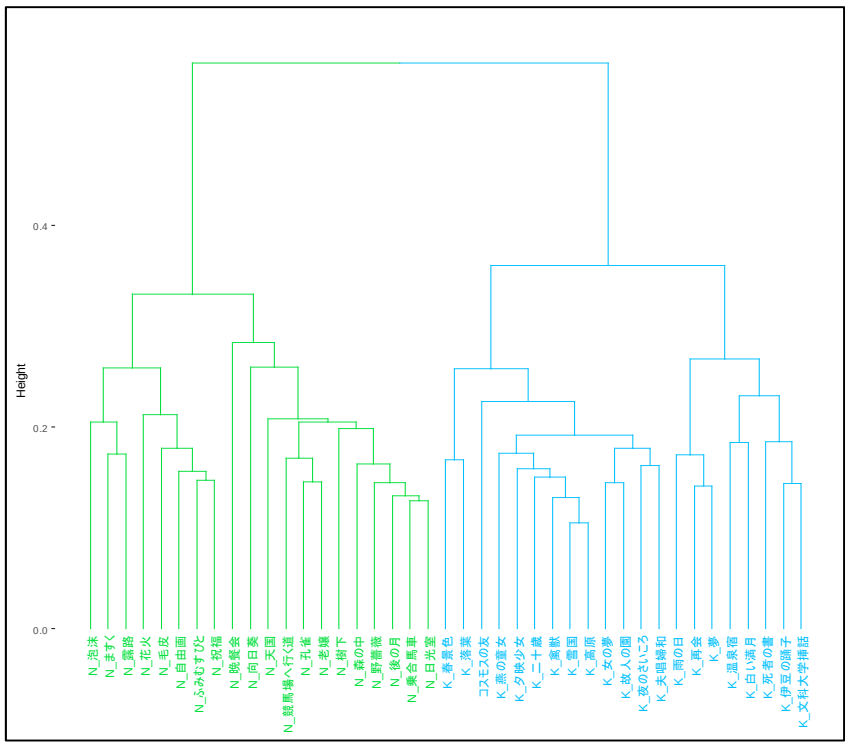

図 5.8 『コスモスの友』タグ付き形態素の階層的クラスター樹形図

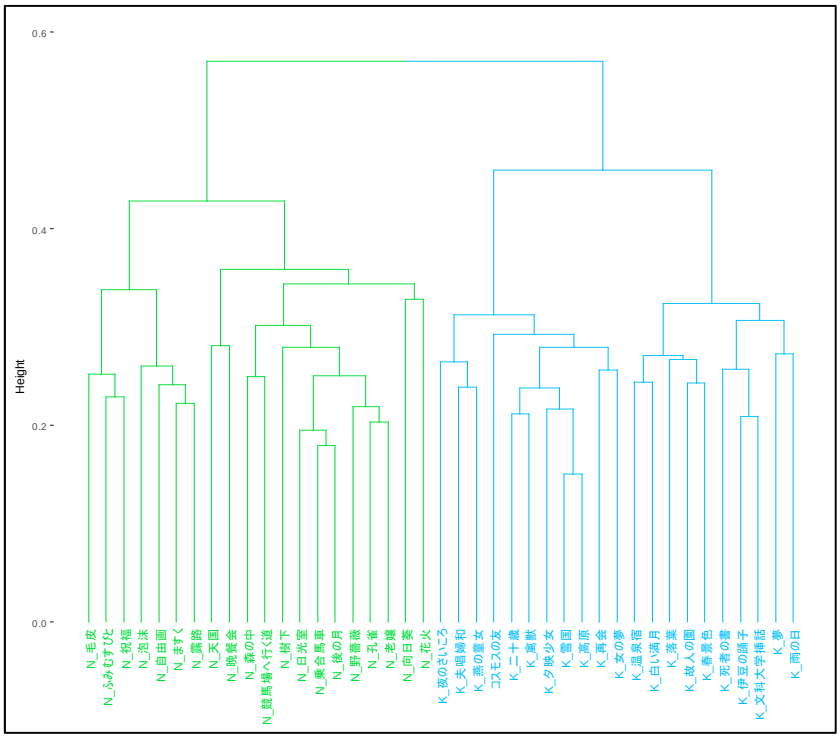

図 5.9 『コスモスの友』文節パターンの階層的クラスター樹形図

5.4 分類器による判別結果

文字記号bi-gram、タグ付き形態素と文節パターンを文体特徴量とし、AdaBoost、HDDA、LMT、RFとSVMによる『コスモスの友』の分類結果と統合結果を表5.1に示す。

表5.1『コスモスの友』の統合判別結果 (K: 川端康成、N: 中里恒子)

分類器/特徴量	AdaBoost	HDDA	LMT	RF	SVM	統合結果
文字記号bi-gram	N	K	K	K	N	K
タグ付き形態素	N	K	N	K	K	K
文節パターン	N	N	N	N	N	N

表5.1では、文字記号bi-gramの場合、AdaBoostとSVMでは中里恒子、HDDA LMTとRFでは川端康成に判別された。タグ付き形態素の場合、AdaBoostとLMTでは中里恒子、HDDA、RFとSVMでは川端康成に判別された。文節パターンではすべて川端康成に判別された。統合結果において文字記号bi-gramでは川端康成、タグ付き形態素と文節パターンでは川端康成に判別された。

5.5 本章のまとめ

『コスモスの友』は川端康成の名義で発表され、中里恒子の代筆と疑われる少女小説である。対応分析と階層的クラスター分析の結果では、『コスモスの友』の文体は川端康成に近いことが示された。また、分類器を用いた判別の結果では、文字記号bi-gramとタグ付き形態素では川端康成に、文節パターンでは中里恒子に判別された。以上の分析より、『コスモスの友』には川端康成と中里恒子の文体が混在しているが、多くの分析結果では、川端康成の文体に近い傾向が現れたので、中里恒子のと比べ、『コスモスの友』の文体は川端康成に近いため、この作品は代筆ではないと推察される。

第 6 章　『古都』の代筆問題研究

川端康成は幼少期から精神状態が不安定で、作家デビューしてから執筆のために昼夜逆転の生活を送り、その影響で精神状態がさらに悪化した。川端康成は当時不眠症に襲われ、やむを得ず睡眠薬に頼るようになった。その睡眠薬依存の最たる時期は1962年で、川端康成はこの年睡眠薬中毒の影響で入院までした。精神科医の栗原 (1982)は、「川端康成の不眠症は亡くなる直前も若干悪化していた形跡がある」と述べ、睡眠薬の影響で川端康成が自殺を図ったのではないかと推測した。このような重度の睡眠薬依存の状態にあったにもかかわらず、川端康成は『古都』や『眠れる美女』などの傑作を次々と書き上げた。その状態での執筆は不可能であると思われ、当時の川端康成名義で発表された『古都』は代筆者が書いたと主張する研究者は少なくない (板坂, 1997; 小谷野, 2013)。

6.1 研究背景

『古都』は川端康成のノーベル文学賞の受賞対象作の一つと言われ、日本だけで 3 回も映画化されるほど絶大な人気を博した作品である (富岡, 2014)。この作品の人気ぶりは日本だけでなく、『古都』は多くの外国語に翻訳され、海外でも知名度を上げている。『古都』は古都・京都を舞台に双子の姉妹の物語を描いた小説である。この小説は『朝日新聞』に昭和 36 年 (1961)10 月から昭和 37 年 (1962)1 月まで計 107 回に渡って連載され、単行本が発行された際に 9 つの章にまとめられた。各章の詳細情報を表 6.1 に示す。

表 6.1 『古都』の各章の詳細

『古都』の各章		文字数
1	春の花	6820
2	尼寺と格子	6430
3	きものの町	6544
4	北山杉	6914
5	祇園祭	9150
6	秋の色	5652
7	松のみどり	4808
8	秋深い姉妹	8860
9	冬の花	5952

川端康成は『東京の人』の執筆 (1954年)から睡眠薬を用い、『古都』の執筆期間が終わる(1962年)までの間使用していた (木幡, 1992)。『古都』の執筆が終わり、睡眠薬をやめようとした川端康成はひどい禁断症状に襲われ、10日間意識不明の状態となり東大沖中内科に入院していた。

川端康成の当時の精神状態は『古都』の内容にも影響をもたらしている。川端康成入院時の担当医であった栗原 (1982)は、「古都の一部にはふわふわと上すべりする感じがする。これはおそらく睡眠薬による影響であって、半覚半醒状態のときにあとからあとからと浮かぶ空想を、筆にしたものともいえる。」と述べた。また、山田 (1980)は次のように述べている。

> 本質的なことは、川端が『古都』という作品において、知らず知らずのうちに霊界との交感をおこなっていたということである。北山杉の村には現世と隔絶した霊界の磁場が張られ、その内奥に〈未生〉および〈死後〉の世界がひそんでいた。その霊界からあらわれたかのような苗子は、主人公千重子を北山杉の村へといざない、千重子に〈未生の時〉をかいま見せるのである。こうした現世と霊界との交感を、川端は眠り薬に侵されたうつつない薄明の世界で、何ものかに促されるように書いていったのである。

栗原 (1982)と山田 (1980)の研究は川端康成の精神状態と『古都』の内容の関連性を示した。川端康成自身もそれに気づき、『古都』の執筆中に自分の精神状態について次のように記している (川端康成全集第33巻 評論5, p. 600-602)。

> 『古都』執筆期間のいろんなことの記憶は多く失われていて、不気味なほどであった。『古都』になにを書いたかもよくはおぼえていなくて、たしかには思い出せなかった。私は毎日『古都』を書き出す前にも、書いているあいだにも、眠り薬を用いた。眠り薬に酔って、うつつないありさまで書いた。眠り薬が書かせたようなものであったろうか。『古都』は「私の異常所産」というわけである。 (中略)
>
> この作品に対する同情と慰謝によって、私は校正にとりかかった。果しておかしいところ、辻褄の合わぬようなところが少なかった。校正でだいぶん直したが、行文のみだれ、調子の狂いが、かえってこの作品の特色となっていると思えるものはそのまま残した。校正は骨が折れた。

「記憶は多く失われる」や「毎日眠り薬を用いて、うつつないありさまで『古都』を書いた」などの内容から、当時川端康成はとても執筆できるとは思えない。『古都』の連載時に使った原稿について、ある出版関係者は「あのときはもう川端さんの原稿は全く使いものにならないものばかりでした。発表されたものは全部第三者が書いたものです。つまりゴーストライターがいたんですよ」と明かしている (板坂, 1997)。また、そのゴーストライターは、川端康成に師事していた澤野久雄、北條誠と三島由紀夫であることも言及している (板坂, 1997)。そのうち最も有力な澤野久雄説について小谷野 (2013)は次のように述べた。

> 『古都』の連載が始まったのは十月八日である。しかし川端は、直前まで何も考えていなかったようで、九月末、澤野久雄に手紙を書いて、何も書くことがないと言った。京都を舞台に描くということで、京都に家を借り、連作を引き受けて、直前になってこ

れである。澤野は驚き、慌てて京都へ行くと、北山杉を川端に教え、これをモティーフにするよう示唆した。

本章では文体計量分析を通して『古都』の代筆問題を明らかにする。代筆問題解明に用いた川端康成、澤野久雄、北條誠と三島由紀夫のコーパスを表6.2に示す。

表 6.2『古都』の代筆問題解明のためのコーパス

川端康成	澤野久雄	北條誠	三島由紀夫
あの国この国	雨しきり	アカシヤの唄 1	遠乗会
たまゆら	炎上	アカシヤの唄 2	鴛鴦
たんぽぽ	遠い音	バラが咲いた	家庭裁判
みづうみ	果樹園の道	花はなんの花	果実
横町	花火	月の砂漠	海と夕焼
岩に菊	花燭	五月の風	近世姑気質
弓浦市	古調	五百マイル	携帯用
故郷	古典	山のむらさき	月
自然	孤客	秋扇	孤閨悶々
小春日	初恋	翠のころ	詩を書く少年
水月	雪譜	赤い夕焼け	修学旅行
日も月も	笛の夜	朝つゆの道 1	女流立志伝
千羽鶴	晩年の石	朝つゆの道 2	食道楽
並木	粉雪	朝つゆの道 3	雛の宿
片腕	未明	朝つゆの道 4	朝顔
雨の日	揺籃	白い夜風	日曜日
無言	落葉樹 1	風のあと	博覧会
名人	落葉樹 2	別れの曲	百万円煎餅
明月	離合	豊かなるもの	憂国
離合	聯璧	緑なる人	離宮の松

6.2 対応分析の結果

『古都』のコーパスから抽出した文字記号bi-gramの次元数は2096である。文字記号bi-gram対応分析個体の第1、2成分の散布図を図6.1に示す。図6.1では、北條誠、三島由紀夫と『朝つゆの道』を除いた澤野久雄の作品は第1スコア軸の負の方向にプロットされた。川端康成作品の大半と『古都』は第1スコア軸の正の方向にプロットされた。『古都』は川端康成の作品に最も近い。この傾向から『古都』の文体は川端康成に似ていることが分かった。

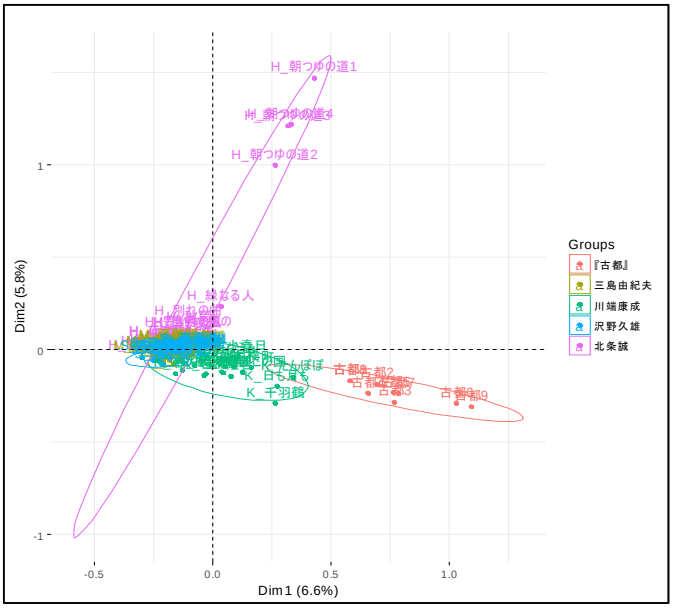

図6.1 『古都』文字記号bi-gram対応分析個体の第1、2スコアの散布図

『古都』の文字記号 bi-gram 変数の第 1 スコアの棒グラフを図 6.2 に示す。第 1 スコア軸の正の方向に現れた変数は「苗子」、「、苗」、「吉郎」、「太吉」、「重子」、「千重」、「秀男」、「郎は」、「、太」と「、秀」で、川端康成作品の一部と『古都』では、このような文字記号 bi-gram が多く用いられている。第 1 スコア軸の負の方向に現れた変数は「端さ」、「川端」、「踊子」、「尚雄」、「朝子」、「、一」、「田島」、「作家」、「更に」と「、川」で、川端康成と北條誠作品の一部、三島由紀夫と澤野久雄の作品にはこういった文字記号 bi-gram が多く用いられている。

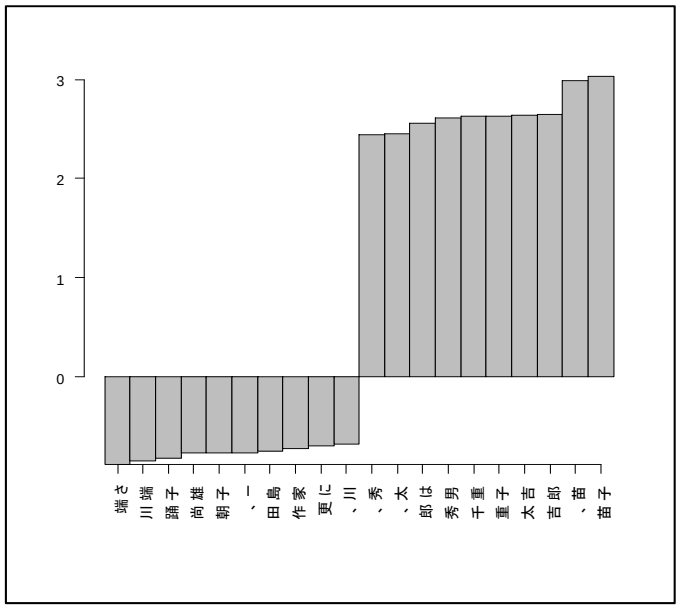

図6.2 『古都』文字記号bi-gramの変数第1スコアの棒グラフ

『古都』のコーパスから抽出したタグ付き形態素の次元数は318である。タグ付き形態素対応分析個体の第1、2成分の散布図を図6.3に示す。図6.3では、『古都』は北條誠、川端康成と澤野久雄作品の一部は原点付近に固まり、『古都』の文体はこの3人に似ていることが見て取れた。

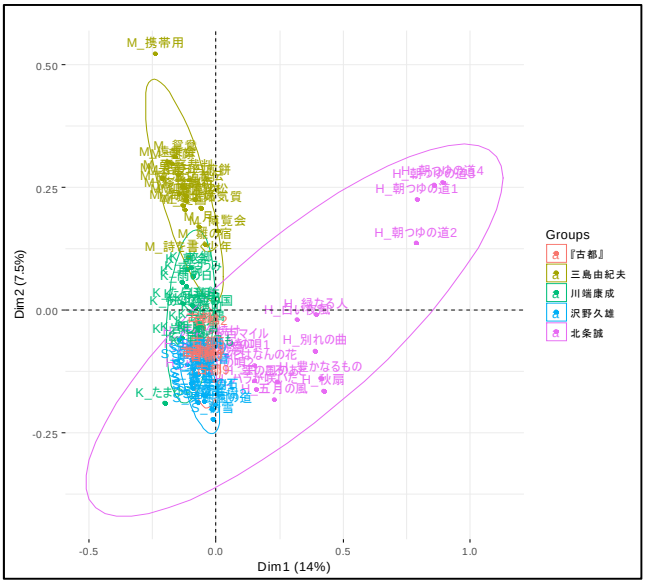

図6.3 『古都』タグ付き形態素の個体第1、2スコアの散布図

『古都』のタグ付き形態素の変数第 1 スコアの棒グラフを図 6.4 に示す。第 1 スコア軸の正の方向に現れた変数は「とう_副詞」、「あ_フィラー」、「じ_助動詞」、「す_接頭詞」、「だが_接続詞」、「…_記号」、「いくら_副詞」、「お_接頭詞」、「どうやら_副詞」と「やっと_副詞」で、北條誠の作品ではこのようなタグ付き形態素が多く用いられている。第 1 スコア軸の負の方向に現れた変数は「尚_接続詞」、「たまゆら_副詞」、「却って_副詞」、「或_連体詞」、「無_接頭詞」、「又_接続詞」、「こう_副詞」、「又_副詞」、「一方_接続詞」と「相_接頭詞」で、川端康成、澤野久雄、三島由紀夫の作品と『古都』では、こういったタグ付き形態素が多く用いられている。

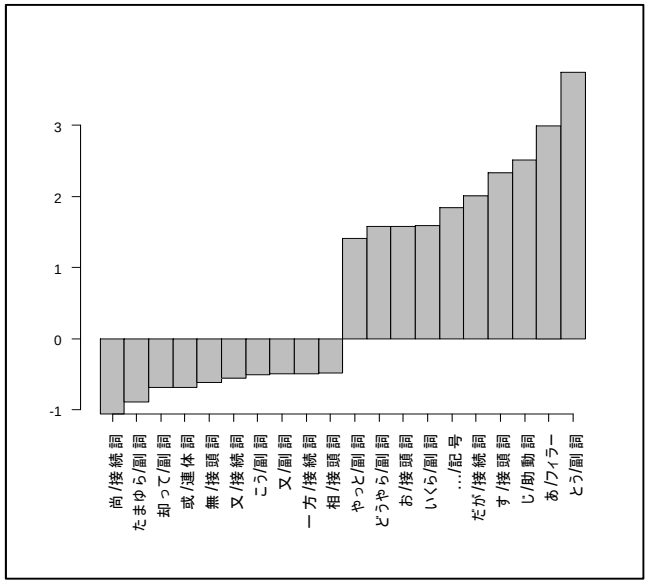

図6.4 『古都』タグ付き形態素の変数第1スコアの棒グラフ

『古都』のタグ付き形態素変数の第 2 スコアの棒グラフを図 6.5 に示す。第 2 スコア軸の正の方向に現れた変数は「尚_接続詞」、「或_連体詞」、「又_接続詞」、「こう_副詞」、「とう_副詞」、「又_副詞」、「却って_副詞」、「すこし_副詞」、「あ_フィラー」と「じ_助動詞」で、三島由紀夫、川端康成と北條誠の作品の一部では、このようなタグ付き形態素が多く用いられている。第 2 スコア軸の負の方向に現れた変数は「たまゆら_副詞」、「けれども_接続詞」、「ー_記号」、「一応_副詞」、「なるほど_感動詞」、「別に_副詞」、「が_接続詞」、「初めて_副詞」、「更に_副詞」と「時々_副詞」で、澤野久雄、『古都』、川端康成と北條誠の作品の一部では、こういったタグ付き形態素が多く用いられている。

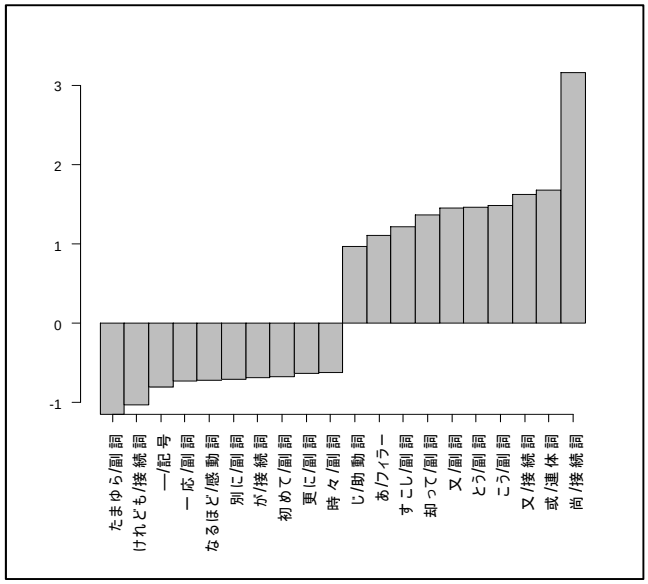

図6.5 『古都』タグ付き形態素の変数第2スコアの棒グラフ

『古都』のコーパスから抽出した文節パターンの次元数は600である。文節パターン対応分析個体の第1、2スコアの散布図を図6.6に示す。図6.6では、『古都』は第4象限、北條誠と澤野久雄作品は第1象限、三島由紀夫の作品は第2象限にそれぞれプロットされた。川端康成の作品は第1スコア軸を跨いで第2、3象限にプロットされ、位置関係では『古都』に最も近い。

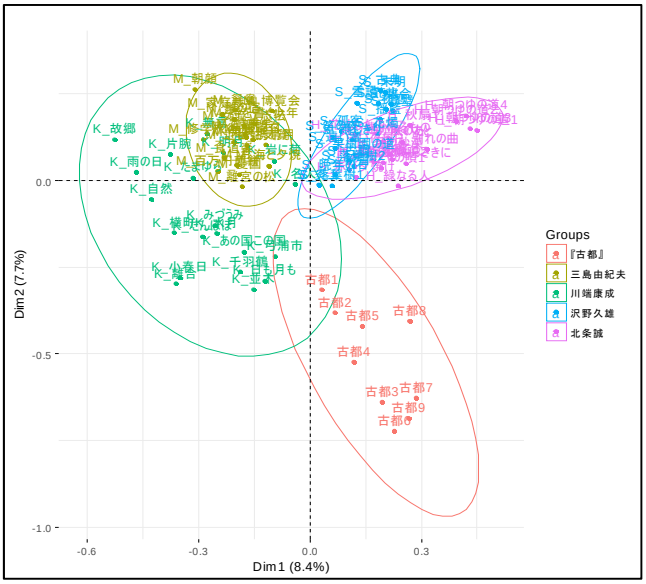

図6.6 『古都』文節パターン対応分析個体の第1、2スコアの散布図

『古都』の文節パターン変数の第 1 スコアの棒グラフを図 6.7 に示す。第 1 スコア軸の正の方向に現れた変数は「接続詞_副詞」、「形容詞_名詞_に」、「「_名詞_」_を」、「連体詞_、」、「代名詞_も_、」、「動詞_て_、_と_、」、「名詞_名詞_と_、」、「副詞_、」、「名詞_まで_、」と「「_名詞_」_の」で、澤野久雄、北條誠の作品と『古都』では、このような文節パターンが多く用いられている。第 1 スコア軸の負の方向に現れた変数は「動詞_て_動詞_て」、「形容詞_ので_、」、「名詞_と_が」、「動詞_だけ_助動詞_助動詞_。」、「動詞_動詞_名詞_に」、「動詞_助動詞_のに_、」、「動詞_て_動詞_助動詞_ので_、」、「動詞_動詞_助動詞_名詞_に」、「動詞_まで」と「助動詞_、」で、川端康成と三島由紀夫の作品では、このような文節パターンが多く用いられている。

『古都』の文節パターン変数の第 2 スコアの棒グラフを図 6.8 に示す。第 2 スコア軸の正の方向に現れた変数は「代名詞_名詞_は」、「動詞_フィラー_名詞_の」、「接頭詞_副詞」、「代名詞_に_は_、」、「「_名詞_」_の」、「「_名詞_」_を」、「代名詞_名詞_の」、「サ変接続_。」、「動詞_助動詞_名詞_かも_動詞_助動詞_。」と「「_名詞_の」で、澤野久雄、北條誠、三島由紀夫の作品と『古都』の一部では、このような文節パターンを多く用いられている。第 2 スコア軸の負の方向に現れた変数は「名詞_にたいする」、「動詞_助動詞_て_、」、「動詞_助動詞_か_、」、「名詞_と_が_、」、「動詞_て_動詞_助動詞_と_、」、「ものの_、」、「名詞_名詞_と_、」、「動詞_助動詞_で_、」、「動詞_助動詞_名詞_も_、」と「形容詞_動

詞_て_、」で、『古都』と川端康成作品の一部では、このような文節パターンが多く用いられている。

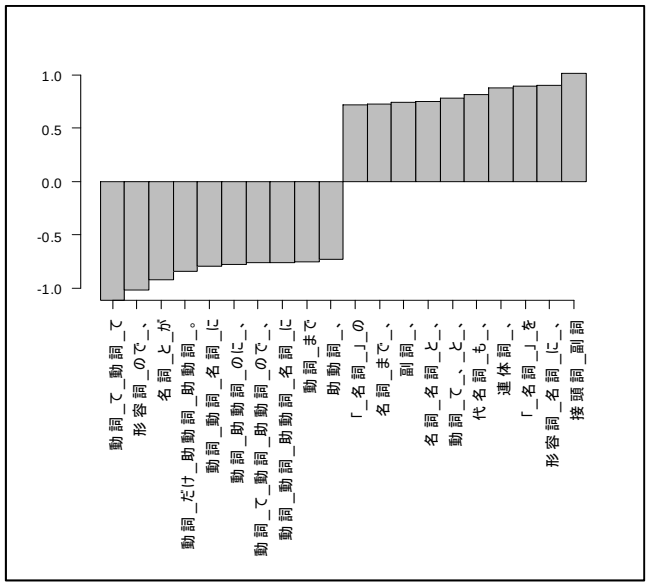

図6.7 『古都』文節パターン変数の第1スコアの棒グラフ

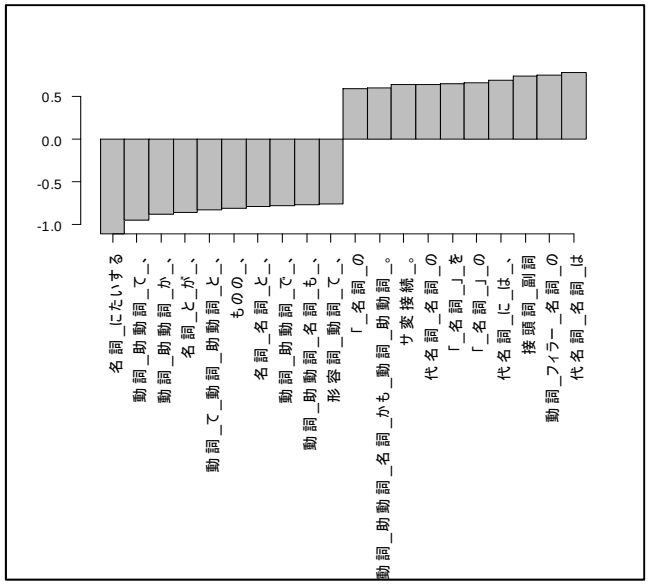

図6.8 『古都』文節パターン変数の第2スコアの棒グラフ

6.3 クラスター分析の結果

『古都』の文字記号bi-gramを用いた階層的クラスター分析の結果を図6.9に示す。作品のクラスターは、大きく北條誠、三島由紀夫、澤野久雄作品クラスターと、『古都』、川端康成クラスターの二つに分かれている。『古都』各章は川端康成の作品と同じクラスターに配置され、文字記号bi-gramでは『古都』の文体は川端康成に近いことが見て取れた。

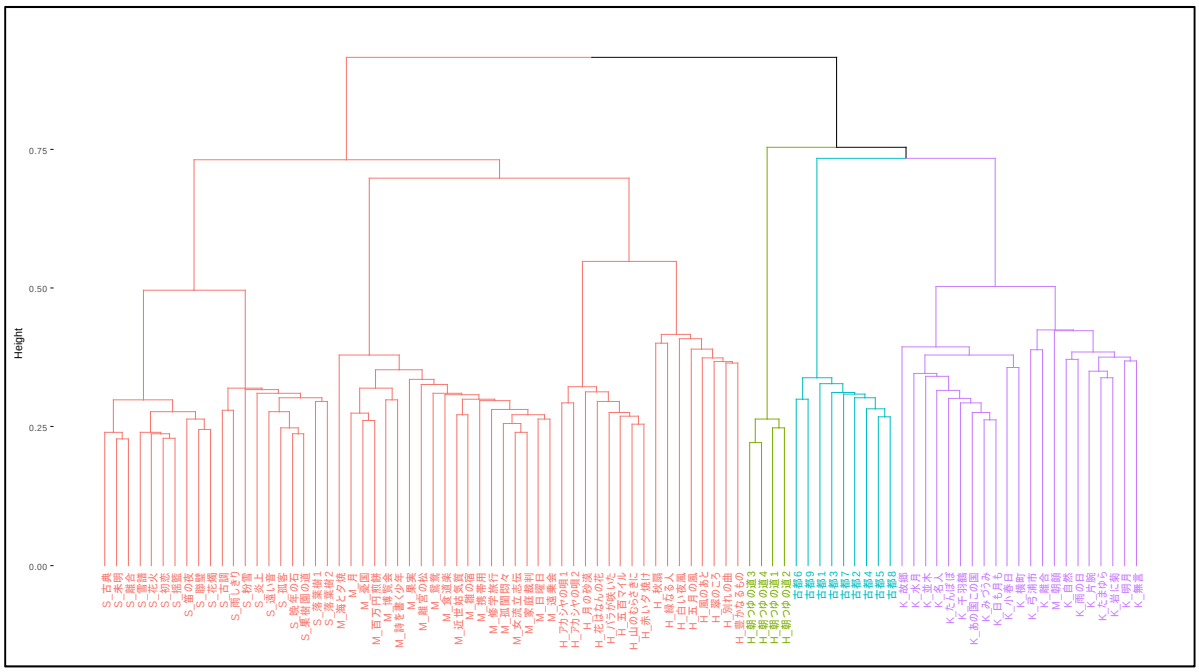

図 6.9 『古都』の文字記号 bi-gram の階層的クラスター樹形図

『古都』のタグ付き形態素を用いた階層的クラスター分析の結果を図6.10に示す。作品のクラスターは、大きく北條誠、三島由紀夫、川端康成、『古都』、澤野久雄作品クラスターの5つに分かれている。『古都』各章のクラスターには北條誠の作品が混在し、また、澤野久雄の作品クラスターとの結合も早い。タグ付き形態素では、『古都』の文体は北條誠と澤野久雄に近いことが見て取れた。

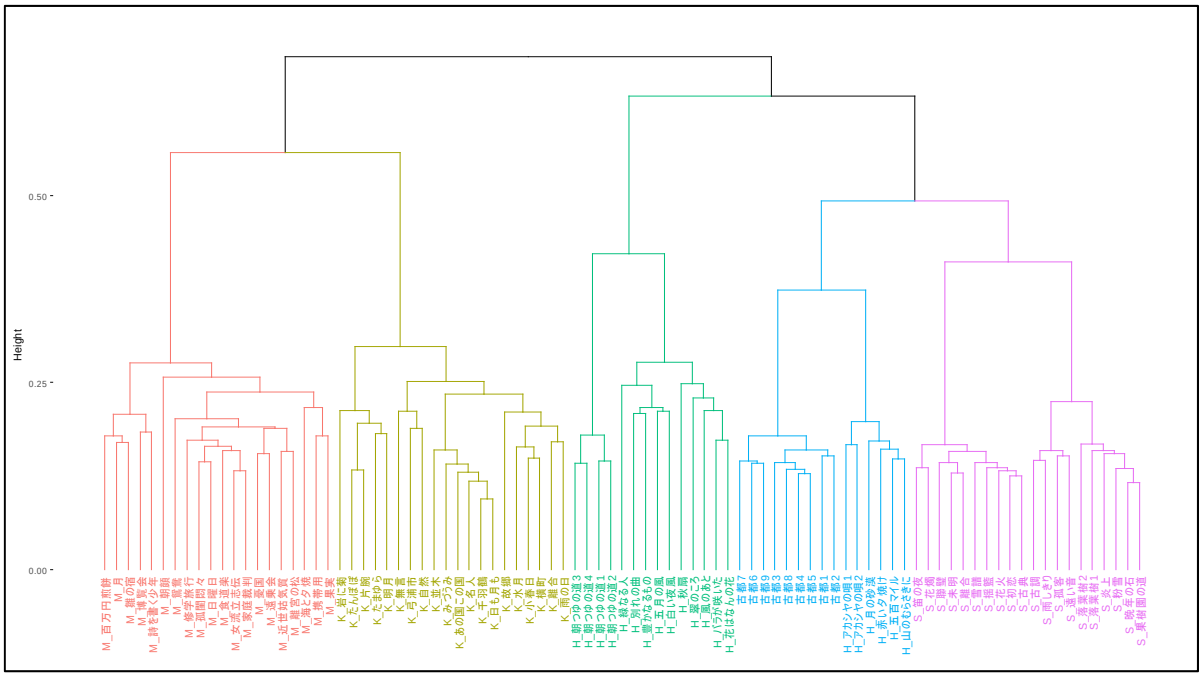

図6.10 『古都』のタグ付き形態素の階層的クラスター樹形図

『古都』の文節パターンを用いた階層的クラスター分析の結果を図6.11に示す。作品のクラスターは、大きく三島由紀夫、川端康成作品クラスターと、『古都』、北條誠、澤野久雄作品クラスターの二つに分かれている。『古都』各章は、北條誠、澤野久雄の作品と同じクラスターに配置され、文節パターンでは、『古都』の文体は北條誠と澤野久雄に近いことが見て取れた。

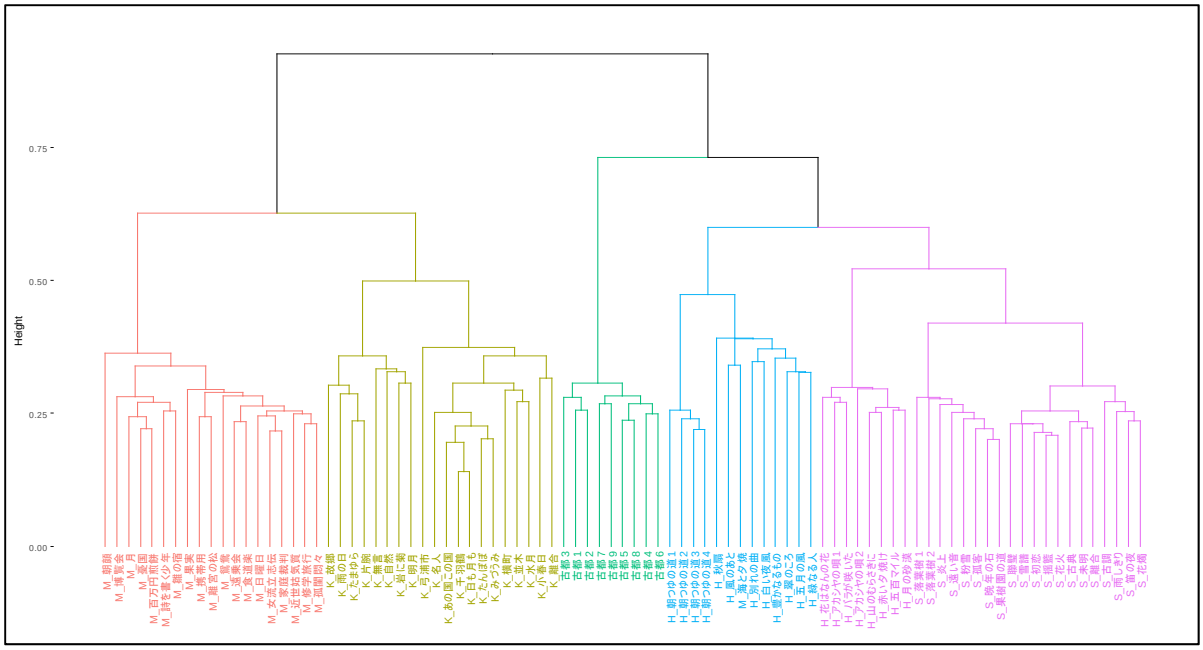

図 6.11 『古都』の文節パターンの階層的クラスター樹形図

6.4 分類器による判別結果

『古都』の分類器を用いた著者推定では、北條誠、川端康成、澤野久雄と三島由紀夫の 4 群分類を行った。分類では、文字記号 bi-gram、タグ付き形態素と文節パターンを学習データとし、AdaBoost、HDDA、LMT、RF と SVM を分類器とした。各文体特徴量における分類器の性能を示す指標を表 6.3 に示す。

6.4.1 文字記号 bi-gram

文字記号bi-gramによる『古都』の分類結果と統合結果を表6.4に示す。AdaBoost、HDDA、LMTとRFでは、すべての章が川端康成に判別された。SVMでは第9章は北條誠、第1、2、3、4、5、6、7、8章は川端康成に判別された。5つの分類器の統合結果では、『古都』の各章は川端康成に判別される結果になった。

表 6.3『古都』の各文体特徴量における分類器の性能評価

特徴量	評価指標	AdaBoost	HDDA	LMT	RF	SVM
文字記号 bi-gram	Precision	1	0.85	0.99	1	0.79
	Recall	1	0.88	0.99	1	0.79
	F-measure	1	0.86	0.99	1	0.79
タグ付き 形態素	Precision	0.96	0.96	0.99	1	1
	Recall	0.96	0.97	0.99	1	1
	F-measure	0.96	0.96	0.99	1	1
文節 パターン	Precision	0.98	0.96	0.98	1	0.95
	Recall	0.98	0.96	0.98	1	0.95
	F-measure	0.98	0.96	0.98	1	0.95

表 6.4 『古都』の文字記号 bi-gram を用いた 5 つの分類器による判別結果
(K:川端康成 S:澤野久雄 H:北條誠 M:三島由紀夫)

古都	AdaBoost	HDDA	LMT	RF	SVM	統合結果
第1章	K	K	K	K	K	K
第2章	K	K	K	K	K	K
第3章	K	K	K	K	K	K
第4章	K	K	K	K	K	K
第5章	K	K	K	K	K	K
第6章	K	K	K	K	K	K
第7章	K	K	K	K	K	K
第8章	K	K	K	K	K	K
第9章	K	K	K	K	H	K

6.4.2 タグ付き形態素

タグ付き形態素による『古都』の分類結果と統合結果を表6.5に示す。AdaBoost、HDDA、LMTとRFではすべての章が川端康成に判別された。SVMでは第1、2、3章は澤野久雄、第4、5、6、7、8、9章は北條誠に判別された。5つの分類器の統合結果では、『古都』のすべて章は川端康成に判別される結果になった。

表 6.5『古都』のタグ付き形態素を用いた 5 つの分類器による判別結果
(K:川端康成 S:澤野久雄 H:北條誠 M:三島由紀夫)

古都	AdaBoost	HDDA	LMT	RF	SVM	統合結果
第1章	K	K	K	K	S	K
第2章	K	K	K	K	S	K
第3章	K	K	K	K	S	K
第4章	K	K	K	K	H	K
第5章	K	K	K	K	H	K
第6章	K	K	K	K	H	K
第7章	K	K	K	K	H	K
第8章	K	K	K	K	H	K
第9章	K	K	K	K	H	K

6.4.3 文節パターン

文節パターンによる『古都』の分類結果と統合結果を表6.6に示す。AdaBoostでは第1、2、3、4、5、6、7、9章は北條誠、第8章は澤野久雄に判別された。HDDAでは第3、7、8章は北條誠、第1、2、4、5、6、9章は川端康成に判別された。LMTでは第3、4、6章は北條誠、第2、5、7、8章は澤野久雄、第1章は川端康成に判別された。RFでは第3、4、5、6、7、8、9章は北條誠、第1、2章は川端康成に判別された。SVMでは第7、8章は北條誠、第1、2、3、4、5、6、9章が川端康成に判別された。5つの分類器の統合結果では、『古都』の第3、4、6、7、8、9章は北條誠、第2、5章は北條誠か川端康成、第1章は川端康成に判別される結果になった。

表 6.6『古都』の文節パターンを用いた 5 つの分類器による判別結果
(K:川端康成 S:澤野久雄 H:北條誠 M:三島由紀夫)

古都	AdaBoost	HDDA	LMT	RF	SVM	統合結果
第1章	H	K	K	K	K	K
第2章	H	K	S	H	K	H/K
第3章	H	H	H	H	K	H
第4章	H	K	H	H	K	H
第5章	H	K	S	H	K	H/K
第6章	H	K	H	H	K	H
第7章	H	H	S	H	H	H
第8章	S	H	S	H	H	H
第9章	H	K	H	H	K	H

6.5 本章のまとめ

『古都』は川端康成の名義で発表され、澤野久雄、北條誠と三島由紀夫の代筆と疑われる小説である。対応分析の結果とクラスター分析の結果では、『古都』は三島由紀夫による代筆の可能性が低いことが分かった。また、5つの分類器を用いた判別分析では、文字記号bi-gramとタグ付き形態素の統合結果では、『古都』の各章はすべて川端康成に判別された。文節パターンの統合結果では、第3、4、6、7、8、9章は北條誠、第2、5章は北條誠か川端康成、第1章は川端康成に判別される結果になった。

以上の結果より、『古都』の文体と最も似ているのは川端康成で、その次は北条誠、最後は沢野久雄である。『古都』についての先行研究では、澤野久雄による代筆の可能性が大きいと言われたが、本章の結論から『古都』の文体は川端康成の作品に最も似ていることが明かとなり、また、代筆者が存在するとすれば、それは北條誠の可能性が最も大きい。

第 7 章　『眠れる美女』の代筆問題研究

7.1 研究背景

『眠れる美女』は1960年に雑誌『新潮』1月号から6月号までと、1961年1月号から11月号まで、約半年の空白期間を挟んで17回に渡って連載された作品である。その17回の詳細情報を表7.1に示す。

表7.1『眠れる美女』各回の詳細

『眠れる美女』各回	発行時間	雑誌	文字数
1	1960 年 1 月	新潮	3942
2	1960 年 2 月	新潮	4227
3	1960 年 3 月	新潮	5540
4	1960 年 4 月	新潮	3369
5	1960 年 5 月	新潮	4753
6	1960 年 6 月	新潮	2093
7	1961 年 1 月	新潮	1976
8	1961 年 2 月	新潮	2190
9	1961 年 3 月	新潮	3926
10	1961 年 4 月	新潮	1429
11	1961 年 5 月	新潮	3067
12	1961 年 6 月	新潮	2137
13	1961 年 7 月	新潮	3053
14	1961 年 8 月	新潮	2020
15	1961 年 9 月	新潮	2341
16	1961 年 10 月	新潮	2283
17	1961 年 11 月	新潮	2357

この作品は昭和37年毎日出版文化賞を受賞したが、昭和35 (1960)年『眠れる美女』が刊行となって間もない頃、川端康成は睡眠薬の禁断症状を起こし、数日間意識不明の状態となった。そのため、『眠れる美女』の執筆中に川端康成は睡眠薬の影響を受けていると先行研究では明かしている。『眠れる美女』には半年間の執筆空白期間があり、今村 (1988)は川端康成が睡眠薬中毒による入院の影響で執筆が滞ったことが一因であるとしている。河野 (1995)は、「題材や内容からすると『眠れる美女』は川端康成の妄想から生んだ小説で、睡眠薬に酔って書いたものである」と述べた。『眠れる美女』の代筆問題に関して、板坂 (1997)は、対談の中で、「『眠れる美女』は三島由紀夫の代筆である」と言及した。『眠れる美女』の原稿を

見たことがあるという安藤[3]は、原稿に書いてある字は川端康成の字ではないと述べた (板坂・鈴木, 2010)。以上の代筆についての諸説に対して、小谷野 (2013)は、著作『川端康成－双面の人』の中で、「『眠れる美女』の代筆はありえない」と代筆説を否定した。このように、『眠れる美女』の代筆問題に関して先行研究では意見が分かれている。本章では、文体解析の観点から『眠れる美女』の代筆問題の解明を試みる。『眠れる美女』は、当初17回にわたって連載され、その後5つの章にまとめられた。毎回の連載の分量は400字詰めの原稿用紙10枚程度で、会話文を除いても文体分析に耐えるデータ量が得られると思われるため、本章では、『眠れる美女』を17回に分けて考察することにした。

『眠れる美女』の代筆問題を検証するに当り、川端康成と三島由紀夫の全集からそれぞれ20編の小説を選んでコーパスを作成した。選ばれた小説のリストを表7.2と7.3に示す。

表 7.2 川端康成作品コーパス

発表時期	作品	文字数
1949 年	千羽鶴	37439
1950 年	雨の日	8330
1951 年	たまゆら	7939
1952 年	岩に菊	5251
1952 年	自然	3079
1952 年	名月	3456
1953 年	無言	3912
1953 年	日も月も	57050
1953 年	水月	6609
1954 年	名人	66496
1954 年	みづうみ	54152
1954 年	横町	5408
1954 年	離合	2927
1954 年	小春日	3871
1955 年	故郷	4146
1956 年	あの国この国	16095
1958 年	弓浦市	2419
1958 年	並木	5012
1964 年	片腕	13864
1968 年	たんぽぽ	30515

表 7.3 三島由紀夫作品コーパス

発表時期	作品	文字数
1950 年	果実	5571
1950 年	日曜日	8866
1950 年	孤閨悶々	11273
1950 年	食道楽	7613
1950 年	鴛鴦	5478
1951 年	家庭裁判	12226
1951 年	携帯用	10908
1951 年	女流立志伝	11853
1951 年	遠乗会	9549
1951 年	離宮の松	9276
1951 年	朝顔	2802
1952 年	近世姑気質	8460
1953 年	修学旅行	8488
1953 年	雛の宿	10769
1953 年	詩を書く少年	8387
1953 年	博覧会	7422
1954 年	海と夕焼	3641
1960 年	百万円煎餅	8317
1961 年	憂国	16206
1962 年	月	10267

[3] 安藤武：三島由紀夫の研究者である。

7.2 対応分析の結果

『眠れる美女』のコーパスから抽出した文字記号bi-gramの次元数は2037である。文字記号bi-gram対応分析個体の第1、2成分のスコアを図7.1に示す。図7.1では、川端康成作品の大多数は第2象限にプロットされた。三島由紀夫の作品は第3象限にプロットされた。『眠れる美女』の各回は両作品群と離れた第1と第2象限にプロットされた。図7.1では『千羽鶴』は川端康成の作品群と遠く離れたところにプロットされた。これは、『千羽鶴』がお茶についての小説で、「茶碗」や「茶室」などのお茶関連の用語はほかの作品より圧倒的に多く用いられた影響と考えられる。

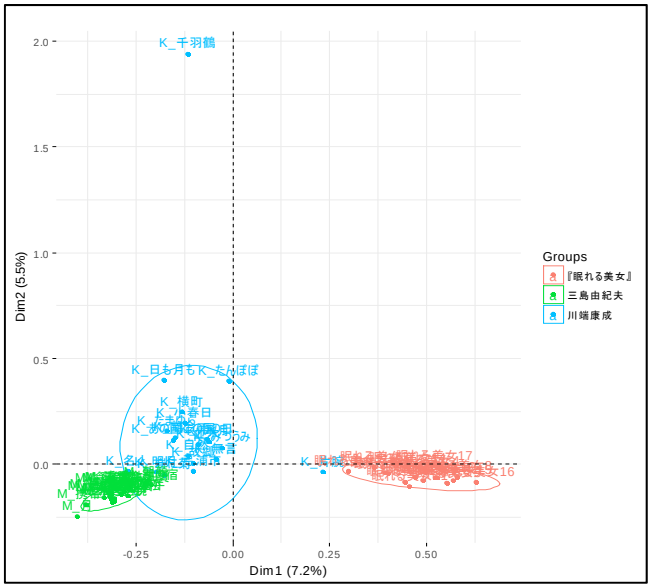

図 7.1 『眠れる美女』文字記号 bi-gram の個体第 1、2 スコアの散布図

『眠れる美女』の文字記号 bi-gram 変数の第 1 スコアの棒グラフを図 7.2 に示す。第 1 スコア軸の正の方向に「い娘」、「。老」、「口老」、「。江」、「眠ら」、「江口」、「と江」、「口は」、「老人」と「眠り」の変数が現れた。『眠れる美女』はこのような変数が多く用いられている。第 1 スコア軸の負の方向に「ミナ」、「—子」、「イミ」、「ハイ」、「ナー」、「タア」、「ピー」、「キー」、「—ラ」と「。尚」の変数が現れた。川端康成と三島由紀夫の作品では、このような変数が多く用いられている。

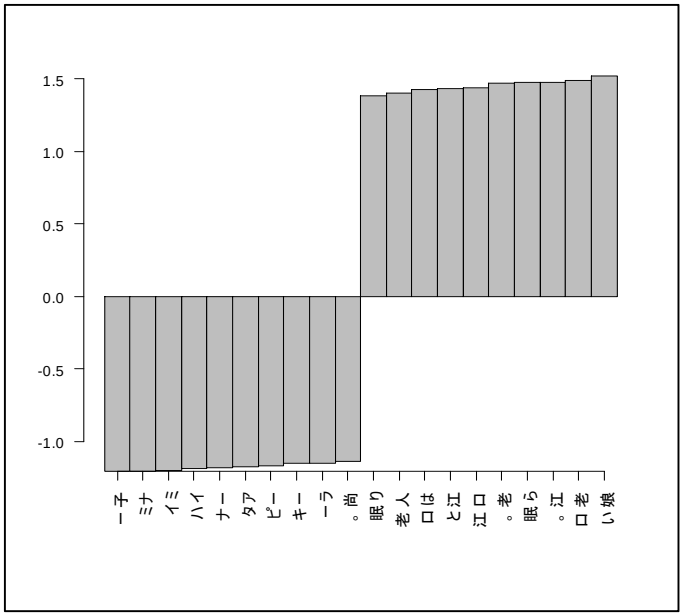

図7.2 『眠れる美女』文字記号bi-gramの第1スコアの棒グラフ

『眠れる美女』のコーパスから抽出したタグ付き形態素の次元数は222である。タグ付き形態素対応分析個体の第1、2成分のスコアを図7.3に示す。図7.3では、三島由紀夫の作品は第1スコア軸の正の方向にプロットされた。川端康成の作品と『眠れる美女』の各回は第2スコア軸の負の方向にプロットされた。この結果から『眠れる美女』の各回の文体は三島由紀夫より川端康成に近いことが見て取れた。川端康成の『たまゆら』という作品は川端康成の作品群から遠く離れたのは、「たまゆら_副詞」という変数の影響と考えられる。

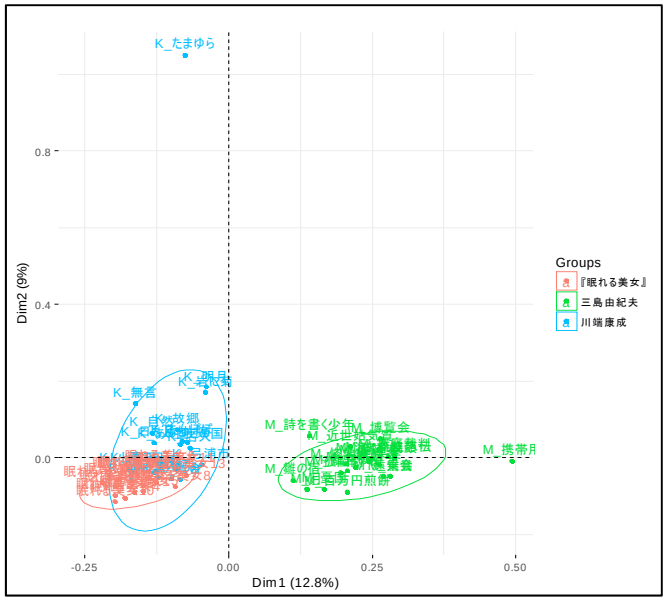

図 7.3 『眠れる美女』タグ付き形態素対応分析個体の第 1、2 スコアの散布図

『眠れる美女』のタグ付き形態素の変数第1スコアの棒グラフを図7.4に示す。第1スコア軸の正の方向に「尚_接続詞」、「或_連体詞」、「こう_副詞」、「まず_副詞」、「すこし_副詞」、「又_接続詞」、「に対する_助詞」、「小さな_連体詞」、「丁度_副詞」と「又_副詞」の変数が現れた。三島由紀夫の作品では、このような変数が多く用いられている。第1スコア軸の負の方向に「いや_接続詞」、「あ_フィラー」、「にたいする_助詞」、「なにか_副詞」、「生き生き_副詞」、「初めて_副詞」、「たかっ_助動詞」、「なお_接続詞」、「なかろ_助動詞」と「直ぐ_副詞」の変数が現れた。『眠れる美女』と川端康成の作品では、このような変数が多く用いられている。

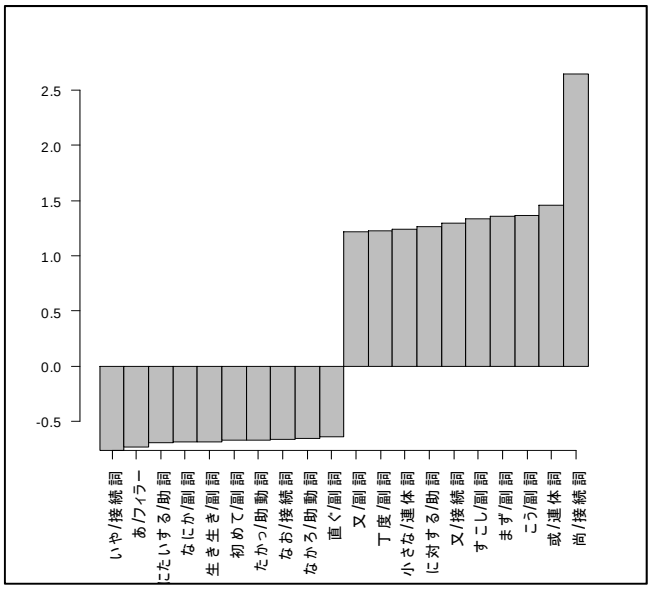

図7.4 『眠れる美女』タグ付き形態素の変数第1スコアの棒グラフ

『眠れる美女』のコーパスから抽出した文節パターンの次元数は424である。文節パターン対応分析個体の第1、2成分のスコアを図7.5に示す。図7.5に示したように、川端康成の作品は第2、3、4象限にプロットされた。三島由紀夫の作品は川端康成の作品と離れ、第1スコア軸を跨いで第1と第2象限にプロットされた。『眠れる美女』の各回は川端康成の作品と重なってプロットされた。この結果から、三島由紀夫より『眠れる美女』の各回の文体川端康成に近いことが見て取れた。

『眠れる美女』の文節パターンの第1スコアの棒グラフを図7.6に示す。第1スコアの正の方向に「連体詞_助動詞」、「接頭詞_名詞_助動詞」、「名詞_や_、」、「動詞_助動詞_に_、」、「代名詞_名詞_の」、「サ変接続_名詞_助動詞」、「動詞_助動詞_は_動詞_助動詞_。」、「サ変接続_動詞_て_、」、「動詞_て_動詞_助動詞_ので_、」と「名詞_名詞_名詞_と」の変数が現れた。三島由紀夫の作品と川端康成作品の一部では、このような変数を多く用いている。第1スコアの負の方向に「動詞_名詞_助動詞_助動詞_か_。」、「名詞_にたいする」、「名詞_の_名詞_助動詞_助動詞_。」、「連体詞_名詞_助動詞」、「動詞_助動詞_名詞_か_、」、「名詞_か_と」、「動詞_助動詞_で_、」、「名詞_など」、「名詞_は_接続詞」と「動詞_も」の

変数が現れた。『眠れる美女』と川端康成作品の一部では、このような変数が多く用いられている。

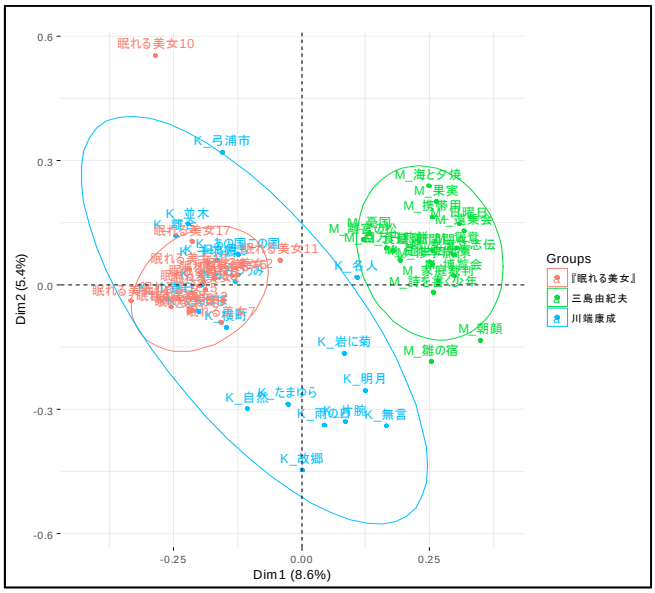

図 7.5 『眠れる美女』文節パターン対応分析個体第 1、2 スコア散布図

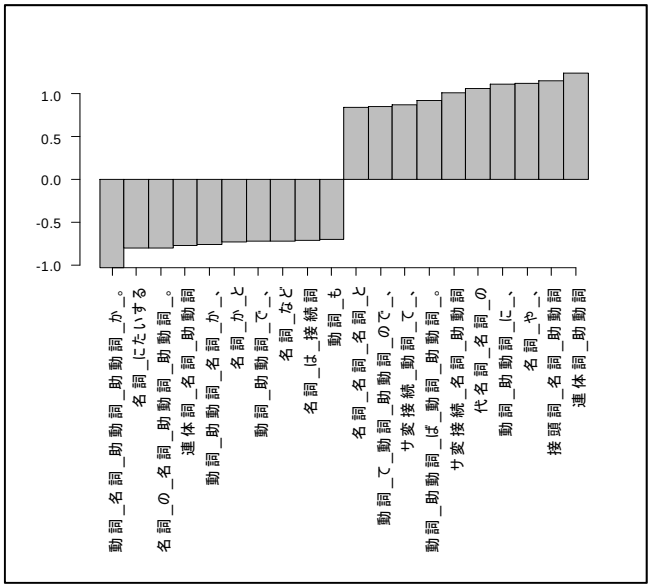

図7.6 『眠れる美女』タグ付き形態素の第1スコアの棒グラフ

7.3 クラスター分析の結果

『眠れる美女』の文字記号bi-gramを用いた階層的クラスター分析の結果を図7.7に示す。作品のクラスターは、大きく『眠れる美女』クラスターと、川端康成、三島由紀夫の作品クラスターの二つに分かれている。図7.7では、『眠れる美女』の各回は川端康成と三島由紀夫の作

品クラスターと離れているため、文字記号のbigramでは、川端康成と三島由紀夫の文体とは異なることが見て取れた。

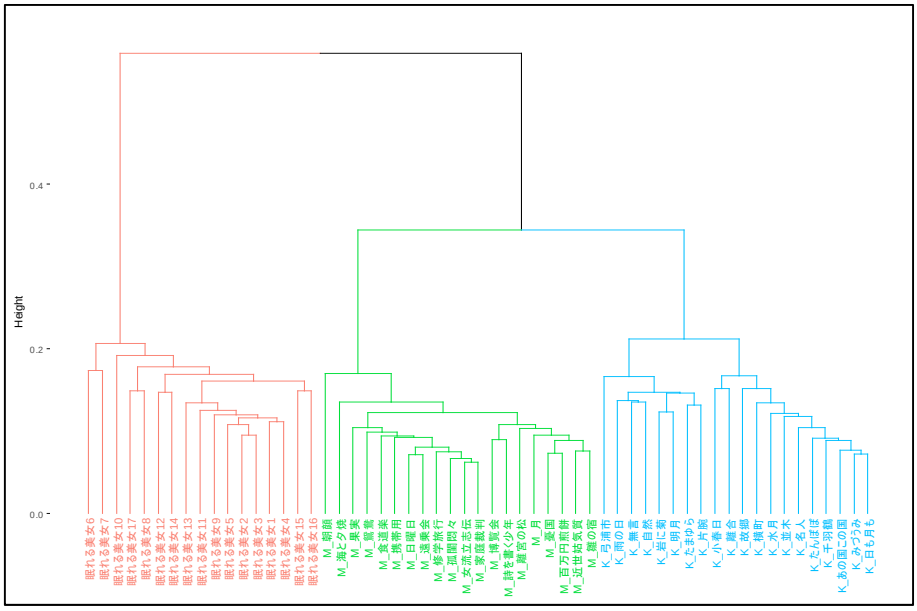

図7.7 『眠れる美女』の文字記号bi-gramの階層的クラスター樹形図

『眠れる美女』のタグ付き形態素を用いた階層的クラスター分析の結果を図7.8に示す。作品のクラスターは、大きく三島由紀夫クラスターと、『眠れる美女』各回、川端康成作品クラスターの二つに分かれている。『眠れる美女』の各回は川端康成のクラスターの近くに配置され、タグ付き形態素では、『眠れる美女』の文体は川端康成に近いことが見て取れた。

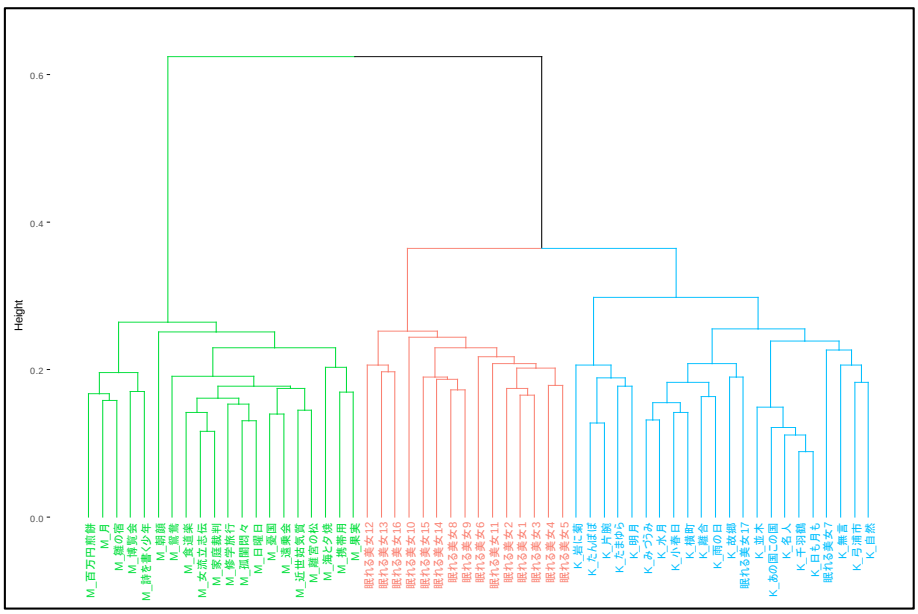

図 7.8 『眠れる美女』のタグ付き形態素の階層的クラスター樹形図

『眠れる美女』の文節パターンを用いた階層的クラスター分析の結果を図7.9に示す。作品のクラスターは大きく二つに分かれているが、一つには川端康成作品の一部と『眠れる美女』の各回、もう一つには川端康成作品の一部と三島由紀夫の作品が入っている。『眠れる美女』の各回は川端康成のクラスターの近くに配置され、文節パターンでは、『眠れる美女』の文体は川端康成に近いことが見て取れた。

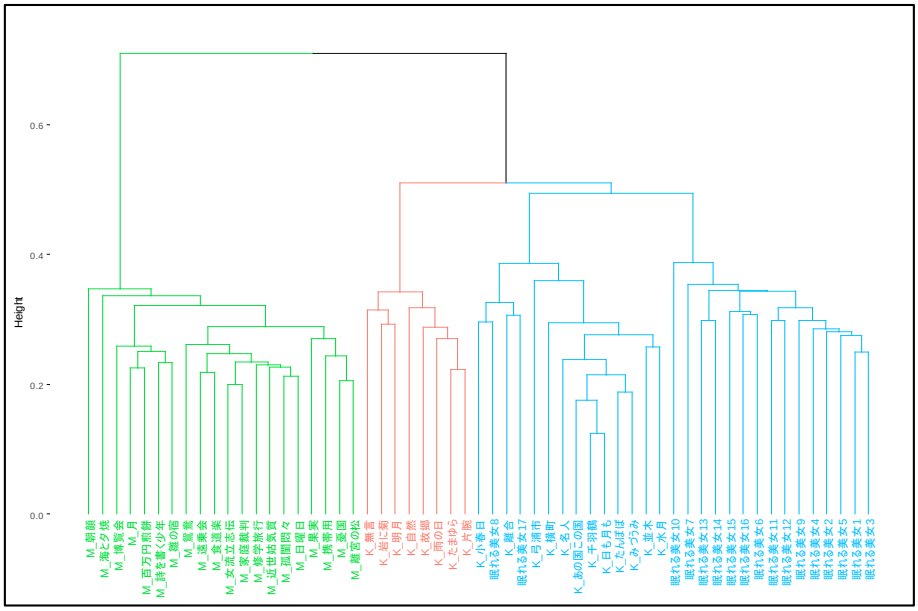

図 7.9 『眠れる美女』の文節パターンの階層的クラスター樹形図

7.4 分類器による判別結果

本節では、『眠れる美女』の判別結果を紹介する。各特徴量における分類器の性能を示す指標を表 7.4 に示す。

表7.4 各特徴データにおける分類器の性能評価

特徴量	評価指標	AdaBoost	HDDA	LMT	RF	SVM
文字記号 bi-gram	Precision	0.95	0.9	1	1	0.8
	Recall	0.95	1	1	1	0.8
	F-measure	0.95	0.95	1	1	0.8
タグ付き形態素	Precision	0.95	0.95	0.95	1	1
	Recall	0.95	1	0.95	1	0.95
	F-measure	0.95	0.98	0.95	1	0.97
文節パターン	Precision	1	1	0.86	1	0.95
	Recall	1	0.95	0.95	1	1
	F-measure	1	0.97	0.9	1	0.98

7.4.1 文字記号 bi-gram

文字記号bi-gramにおける『眠れる美女』の分類結果と統合結果を表7.5に示す。AdaBoostでは第9回は三島由紀夫、第1、2、3、4、5、6、7、8、10、11、12、13、14、15、16、17回は川端康成に判別された。HDDA、LMT、RFとSVMではすべての回が川端康成に判別された。5つの分類器の統合結果では、『眠れる美女』の各回は川端康成に判別される結果になった。

表7.5 『眠れる美女』の文字記号bi-gramを用いた5つの分類器による判別結果
(K: 川端康成 M: 三島由紀夫)

『眠れる美女』各回	AdaBoost	HDDA	LMT	RF	SVM	統合結果
1	K	K	K	K	K	K
2	K	K	K	K	K	K
3	K	K	K	K	K	K
4	K	K	K	K	K	K
5	K	K	K	K	K	K
6	K	K	K	K	K	K
7	K	K	K	K	K	K
8	K	K	K	K	K	K
9	M	K	K	K	K	K
10	K	K	K	K	K	K
11	K	K	K	K	K	K
12	K	K	K	K	K	K
13	K	K	K	K	K	K
14	K	K	K	K	K	K
15	K	K	K	K	K	K
16	K	K	K	K	K	K
17	K	K	K	K	K	K

7.4.2 タグ付き形態素

タグ付き形態素における『眠れる美女』の分類結果と統合結果を表7.6に示す。AdaBoostでは第1、2、3、5、7、9、10、11、12、13、15、17回は川端康成、第4、6、8、14、16回は三島由紀夫に判別された。HDDA、RFとSVMではすべての回が川端康成に判別された。LMTでは、第1、2、3、4、5、6、7、9、10、11、12、13、15、16、17回は川端康成、第8回は三島由紀夫に判別された。5つの分類器の統合結果では、『眠れる美女』の各章は川端康成に判別される結果になった。

表7.6『眠れる美女』のタグ付き形態素を用いた5つの分類器による判別結果
(K: 川端康成 M: 三島由紀夫)

『眠れる美女』各回	AdaBoost	HDDA	LMT	SVM	RF	統合結果
1	K	K	K	K	K	K
2	K	K	K	K	K	K
3	K	K	K	K	K	K
4	M	K	K	K	K	K
5	K	K	K	K	K	K
6	M	K	K	K	K	K
7	K	K	K	K	K	K
8	M	K	M	K	K	K
9	K	K	K	K	K	K
10	K	K	K	K	K	K
11	K	K	K	K	K	K
12	K	K	K	K	K	K
13	K	K	K	K	K	K
14	M	K	K	K	K	K
15	K	K	K	K	K	K
16	M	K	K	K	K	K
17	K	K	K	K	K	K

7.4.3 文節パターン

文節パターンにおける『眠れる美女』の分類結果と統合結果を表7.7に示す。AdaBoostでは、第3、4、5、6、7、8、9、10、11、12、13、14、15、16、17回は川端康成、第1、2回は三島由紀夫に判別された。HDDA、LMT、RFとSVMでは、すべての回が川端康成に判別された。5つの分類器の統合結果では、『眠れる美女』の各章は川端康成に判別される結果になった。

表7.7『眠れる美女』の文節パターンを用いた5つの分類器による判別結果
(K: 川端康成 M: 三島由紀夫)

『眠れる美女』各回	AdaBoost	HDDA	LMT	SVM	RF	統合結果
1	M	K	K	K	K	K
2	M	K	K	K	K	K
3	K	K	K	K	K	K
4	K	K	K	K	K	K
5	K	K	K	K	K	K
6	K	K	K	K	K	K
7	K	K	K	K	K	K
8	K	K	K	K	K	K
9	K	K	K	K	K	K
10	K	K	K	K	K	K
11	K	K	K	K	K	K
12	K	K	K	K	K	K
13	K	K	K	K	K	K
14	K	K	K	K	K	K
15	K	K	K	K	K	K
16	K	K	K	K	K	K
17	K	K	K	K	K	K

7.5 本章のまとめ

『眠れる美女』は川端康成の名義で発表され、三島由紀夫の代筆と疑われる小説である。タグ付き形態素と文節パターンを用いた『眠れる美女』の対応分析とクラスター分析の結果では、『眠れる美女』の各回は三島由紀夫より川端康成に近いケースが多い。5つの分類器を用いた判別の場合、各文体特徴量と分類器の統合結果では、『眠れる美女』の各回はすべて川端康成に判別された。以上の分析に基づき、『眠れる美女』は三島由紀夫による代筆の可能性は非常に低いという結論に至った。

第 8 章　『山の音』の代筆問題研究

川端康成の代筆問題の中で、第3~5章で取り上げた少女小説と第6~7章で取り上げた睡眠薬中毒時期の小説のほかに、名作『山の音』にも代筆疑惑がもたれている。

8.1 研究背景

『山の音』は戦後日本文学の最高峰とされた川端康成の長編小説で、第7回野間文芸賞の受賞作でもある。『山の音』の各章は、1949年から1954年にかけて複数の雑誌に断続的に発表された。川端康成は51歳の時にこの作品を書き始め、ちょうど『山の音』の執筆の前半に新潮社から『川端康成全集』全16巻が刊行された。川端康成は刊行の「あとがき」に亡くなった友であった片岡鉄兵、横光利一と菊池寛に哀悼の意を述べ、50歳は自分の生涯の谷であると述べた。完結版の『山の音』の各章の詳細情報を表8.1にまとめる。発表当時は17章からなる小説であったが、その後の改版で16章にまとめられた。また、全集を編集する際に初出各章の改正が行われ、改正前各章のタイトルを括弧の中に示す。

表8.1　『山の音』各章の詳細

『山の音』の各章		発行日	発行先雑誌	文字数
1	山の音	1949 年 9 月	『改造文藝』第 1 巻第 3 号	6538
2	日まわり (蝉の羽)	1949 年 10 月	『群像』第 46 巻第 1 号	7663
3	雲の炎	1949 年 10 月	『新潮』第 46 巻第 10 号	3698
4	栗の実	1949 年 12 月	『世界春秋』第 1 巻第 2 号	7560
5	女の家 (栗の実の続き)	1950 年 1 月	『世界春秋』第 2 巻第 1 号	
6	島の夢	1950 年 5 月	『改造』第 31 巻第 4 号	6600
7	冬の桜	1950 年 10 月	『新潮』第 47 巻第 5 号	5287
8	朝の水	1951 年 10 月	『文學界』第 5 巻第 10 号	3952
9	夜の聲	1951 年 3 月	『群像』第 7 巻第 3 号	6341
10	春の鐘	1951 年 10 月	『別冊文藝春秋』第 28 号	5799
11	鳥の家	1952 年 10 月	『新潮』第 10 号	4197
12	傷の後	1952 年 12 月	『別冊文藝春秋』第 31 号	5611
13	都の苑	1953 年 4 月	『新潮』第 50 巻第 1 号	6693
14	雨の中	1953 年 4 月	『改造』第 34 巻第 4 号	3965
15	蚊の夢 (蚊の群)	1953 年 10 月	『別冊文藝春秋』第 33 号	4083
16	蛇の卵	1953 年 10 月	『別冊文藝春秋』第 36 号	5524
17	鳩の音 (秋の魚)	1954 年 4 月	『オール讀物』第 9 巻第 4 号	5743

川端康成の他の小説と比べ、『山の音』は異常に長く、しかも中断も挟んで5年間をかけて複数の雑誌に発表されたことで代筆疑惑を持たれている。『山の音』の代筆問題をめぐり、三島由紀夫の妻は「『山の音』は自分の旦那の代筆です。」と証言している (板坂・鈴木, 2010)。これに対して、小谷野 (2013)は「『山の音』の代筆はありえない」と代筆説を否定した。本節では、このよう対立した先行研究に基づき、『山の音』の代筆問題を解明する。

『山の音』の代筆問題を検証するに当り、川端康成と三島由紀夫の全集からそれぞれ20編の小説を選んでコーパスを作成した。選ばれた小説のリストを表8.2と8.3に示す。

表 8.2 川端康成作品コーパス

発表時期	作品	文字数
1947 年	夢	3806
1948 年	再婚者	39764
1949 年	少年	16425
1949 年	千羽鶴	37439
1950 年	岩に菊	5251
1950 年	北の海から	8330
1951 年	舞姫	59050
1951 年	たまゆら	7939
1951 年	虹いくたび	53708
1952 年	自然	3079
1952 年	名月	3456
1953 年	無言	3912
1953 年	日も月も	50750
1953 年	水月	6609
1954 年	名人	66469
1954 年	みづうみ	54152
1954 年	横町	5408
1954 年	離合	2927
1954 年	小春日	3871
1955 年	故郷	4146

表 8.3 三島由紀夫作品コーパス

発表時期	作品	文字数
1950 年	果実	5571
1950 年	日曜日	8866
1950 年	孤閨悶々	11273
1950 年	食道楽	7613
1950 年	鴛鴦	5478
1951 年	家庭裁判	12226
1951 年	携帯用	10908
1951 年	女流立志伝	11853
1951 年	遠乗会	9549
1951 年	離宮の松	9276
1951 年	朝顔	2802
1952 年	近世姑気質	8460
1953 年	修学旅行	8488
1953 年	雛の宿	10769
1953 年	詩を書く少年	8387
1953 年	博覧会	7422
1954 年	海と夕焼	3641
1960 年	百万円煎餅	8317
1961 年	憂国	16206
1962 年	月	10267

8.2 対応分析の結果

『山の音』のコーパスから抽出した文字記号bi-gramの次元数は2004である。文字記号bi-gram対応分析個体の第1、2成分のスコアを図8.1に示す。図8.1では、川端康成の作品は第1と第2象限にプロットされた。川端康成作品グループ三島由紀夫の作品は第1スコア軸を跨ぎ、

第2と第3象限にプロットされた。『山の音』の各章は両者と大きく離れた第4象限にプロットされた。

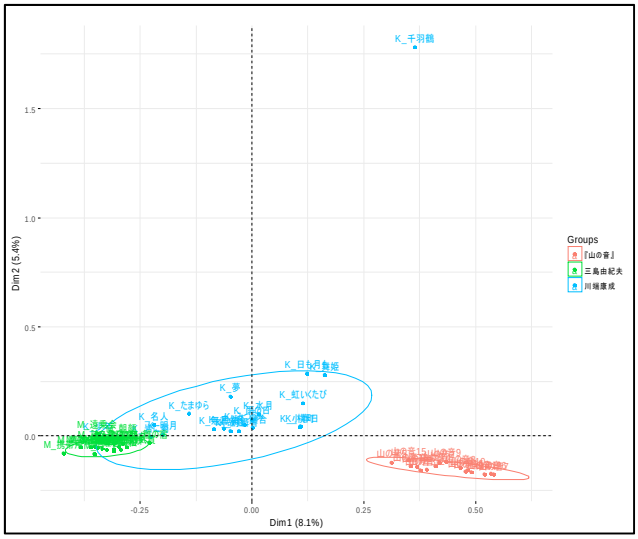

図 8.1 『山の音』文字記号 bi-gram の個体第 1、2 スコア散布図

『山の音』文字記号bigram変数の第1スコアの棒グラフを図8.2に示す。第1スコア軸の正の方向に「英子」、「里子」、「。英」、「吾が」、「、信」、「修一」、「菊子」、「吾に」、「信吾」と「吾は」の変数が現れた。川端康成作品の一部と『山の音』では、このような変数が多く用いられている。第1スコア軸の負の方向に「朝子」、「尚雄」、「合子」、「右近」、「麗子」、「尉は」、「中尉」、「ハイ」、「ナー」と「タア」の変数が現れた。川端康成作品の一部と三島由紀夫の作品では、こういった変数が多く用いられている。

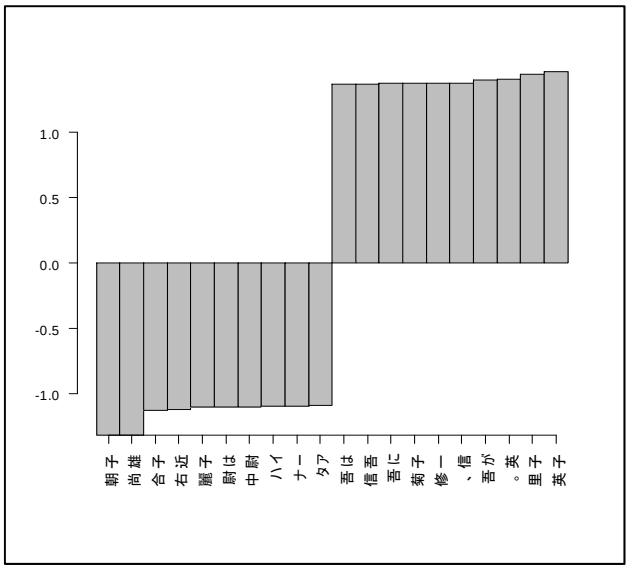

図 8.2 『山の音』文字記号 bigram 変数の第 1 スコアの棒グラフ

『山の音』のコーパスから抽出したタグ付き形態素の次元数は244である。タグ付き形態素対応分析個体の第1、2成分のスコアを図8.3に示す。図8.3では、川端康成の作品は第1、第2と第3象限にプロットされた。三島由紀夫の作品は第1と第4象限にプロットされた。『山の音』の各章は第2と第3象限の川端康成の作品と重なっているところにプロットされた。このような傾向は、三島由紀夫と比べ、『山の音』の文体は川端康成に近いことを示した。

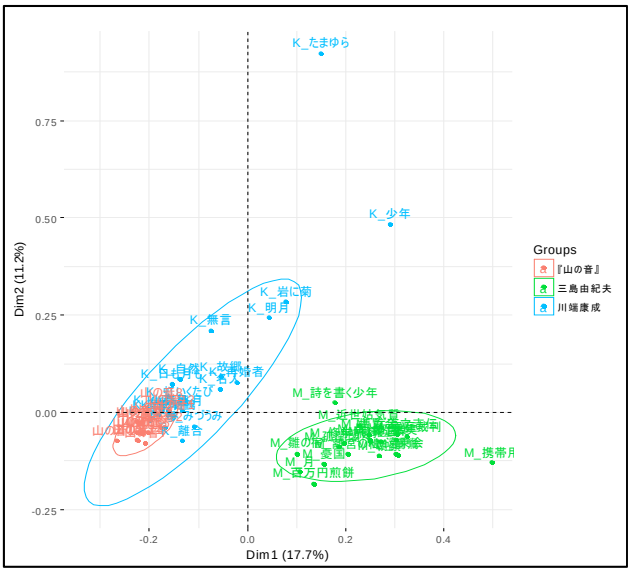

図 8.3 『山の音』のタグ付き形態素の個体の第 1、2 スコアの散布図

『山の音』の文字記号bigram変数の第1スコアの棒グラフを図8.4に示す。第1スコア軸の正の方向に「尚_接続詞」、「或_連体詞」、「こう_副詞」、「却って_副詞」、「まず_副詞」、「すこし_副詞」、「すでに_副詞」、「さらに_副詞」、「すると_接続詞」と「決して_副詞」の変数が現れた。川端康成作品の一部と三島由紀夫の作品ではこのような変数が多く用いられている。第1スコア軸の負の方向に「ん_助詞」、「後で_副詞」、「はっと_副詞」、「ふっと_副詞」、「つい_副詞」、「なんとなく_副詞」、「無論_副詞」、「っと_助詞」、「にたいして_助詞」と「なかろ_助動詞」の変数が現れた。川端康成作品の一部と『山の音』の作品では、こういった変数が多く用いられている。

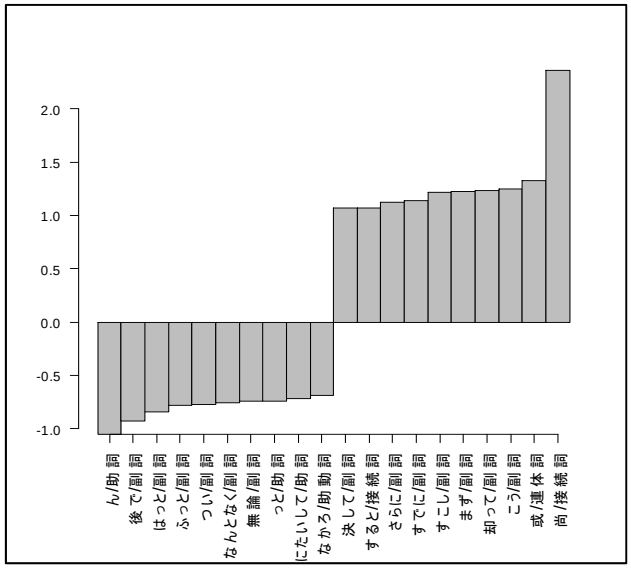

図 8.4 『山の音』タグ付き形態素の変数第 1 スコアの棒グラフ

『山の音』のコーパスから抽出した文節パターンの次元数は485である。文節パターンの対応分析個体の第1、2成分のスコアを図8.5に示す。図8.5では、川端康成の作品は第1、第2、第3象限にプロットされた。三島由紀夫の作品は第2、第3象限に川端康成の作品と重なってプロットされた。三島由紀夫の作品は第1と第4象限にプロットされた。このような傾向は、三島由紀夫と比べ、『山の音』の文体は川端康成に近いことを示した。

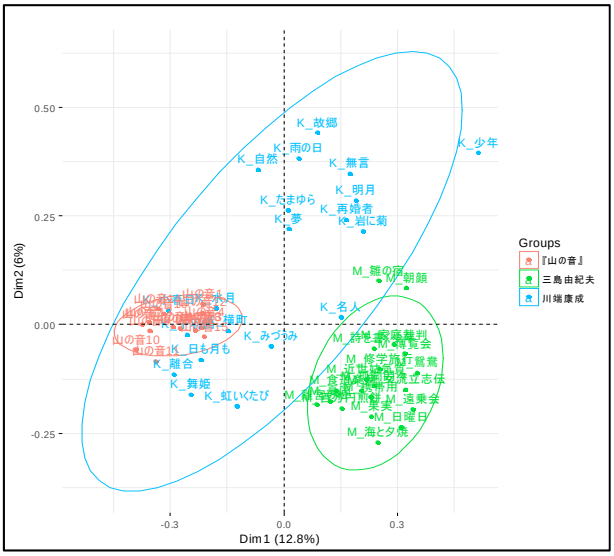

図 8.5 『山の音』文節パターンの個体第 1、2 スコアの散布図

『山の音』文節パターンの第1スコアの棒グラフを図8.6に示す。第1スコア軸の正の方向に「「_名詞_名詞_名詞_で_の」、「名詞_」_の」、「名詞_名詞_名詞_名詞」、「連体詞_助動詞」、「「_名詞_の」、「代名詞_は_、」、「代名詞_名詞_の」、「名詞_名詞_助動詞_助動詞_助動詞_。」、「名詞_名詞_名詞_名詞_を」と「名詞_動詞_を」の変数が現れた。川端康成

作品の一部と三島由紀夫の作品では、このような変数が多く用いられている。第1スコア軸の負の方向に「名詞_ん」、「動詞_名詞_助動詞_助動詞_か_。」、「動詞_て_動詞_助動詞_名詞_助動詞_助動詞_か_。」、「と_、」、「名詞_にたいする」、「動詞_助動詞_名詞_を_、」、「名詞_まで」、「動詞_助動詞_て_、」、「副詞_動詞_て_動詞_助動詞_。」と「動詞_助動詞_と_、」の変数が現れた。川端康成作品の一部と『山の音』では、こういった変数が多く用いられている。

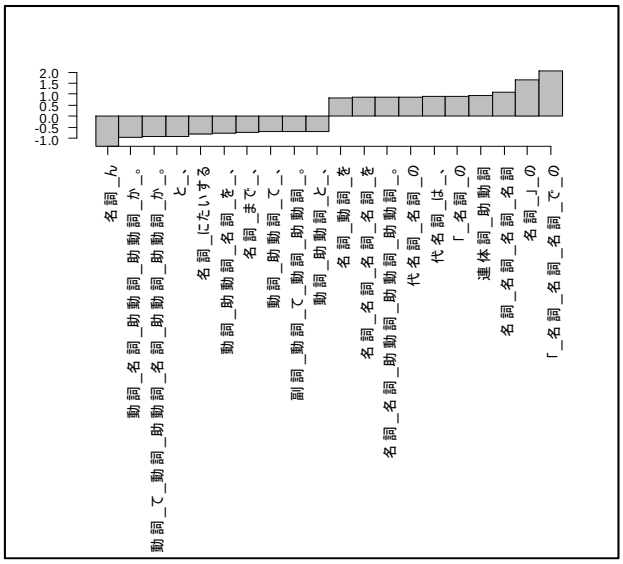

図 8.6 『山の音』文節パターン変数の第 1 スコアの棒グラフ

8.3 クラスター分析の結果

『山の音』の文字記号bi-gramを用いた階層的クラスター分析の結果を図8.7に示す。作品のクラスターは、大きく『山の音』各章クラスター、三島由紀夫作品クラスターと川端康成作品クラスター、川端康成作品クラスターの三つに分かれている。図8.7では、『山の音』の各回は川端康成と三島由紀夫の作品クラスターと離れているため、文字記号のbigramでは、川端康成と三島由紀夫の文体とは異なることが見て取れた。

『山の音』のタグ付き形態素を用いた階層的クラスター分析の結果を図8.8に示す。作品のクラスターは、大きくの三つのクラスターに分かれている。左側のクラスターには『山の音』の各章と川端康成大半の作品が入っている。右側のクラスターには三島由紀夫の作品と4編の川端康成の作品が入っている。この結果から、タグ付き形態素では『山の音』の文体は大まかに川端康成に似ていることが見て取れた。

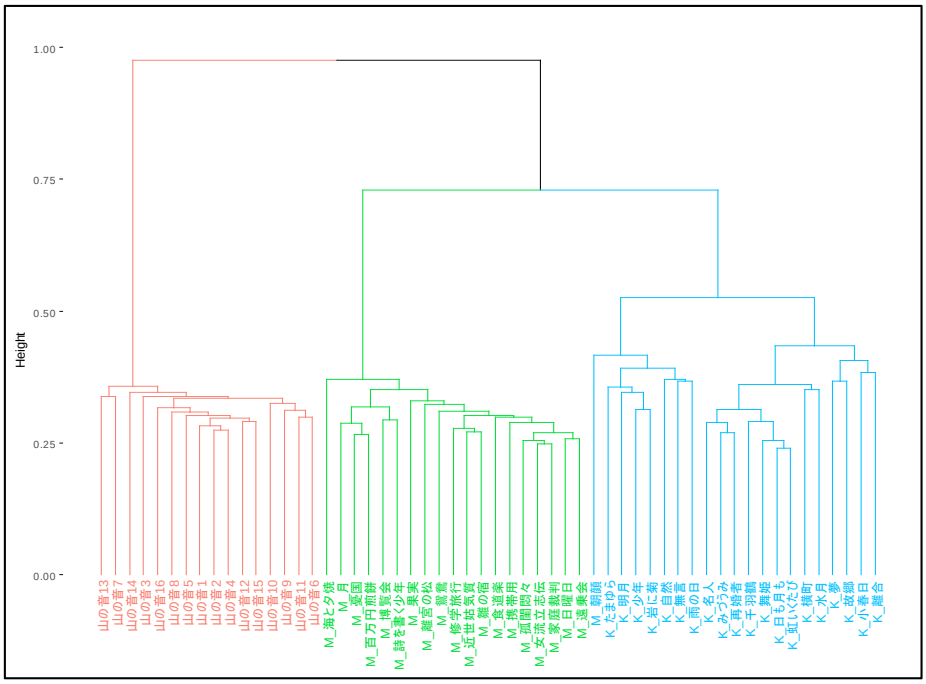

図 8.7 『山の音』の文字記号 bi-gram の階層的クラスター樹形図

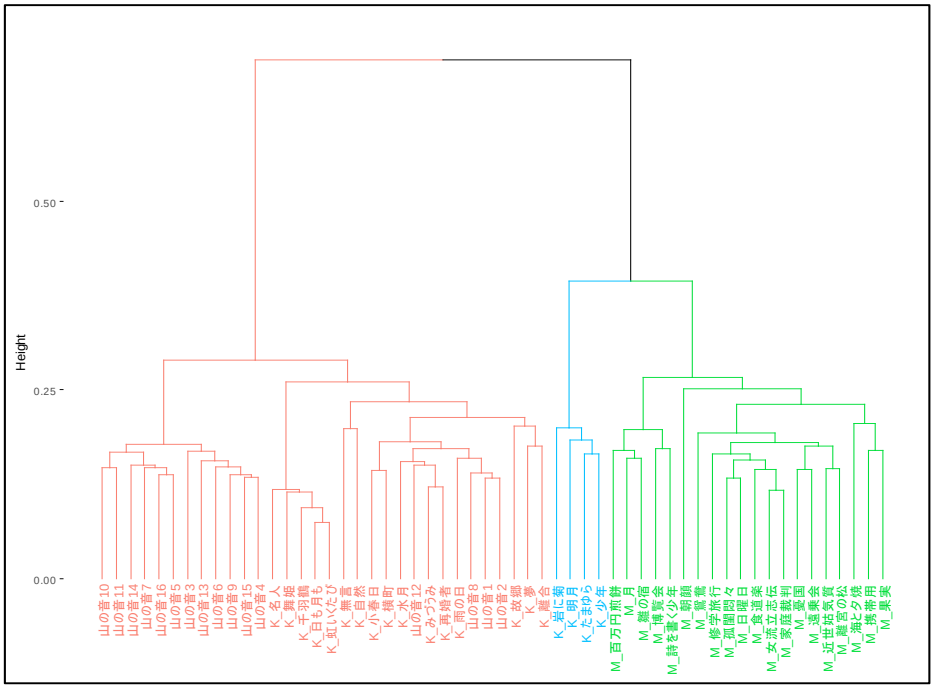

図 8.8 『山の音』のタグ付き形態素の階層的クラスター樹形図

『山の音』の文節パターンを用いた階層的クラスター分析の結果を図8.9に示す。作品のクラスターは、大きくの二つのクラスターに分かれている。左側のクラスターには『山の音』の各章と川端康成半分の作品が入っている。右側のクラスターには三島由紀夫の作品と川端

康成残りの半分の作品が入っている。文節パターンでは『山の音』は川端康成作品の一部に似ていることが見て取れた。

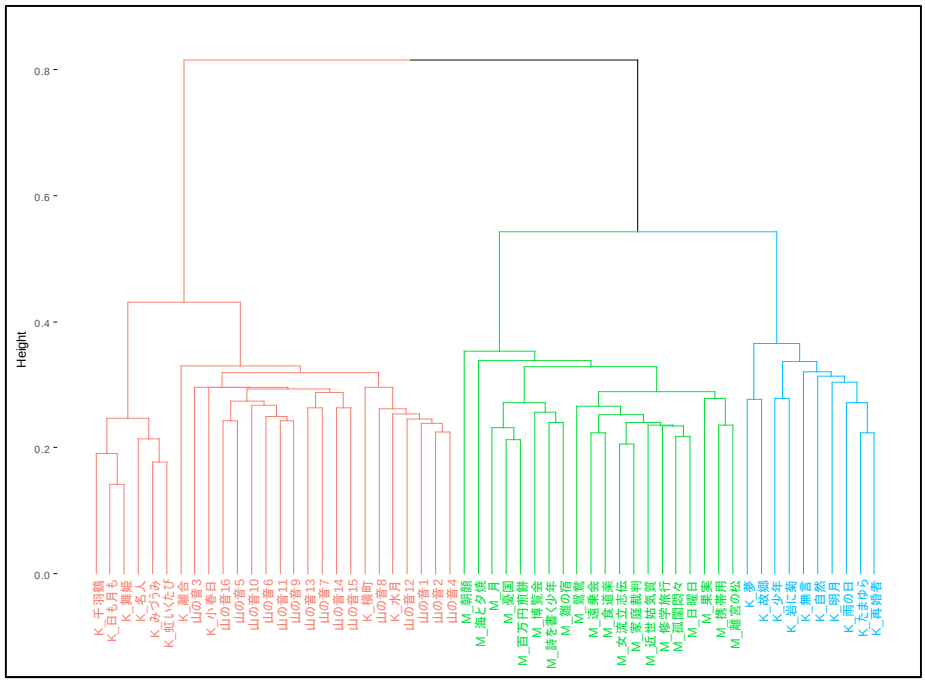

図 8.9 『山の音』の文節パターンの階層的クラスター樹形図

8.4 分類器による判別結果

文字記号 bi-gram、タグ付き形態素と文節パターンの三つの特徴量と、AdaBoost、HDDA、LMT、RF と SVM の 5 つの分類器を用いて解析を行った。各特徴量における分類器の性能を示す適合率 (Precision)、再現率 (Recall)と F-尺度 (F-measure)を表 8.4 に示す。

表8.4各特徴データにおける分類器の性能評価

特徴量	評価指標	AdaBoost	HDDA	LMT	RF	SVM
文字記号 bi-gram	Precision	0.95	0.87	0.87	1	0.89
	Recall	0.95	1	1	1	0.8
	F-measure	0.95	0.93	0.93	1	0.84
形態素タグ bi-gram	Precision	0.95	0.95	1	1	1
	Recall	0.95	1	0.95	1	1
	F-measure	0.95	0.98	0.97	1	1
文節 パターン	Precision	0.78	0.9	0.95	1	1
	Recall	0.9	0.95	0.95	1	1
	F-measure	0.84	0.93	0.95	1	1

8.4.1 文字記号 bi-gram

文字記号 bi-gram において、『山の音』16 章の分類器による判別結果と統合結果を表 8.5 に示す。表 8.5 では、すべての特徴量と分類器の組み合わせでは、『山の音』の 16 章は川端康成に判別され、統合結果も川端康成になっている。

表8.5 『山の音』各章の文字記号bi-gramを用いた5つの分類器による判別結果
(K: 川端康成 M: 三島由紀夫)

『山の音』	AdaBoost	HDDA	LMT	SVM	RF	統合結果
1	K	K	K	K	K	K
2	K	K	K	K	K	K
3	K	K	K	K	K	K
4	K	K	K	K	K	K
5	K	K	K	K	K	K
6	K	K	K	K	K	K
7	K	K	K	K	K	K
8	K	K	K	K	K	K
9	K	K	K	K	K	K
10	K	K	K	K	K	K
11	K	K	K	K	K	K
12	K	K	K	K	K	K
13	K	K	K	K	K	K
14	K	K	K	K	K	K
15	K	K	K	K	K	K
16	K	K	K	K	K	K

8.4.2 タグ付き形態素

タグ付き形態素において、『山の音』16 章の分類器による判別結果と統合結果を表 8.6 に示す。表 8.6 では、すべての特徴量と分類器の組み合わせでは、『山の音』の 16 章は川端康成に判別され、統合結果も川端康成になっている。

表 8.6 『山の音』各章のタグ付き形態素を用いた 5 つの分類器による判別結果
(K: 川端康成 M: 三島由紀夫)

『山の音』	AdaBoost	HDDA	LMT	SVM	RF	統合結果
1	K	K	K	K	K	K
2	K	K	K	K	K	K
3	K	K	K	K	K	K
4	K	K	K	K	K	K
5	K	K	K	K	K	K
6	K	K	K	K	K	K
7	K	K	K	K	K	K
8	K	K	K	K	K	K
9	K	K	K	K	K	K
10	K	K	K	K	K	K
11	K	K	K	K	K	K
12	K	K	K	K	K	K
13	K	K	K	K	K	K
14	K	K	K	K	K	K
15	K	K	K	K	K	K
16	K	K	K	K	K	K

8.4.3 文節パターン

文節パターンにおいて、『山の音』16章の分類器による判別結果と統合結果を表8.7に示す。表8.7では、すべての特徴量と分類器の組み合わせでは、『山の音』の16章は川端康成に判別され、統合結果も川端康成になっている。

表8.7 『山の音』各章の文節パターンを用いた5つの分類器による判別結果
(K: 川端康成 M: 三島由紀夫)

『山の音』	AdaBoost	HDDA	LMT	SVM	RF	統合結果
1	K	K	K	K	K	K
2	K	K	K	K	K	K
3	K	K	K	K	K	K
4	K	K	K	K	K	K
5	K	K	K	K	K	K
6	K	K	K	K	K	K
7	K	K	K	K	K	K
8	K	K	K	K	K	K
9	K	K	K	K	K	K
10	K	K	K	K	K	K
11	K	K	K	K	K	K
12	K	K	K	K	K	K
13	K	K	K	K	K	K
14	K	K	K	K	K	K
15	K	K	K	K	K	K
16	K	K	K	K	K	K

8.5 本章のまとめ

『山の音』は川端康成の名義で発表され、三島由紀夫の代筆と疑われる小説である。『山の音』のタグ付き形態素と文節パターンにおける対応分析の結果では、『山の音』の各章は三島由紀夫より川端康成近いケースが多い。クラスター分析の結果において、タグ付き形態素では、『眠れる美女』は川端康成と三島由紀夫のグループから離れていたが、タグ付き形態素では、川端康成作品のクラスターに入っている。5つの分類器を用いた判別では、すべての文体特徴量と分類器の組み合わせでは、川端康成に判別された。以上の分析より、『山の音』は三島由紀夫による代筆の可能性は極めて低いという結論になった。

第 9 章 川端康成の文体の存在問題

第 9~11 章では、川端康成に纏わる三つの文体問題を取り上げる。一つ目は川端康成の文体存在問題、つまり、川端康成は文体を持つかの問題である。二つ目は川端康成の文体変化問題、つまり、川端康成の文体は終戦を境に変化が生じたかどうかの問題である。三つ目は川端康成の語彙問題である。本章では、川端康成の文体存在問題を取り上げる。

「文体とは何か」の議論は昔から行われていたが、学者の数だけ定義があると言われている (揚妻他, 2015)。数多くの定義の中で広く知られているのは、次に示した文体学者の中村 (1993)による文体の定義である。

> 文体は表現主体によって開かれた文章が、受容主体の参加によって展開する過程で、異質性としての印象、効果を果たす時に、その動力となった作品形成上の言語的な性格の統合である。

また、『大辞林』 (第三版)では、文体を「①文章の形式・様式。②語句・語法・修辞などにみられる、その作者特有の文章表現。」と定義している (松村(編), 2006)。この「文章表現上著者らしい特色」の詳細について、文体研究者の中村 (2010)は、次のようにまとめた。

> 文章の表現上の性格を他と対比的にとらえた特殊性。文体を類型面でとらえるか個性面でとらえるかによって大きく二分され、現実に次のように多様な意味で用いられている。 (1)文字表記の違い (2)使用語彙の違い (3)語法の違い (4)文末表現の違い (5)文章の種類の違い (6)文章の用途の違い (7)ジャンルの違い (8)調子の違い (9)修辞の違い (10)文章の性格の違い (11)時代の違い (12)使用言語の違い (13)表現主体の属性の違い (14)文学史上の流派の違い (15)作家ごとの文章や表現の違い (16)執筆時期の違い (17)作品ごとの文章や表現の特徴の違い。

以上の文体項目に基づいた文体研究は長年文学専門家の内省によって行われてきた。文体研究における内省の方法は「伝統的な方法」として知られ、それに対して、「非伝統的な方法」は文体分析のための統計的手法を指す。イビィッチ (1974)は、統計的手法の言語・文体研究に貢献できる分野として次のように挙げている。

> 文献の著者判定：どのような単語が、どの程度の頻度で用いられているかを調べることにより、作者がはっきりしていない文献の著者判定や、文献の執筆年代の推定が可能となってきた。
>
> 文体論：統計的手法の利用により、文体の研究は、客観的かつ精密になった。ある表現がどの程度陳腐であるかは、その表現の頻度の大きさと関係している。

イビィッチ (1974)が述べたように、伝統的な内省の方法を補い、著者判定と文体の精密分析統計的手法が役立つ。本論文の第 3 章から第 8 章までは著者判定の問題に統計的手法を適用した。本章ではより進んだ統計的手法を用いて川端康成の文体存在問題を解明する。

川端康成の文体特徴について、佐伯 (1959)は次のように述べている。

> 本質的に抒情的な散文であって、繊細微妙な美しさにみちている。といって、感傷的な情緒性に流れすぎることはなくて、むしろ知的な透徹した味わいを底に感じさせる。知性によって規制された、冷たい抒情美の文体である。また、自然描写の豊富さ、季節感の鮮かさ、その反面に人間描写、性格や心理描写における著しい省筆ぶりなどわが国の伝統的な文学感覚に深く根をおろして、この上なく典型的な「日本的」文体と見えながら、不思議と古さを感じさせない。一面、ひどくモダーンで、新鮮だ。鋭角的な近代感覚が随所に閃いていて、扱われる対象、用いられるイメージに大胆斬新な組合せが目立つ。

このような文章の深読みを通して川端康成の文体のイメージを掴めた研究者がいる一方で、川端康成の作品を読んでもその文体を全く感じ取れないという意見を唱える研究者もいる。文学批評家の寺田 (1949)が川端康成の文体について次のように述べている。

> 百五十編近い彼の創作を通読する間、僕には絶えず芥川龍之介と泉鏡花と徳田秋聲と横光利一の口許、口調、聲音が思い出されてならなかった。文章の姿に執念深い注意を怠らないように見える川端が、殆ど独自な文体を持っていないことを知ったのは僕の驚きで…

このような研究は、いずれも研究者の文章に対する「感覚」から結論を導きだしたものである。しかし、同じ文章を読んでも研究者によって感じたものも異なるため、十人十色の結論になりかねない。このような先行研究を補い、データで客観的に示す発想から、本研究では、統計的手法を川端康成の文体問題に導入した。精神科医の栗原 (1982)は、川端康成小説の題名の文字数を分析した。川端康成の小説の題名は、谷崎潤一郎、三島由紀夫と夏目漱石の小説の題名より短いものが多く、統計学的に有意の差があることも示された。

統計的手法を適用する前に研究対象を決めなければならない。寺田 (1949)が川端康成の文体は存在しないと指摘したのは、他の作家の作品を読んでいるうちにその作家ならではの特徴を捉えることはできたが、川端康成の場合、何も感じとれていないからである。つまり、「川端康成文体の存在」問題に関して、ほかの作家と比べて川端康成は自分の文体特徴を持つかという文体比較の問題と見なせる。換言すれば、文体特徴量を用いて、川端康成の作品とほかの作家の作品の分類を行う際に、はっきりと分かれた場合、川端康成は自分の文体を持つと言え、そうではない場合、川端康成は自分の文体を持たないということになる。寺田 (1949)は、川端康成の文章を芥川龍之介、泉鏡花、徳田秋聲と横光利一の文章と比較したが、金 (2009)は、芥川龍之介の文体は安定しないと指摘したため、本論文では、比較対象から芥

川龍之介を外し、泉鏡花、徳田秋聲と横光利一の文章を用いることにした。また、文体比較に用いた特徴量を文字記号 bi-gram、タグ付き形態素と文節パターンとし、統計的手法として対応分析とクラスター分析を用いた。

9.1 川端康成の文体存在研究のためのコーパス

本章では、川端康成の文章は川端康成 1969 年の全集から 20 編を抽出し、泉鏡花、徳田秋聲と横光利一の 3 人の小説それぞれ 20 編を青空文庫からダウンロードしてコーパスを作成した。用いた文章のリストを表 9.1 に示す。

表 9.1 文体存在計量分析のためのコーパス

川端康成	泉鏡花	徳田秋聲	横光利一
たまゆら	瓜の涙	チビの魂	火
みづうみ	怨霊借用	のらもの	花園の思想
雨の日	縁結び	或売笑婦の話	街の底
横町	菎蒻本	仮想人物 1	機械
岩に菊	歌行灯	仮想人物 2	御身
故郷	海域発電	花が咲く	時間
再婚者	絵本の春	縮図 1	春は馬車に乗って
自然	外科室	縮図 2	笑われた子
小春日	義血侠血	新世帯	上海
少年	国貞えがく	挿話	睡蓮
水月	七宝の柱	白い月	赤い着物
虹いたび	女客	足跡 1	鳥
日も月も	小春の狐	足跡 2	南北
千羽鶴	雛がたり	町の踊り場	日輪
舞姫	茸の舞姫	風呂桶	比叡
北の海から	木の子説法	霧ヶ峰から鷲ヶ峰へ	微笑
無言	栃の実	和解	旅愁 1
名人	売色鴨南蛮	佗しい放浪の旅	旅愁 2
明月	伯爵の釵	爛	旅愁 3
離合	眉かくしの霊	黴	罌粟の中

9.2 一対比較による結果

9.2.1 川端康成と泉鏡花の文体

本節では、川端康成と泉鏡花の文体を文字記号 bi-gram、タグ付き形態素と文節パターンの三つの文体特徴量及び、対応分析とクラスター分析を用いた結果を紹介する。

川端康成と泉鏡花の作品における文字記号 bi-gram 対応分析個体の第 1、2 スコアの散布図を図 9.1 に示す。川端康成と泉鏡花の作品が大きく二つのグループに分かれ、泉鏡花の作品は第 1 スコア軸の負の方向、川端康成の作品は第 1 スコア軸の正の方向にそれぞれプロットされた。両者ははっきり分かれている。

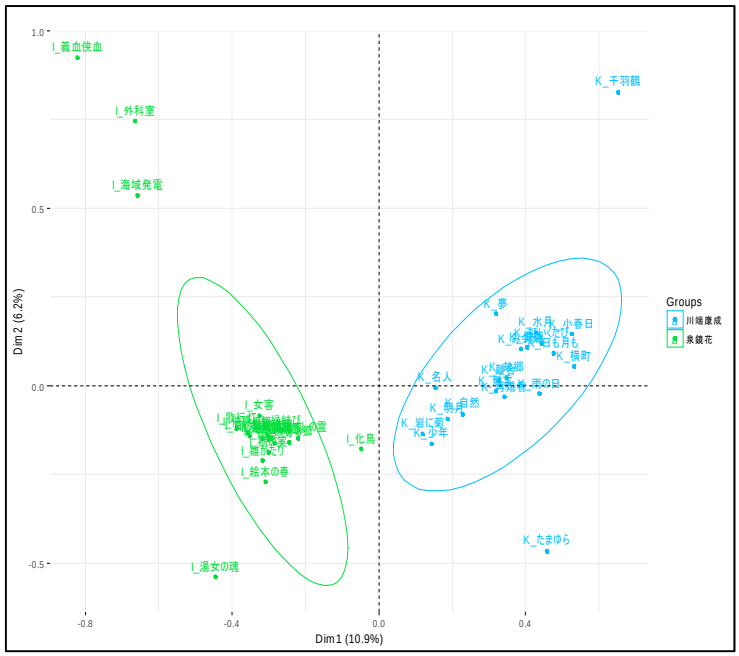

図 9.1 川端康成と泉鏡花の文字記号 bi-gram 対応分析個体の第 1、2 スコアの散布図

川端康成と泉鏡花の作品における文字記号bigram変数の第1スコアの棒グラフを図9.2に示す。第1スコア軸の正の方向に「治が」、「菊治」、「田夫」、「文子」、「太田」、「治は」、「か子」、「。菊」、「治の」と「、菊」の変数が現れた。川端康成の作品では、このような変数が多く用いられている。第1スコア軸の負の方向に「糸は」、「白糸」、「馭者」、「せり」、「渠は」、「き。」、「。渠」、「るを」、「ざる」と「渠の」が現れた。泉鏡花の作品では、こういった変数が多く用いられている。

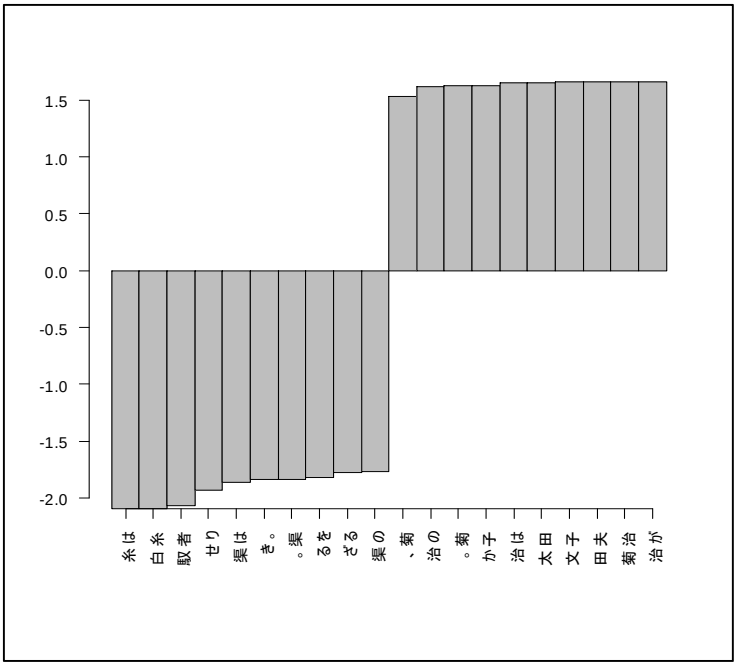

図 9.2 川端康成と泉鏡花の文字記号 bi-gram 変数の第 1 スコアの棒グラフ

川端康成と泉鏡花の作品におけるタグ付き形態素対応分析個体の第 1、2 スコアの散布図を図 9.3 に示す。川端康成と泉鏡花の文章が大きく二つのグループに分かれ、泉鏡花の文章は第 1 スコア軸の正の方向、川端康成の文章は第 1 スコア軸の負の方向にそれぞれプロットされた。

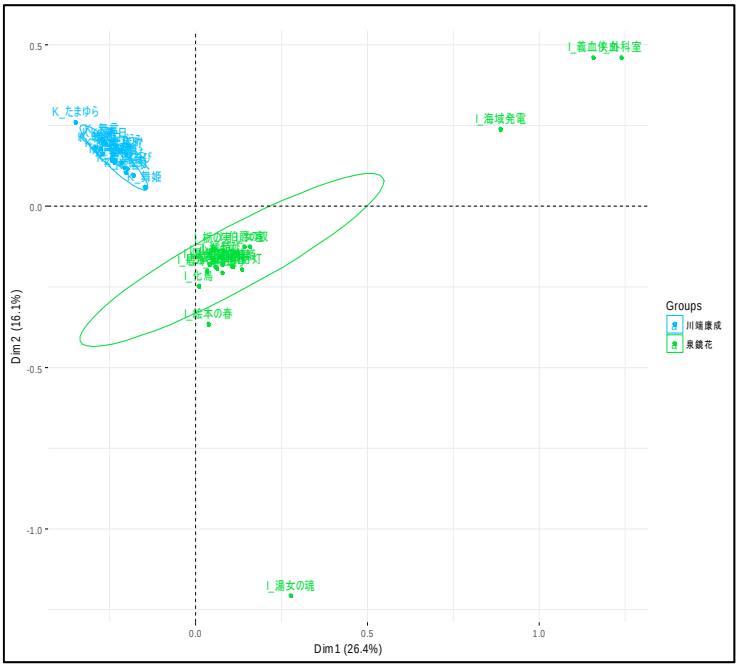

図 9.3 川端康成と泉鏡花のタグ付き形態素対応分析個体の第 1, 2 スコアの散布図

川端康成と泉鏡花の作品におけるタグ付き形態素変数の第1スコアの棒グラフを図9.4に示す。第1スコア軸の正の方向に「べから_助動詞」、「ざる_助動詞」、「なり_助動詞」、「し_助動詞」、「ざり_助動詞」、「たる_助動詞」、「けり_助動詞」、「たり_助動詞」、「にて_助詞」と「において_助詞」の変数が現れた。泉鏡花の作品では、このような変数が多く用いられている。第1スコア軸の負の方向に「たまゆら_副詞」、「「_記号」、「ずいぶん_副詞」、「どんな_連体詞」、「やはり_副詞」、「もっと_副詞」、「」_記号」、「なにか_副詞」、「はっきり_副詞」と「初めて_副詞」が現れた。川端康成作品では、こういった変数が多く用いられている。

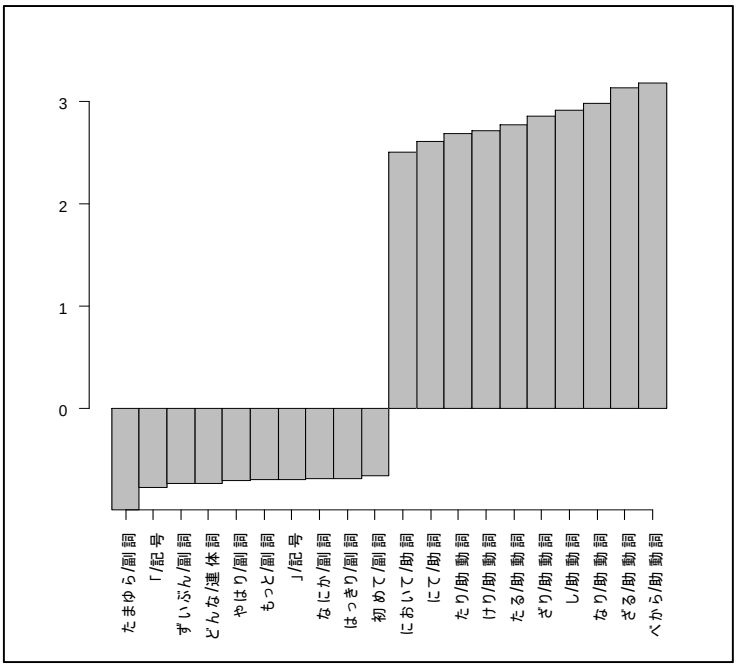

図 9.4 川端康成と泉鏡花のタグ付き形態素の変数第 1 スコアの棒グラフ

川端康成と泉鏡花の作品における文節パターン対応分析個体の第 1、2 スコアの散布図を図 9.5 に示す。川端康成と泉鏡花の文章が大きく二つのグループに分かれ、川端康成の文章は第 1 スコア軸の負の方向、泉鏡花の文章は第 1 スコア軸の正の方向にそれぞれプロットされた。

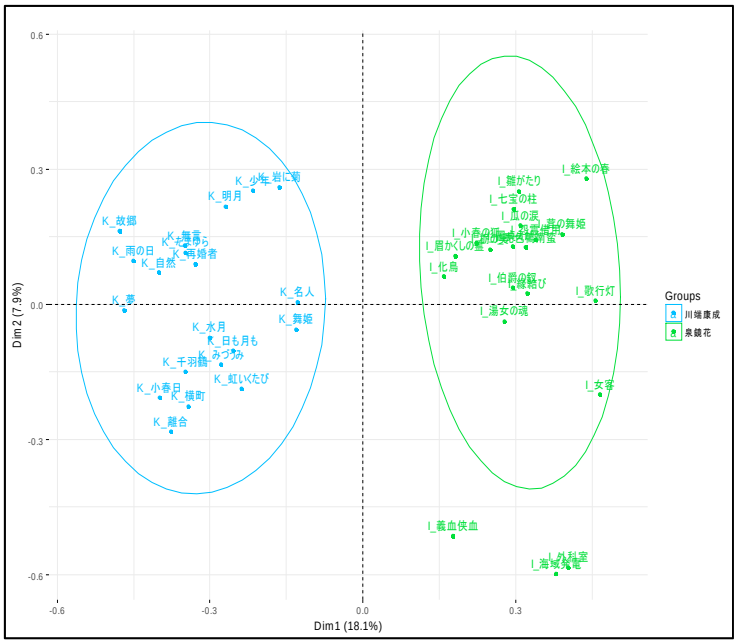

図 9.5 川端康成と泉鏡花の文節パターン対応分析個体の第 1、2 スコアの散布図

川端康成と泉鏡花の作品における文節パターン変数の第1スコアの棒グラフを図9.6に示す。第1スコア軸の正の方向に「副詞_と_、」、「名詞_動詞_、」、「名詞_の_助動詞_、」、「名詞_の_助動詞」、「動詞_まで_、」、「動詞_たり_。」、「名詞_々_と」、「名詞_。」、「名詞_動詞_。」と「動詞_つつ_、」の変数が現れた。泉鏡花の作品では、このような変数が多く用いられている。第1スコア軸の負の方向に「動詞_て_動詞_名詞_助動詞_。」、「動詞_て_動詞_名詞_助動詞」、「と_、」、「名詞_でも_動詞_助動詞_。」、「動詞_て_から」、「動詞_て_動詞_名詞_は_、」、「動詞_助動詞_名詞_助動詞_助動詞_か_。」、「動詞_て_動詞_名詞_助動詞_助動詞_。」、「動詞_助動詞_名詞_か_、」と「形容詞_動詞_て_動詞_助動詞_。」が現れた。川端康成作品では、このような変数が多く用いられている。

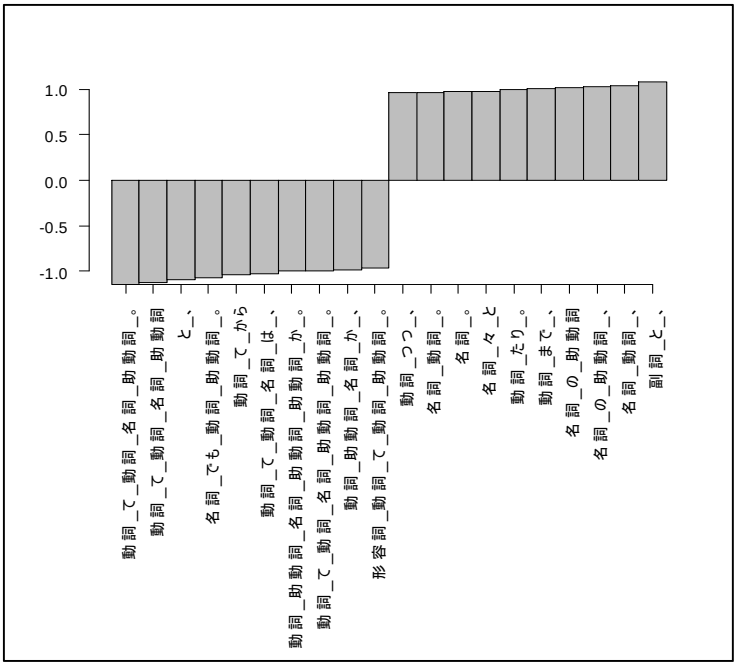

図 9.6 川端康成と泉鏡花の文節パターンの第 1 スコアの棒グラフ

川端康成と泉鏡花作品における文字記号 bi-gram、タグ付き形態素と文節パターンを用いたクラスター分析の結果を図 9.7~9.9 に示す。図 9.7~9.9 では、いずれも川端康成と泉鏡花で大きく二つのクラスターを形成し、川端康成の作品は左側のクラスター、泉鏡花の作品は右側のクラスターに入っている。

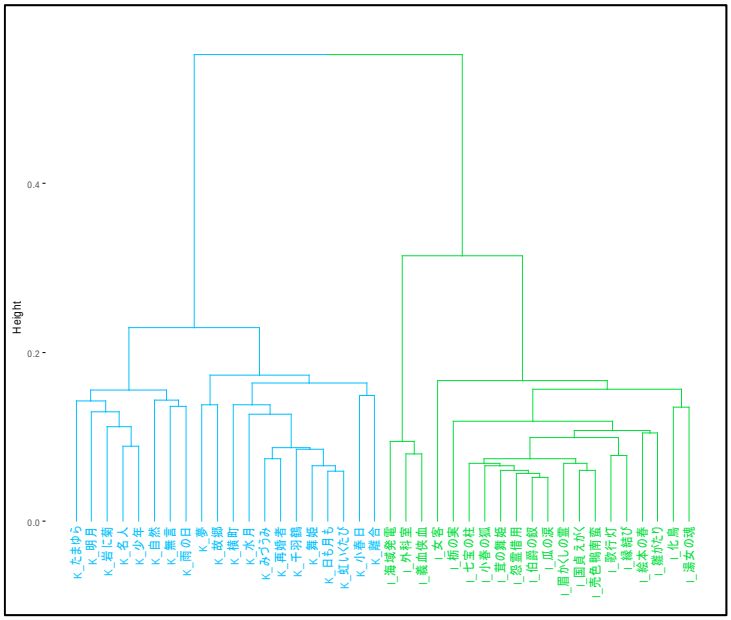

図 9.7 川端康成と泉鏡花の文字記号 bi-gram の階層的クラスター樹形図

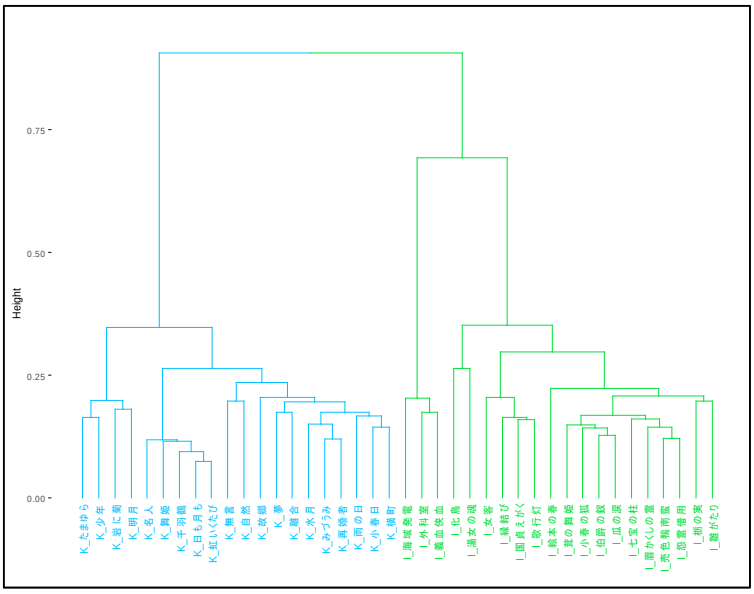

図 9.8 川端康成と泉鏡花のタグ付き形態素の階層的クラスター樹形図

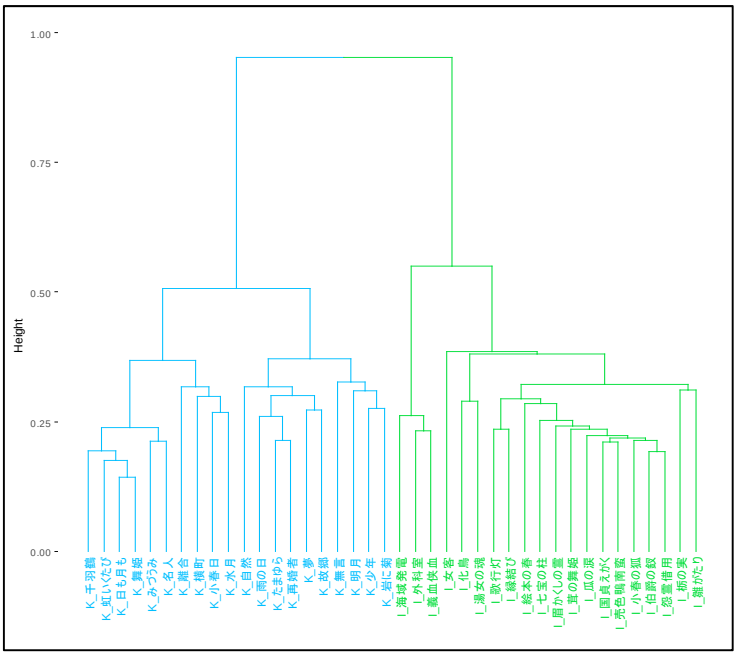

図 9.9 川端康成と泉鏡花の文節パターンの階層的クラスター樹形図

9.2.2 川端康成と徳田秋聲の文体

本節では、川端康成と徳田秋聲の文体を文字記号 bi-gram、タグ付き形態素と文節パターンの三つの特徴量及び、対応分析、クラスター分析を用いた分析結果を紹介する。

川端康成と徳田秋聲の作品における文字記号 bi-gram 対応分析個体の第 1、2 スコアの散布図を図 9.10 に示す。図 9.10 では、川端康成と泉鏡花の文章が分かれ、川端康成の作品は第 1 スコア軸の正の方向、徳田秋聲の作品は第 1 スコア軸の負の方向にプロットされた。

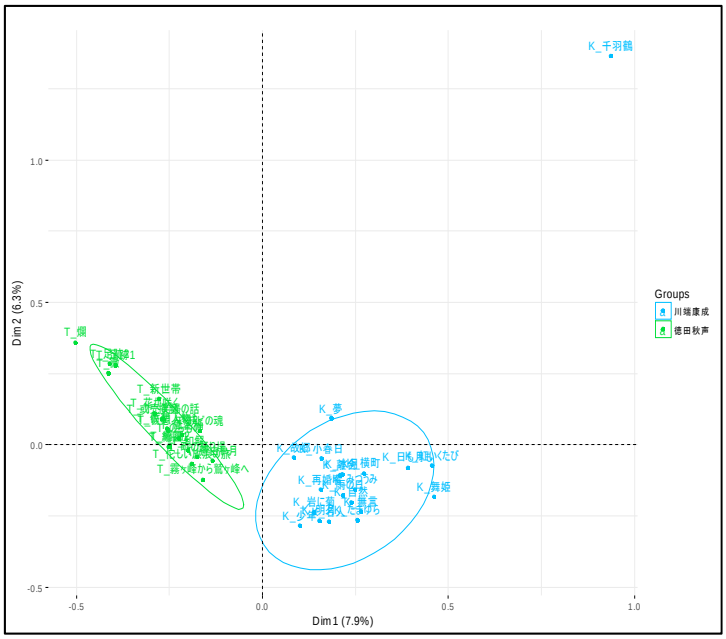

図 9.10 川端康成と徳田秋聲の文字記号 bi-gram 対応分析個体の第 1、2 スコア散布図

川端康成と徳田秋聲の作品における文字記号bigram変数の第1スコアの棒グラフを図9.11に示す。第1スコア軸の正の方向に「菊治」、「文子」、「治は」、「治の」、「。菊」、「か子」、「、菊」、「令嬢」、「。文」と「、文」の変数が現れた。泉鏡花の作品ではこのような変数が多く用いられている。第1スコア軸の負の方向に「お今」、「浅井」、「増は」、「お増」、「お雪」、「。浅」、「増の」、「井の」、「、浅」と「銀は」が現れた。川端康成作品では、このような変数が多く用いられている。

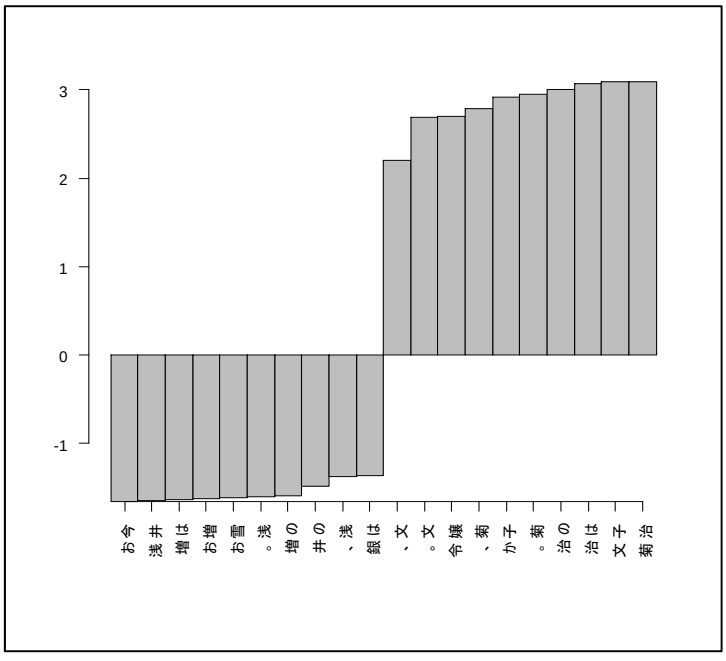

図 9.11 川端康成と徳田秋聲の文字記号 bi-gram の変数第 1 スコアの棒グラフ

川端康成と徳田秋聲の作品におけるタグ付き形態素対応分析個体の第 1、2 スコアの散布図を図 9.12 に示す。図 9.12 では、川端康成と泉鏡花の文章が分かれ、川端康成の作品は第 1 スコア軸の負の方向、徳田秋聲の作品は第 1 スコア軸の正の方向にプロットされた。

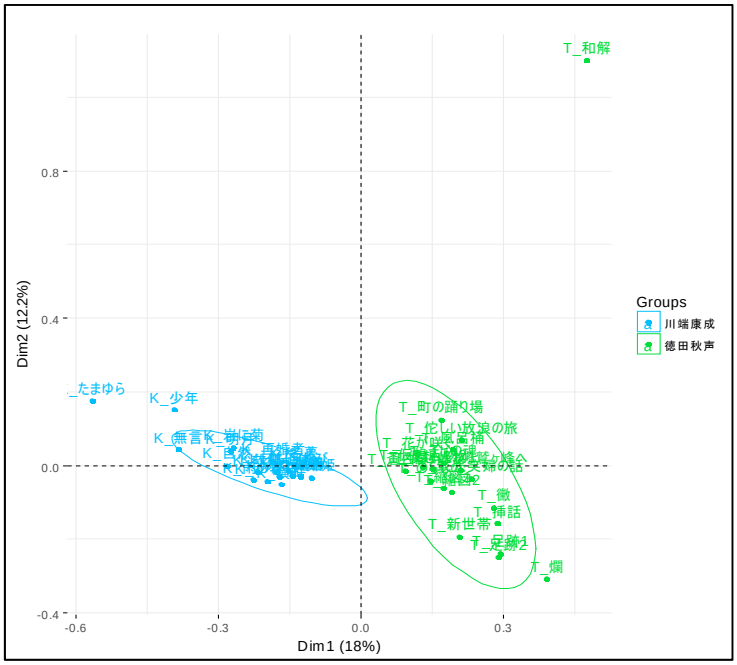

図 9.12 川端康成と徳田秋聲のタグ付き形態素対応分析個体の第 1、2 スコア散布図

川端康成と徳田秋聲の作品におけるタグ付き形態素変数の第1スコアの棒グラフを図9.13に示す。第1スコア軸の正の方向に「T_記号」、「O_記号」、「K_記号」、「M_記号」、「ー_記号」、「にやにや_副詞」、「じきに_副詞」、「ちょいちょい_副詞」、「じろじろ_副詞」と「お_接頭詞」の変数が現れた。徳田秋聲の作品では、このような変数が多く用いられている。第1スコア軸の負の方向に「たまゆら_副詞」、「「_記号」、「」_記号」、「けれども_助詞」、「あるいは_接続詞」、「まったく_副詞」、「無_接頭詞」、「あろ_助動詞」、「なにか_副詞」と「おそらく_副詞」が現れた。川端康成作品では、このような変数が多く用いられている。

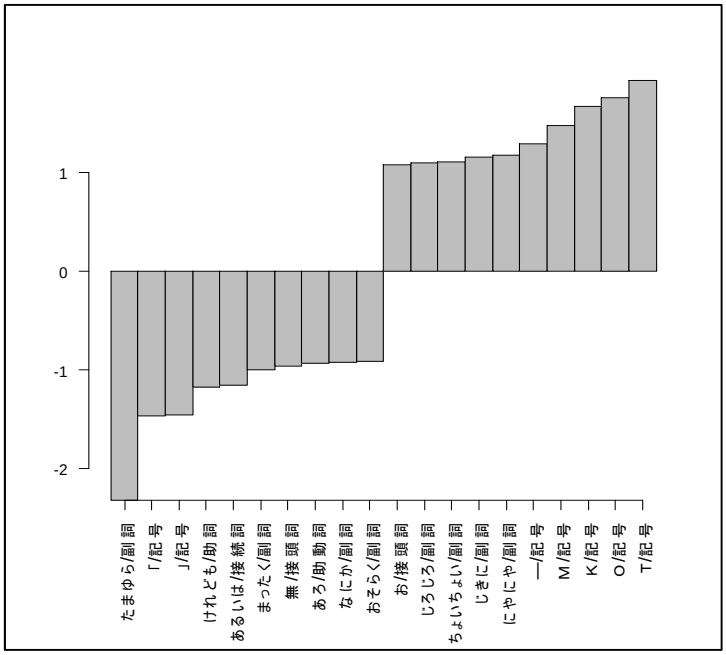

図 9.13 川端康成と徳田秋聲のタグ付き形態素の変数第 1 スコアの棒グラフ

川端康成と徳田秋聲の作品における文節パターン対応分析個体の第 1、2 スコアの散布図を図 9.14 に示す。図 9.14 では、川端康成と泉鏡花の文章が分かれ、川端康成の作品は第 1 スコア軸の負の方向、徳田秋聲の作品は第 1 スコア軸の正の方向にプロットされた。

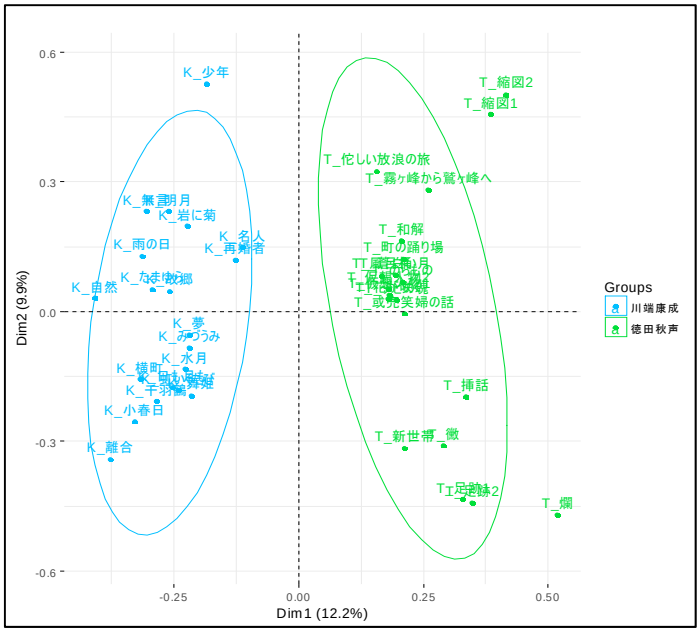

図 9.14 川端康成と徳田秋聲の文節パターン対応分析個体の第 1、2 スコア散布図

川端康成と徳田秋聲の作品における文節パターン変数の第1スコアの棒グラフを図9.15に示す。第1スコア軸の正の方向に「動詞_名詞_名詞」、「名詞_動詞_も」、「動詞_って」、「動

詞_動詞_たり_動詞_助動詞_。」、「名詞_名詞_動詞_に」、「名詞_動詞_へ」、「名詞_動詞_を」、「名詞_名詞_動詞_を」、「サ変接続_動詞_を」と「接頭詞_名詞_は_、」の変数が現れた。徳田秋聲の作品では、このような変数が多く用いられている。第1スコア軸の負の方向に「動詞_て_動詞_名詞_助動詞_。」、「動詞_助動詞_名詞_助動詞_。」、「動詞_助動詞_名詞_助動詞_か_。」、「名詞_助動詞_名詞_か_、」、「動詞_て_動詞_名詞_は_、」、「名詞_名詞_助動詞_助動詞_。」、「動詞_名詞_助動詞_。」、「名詞_の_名詞_助動詞_助動詞_。」、「動詞_て_動詞_助動詞_名詞_助動詞_。」と「形容詞_。」の変数が現れた。川端康成作品では、このような変数が多く用いられている。

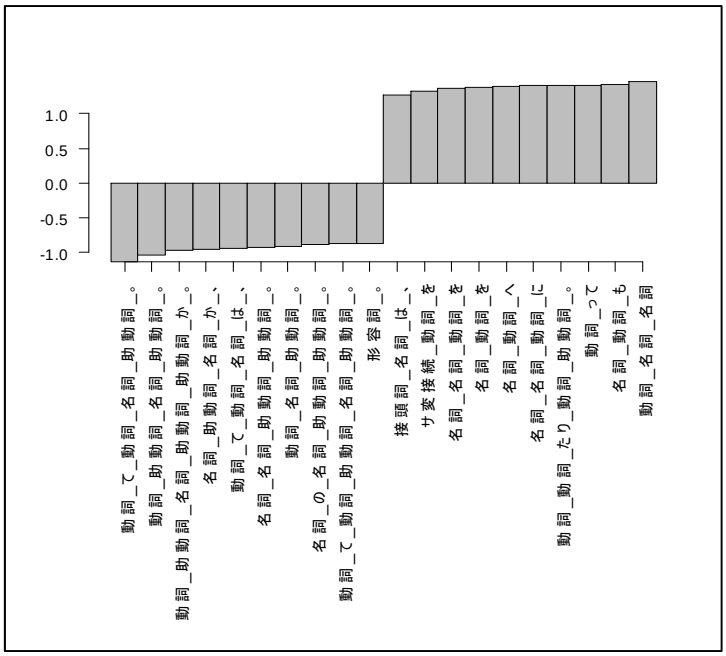

図 9.15 川端康成と徳田秋聲の文節パターンの変数第 1 スコアの棒グラフ

文字記号 bi-gram、タグ付き形態素と文節パターンを用いたクラスター分析の結果を図 9.16~9.18 に示す。図 9.16~9.18 では、いずれも川端康成と泉鏡花で大きく二つのクラスターを形成し、川端康成の作品は左側のクラスター、徳田秋聲の作品は右側のクラスターに入っている。

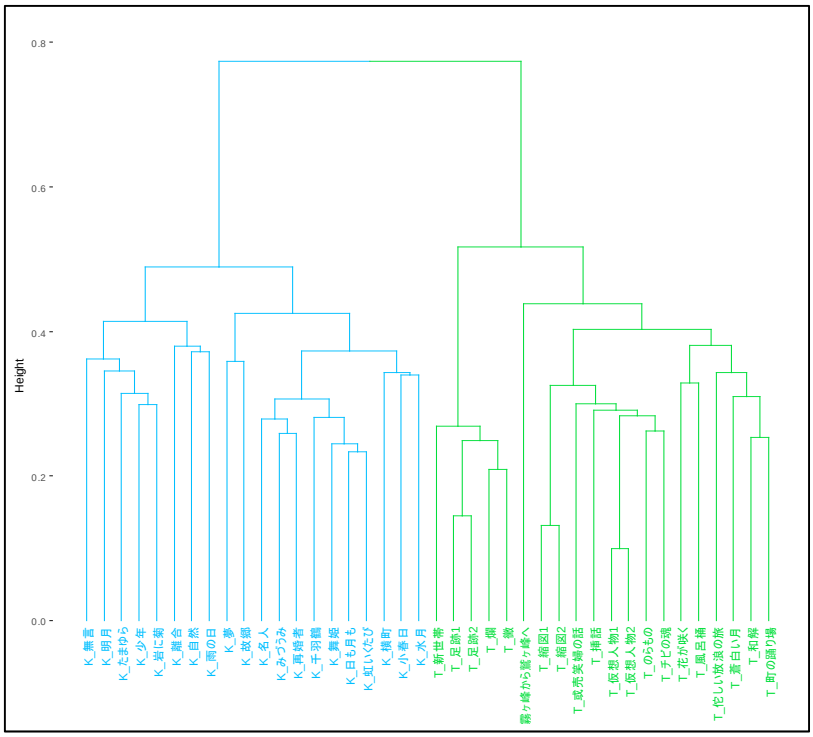

図 9.16 川端康成と徳田秋聲の文字記号 bi-gram の階層的クラスター樹形図

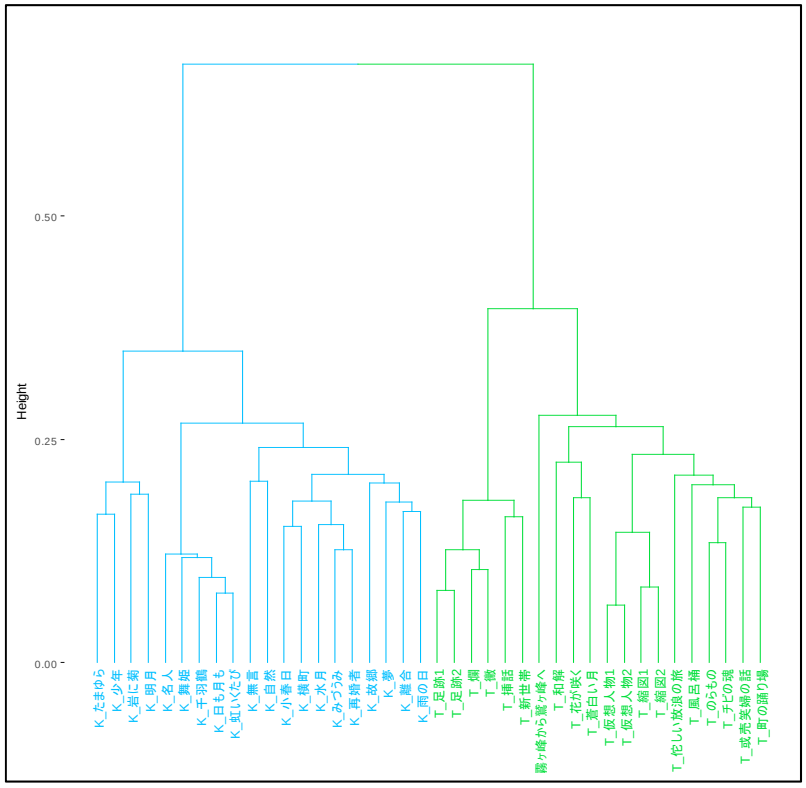

図 9.17 川端康成と徳田秋聲のタグ付き形態素の階層的クラスター樹形図

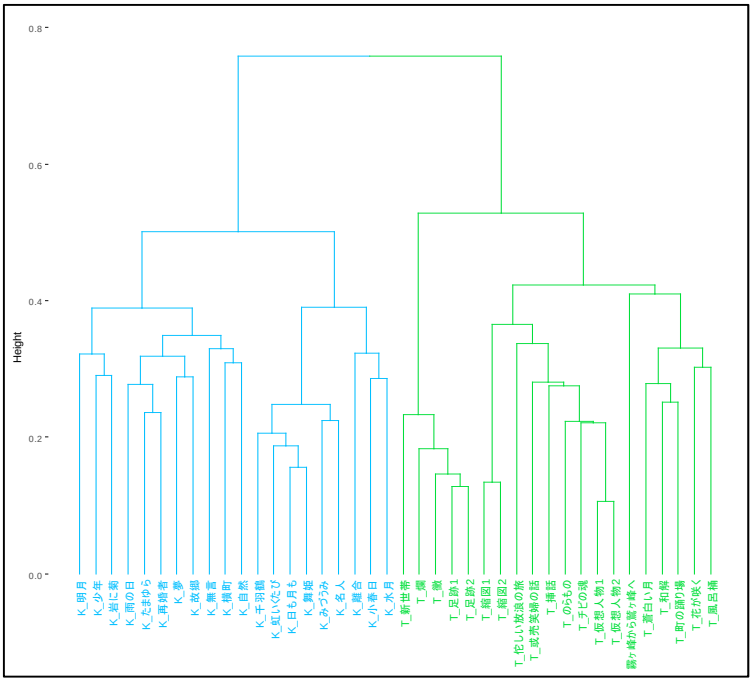

図 9.18 川端康成と徳田秋聲の文節パターンの階層的クラスター樹形図

9.2.3 川端康成と横光利一の文体

本節では、川端康成と横光利一の文体を文字記号 bi-gram、タグ付き形態素と文節パターンの三つの特徴量及び、対応分析とクラスター分析を用いた分析結果を紹介する。

川端康成と横光利一の作品における文字記号 bi-gram 対応分析個体の第 1、2 スコアの散布図を図 9.19 示す。図 9.19 では、川端康成と横光利一の文章が分かれ、川端康成の文章は第 1 スコア軸の正の方向、横光利一の文章は第 1 スコア軸の負の方向にプロットされた。

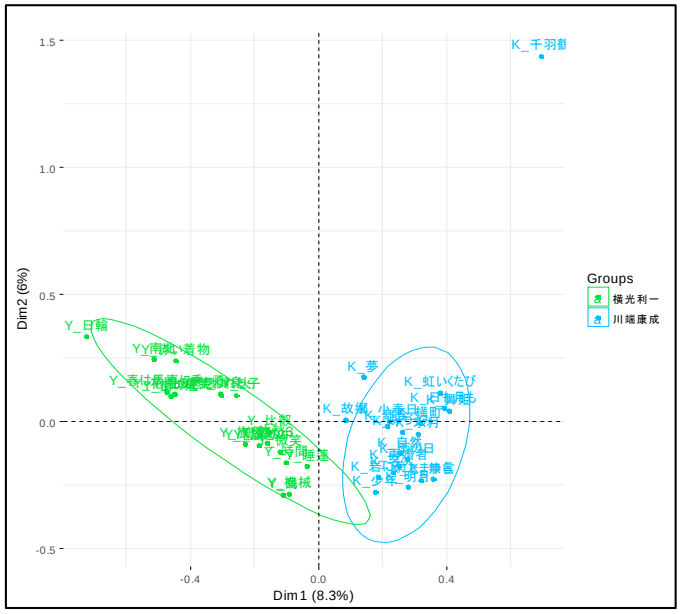

図 9.19 川端康成と横光利一の文字記号 bi-gram 対応分析個体の第 1、2 スコア散布図

川端康成と横光利一の文章における文字記号bigramの第1スコアの棒グラフを図9.20に示す。第1スコア軸の正の方向に「菊治」、「文子」、「治は」、「令嬢」、「治の」、「、文」、「。菊」、「か子」、「。ち」と「。文」の変数が現れた。川端康成の作品では、このような変数が多く用いられている。第1スコア軸の負の方向に「呼の」、「長羅」、「反絵」、「呼は」、「卑弥」、「弥呼」、「和郎」、「訶和」、「兵士」と「。反」が現れた。川端康成作品では、このような変数が多く用いられている。

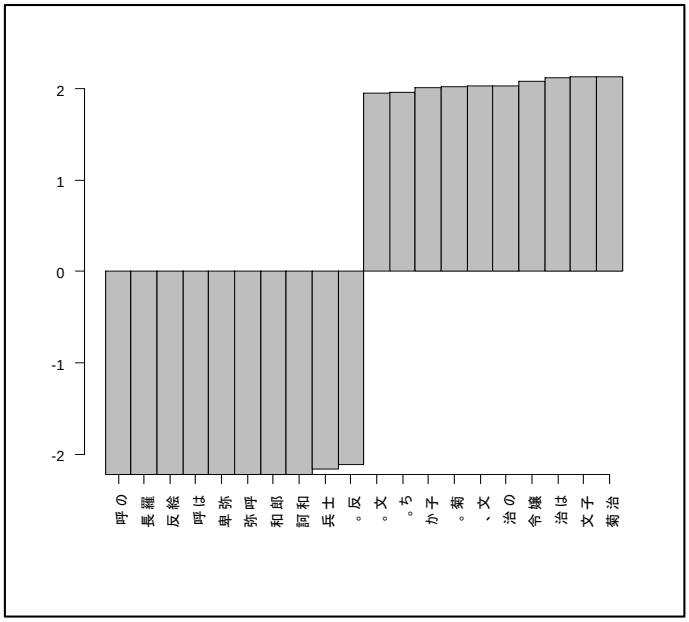

図 9.20 川端康成と横光利一の文字記号 bi-gram 変数の第 1 スコアの棒グラフ

川端康成と横光利一の作品におけるタグ付き形態素対応分析個体の第 1、2 スコアの散布図を図 9.21 示す。図 9.21 では、川端康成と横光利一の文章が分かれ、川端康成の文章は第 1 スコア軸の負の方向、横光利一の文章は第 1 スコア軸の正の方向にプロットされた。

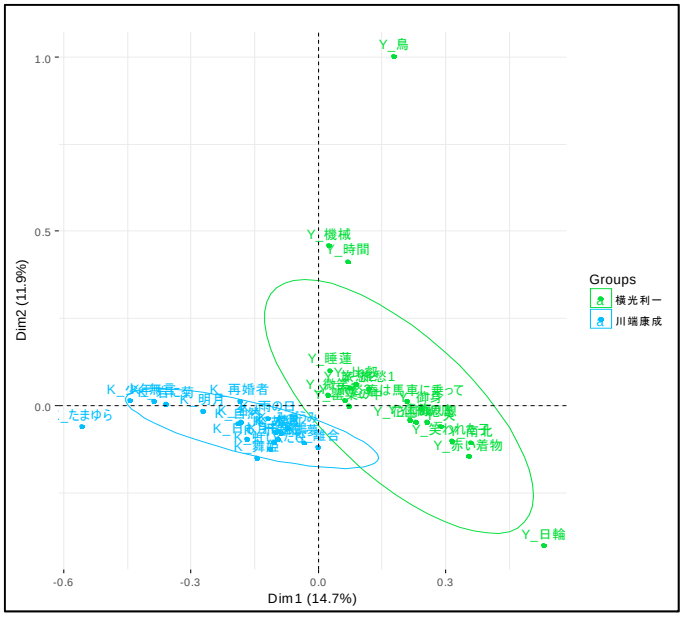

図 9.21 川端康成と横光利一のタグ付き形態素対応分析個体の第 1、2 スコア散布図

川端康成と横光利一の文章におけるタグ付き形態素の第1スコアの棒グラフを図9.22に示す。第1スコア軸の正の方向に「反_接頭詞」、「長_接頭詞」、「が_接続詞」、「再び_副詞」、「不_接頭詞」、「そうして_接続詞」、「次第に_副詞」、「ぜ_助詞」、「忽ち_副詞」と「時々_副詞」の変数が現れた。横光利一の作品では、このような変数が多く用いられている。第1スコア軸の負の方向に「たまゆら_副詞」、「「_記号」、「」_記号」、「無_接頭詞」、「けれども_助詞」、「たいてい_副詞」、「あるいは_接続詞」、「くらい_助詞」、「おそらく_副詞」と「だろ_助動詞」の10個が現れた。川端康成作品では、このような変数が多く用いられている。

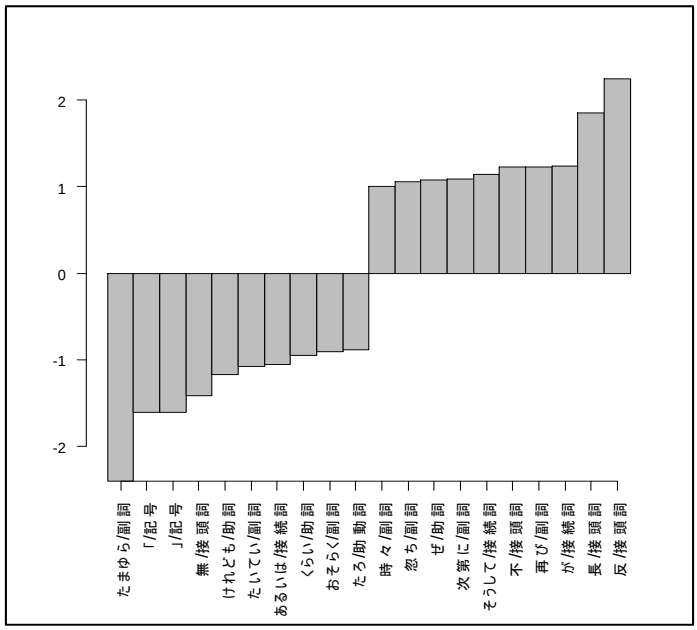

図 9.22 川端康成と横光利一のタグ付き形態素の変数第 1 スコアの棒グラフ

タグ付き形態素対応分析個体の第 1、2 成分のスコアを図 9.23 示す。図 9.23 では、川端康成と横光利一の文章が分かれ、川端康成の文章は第 1 スコア軸の正の方向、横光利一の文章は第 1 スコア軸の負の方向にプロットされた。

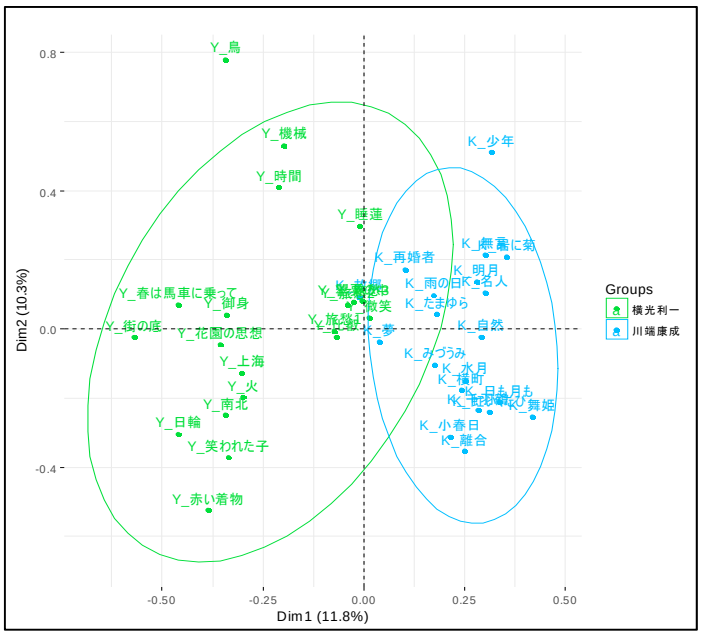

図 9.23 川端康成と横光利一の文節パターン対応分析個体の第 1、2 スコア散布図

川端康成と横光利一の文章における文節パターンの第1スコアの棒グラフを図9.24に示す。第1スコア軸の正の方向に「動詞_助動詞_名詞_が_、」、「名詞_名詞_名詞_、」、「名詞_名詞_名詞_名詞_、」、「名詞_にたいする」、「動詞_動詞_て_動詞_。」、「名詞_名詞_名詞_名詞_名詞_の」、「名詞_名詞_名詞_助動詞_助動詞_。」、「ほど_助動詞_助動詞_。」、「動詞_から_、」と「サ変接続_名詞_に_、」の変数が現れた。川端康成の作品では、このような変数が多く用いられている。第1スコア軸の負の方向に「名詞_々_と」、「動詞_動詞_助動詞_名詞」、「名詞_け_て_動詞_助動詞_。」、「代名詞_で_は」、「連体詞_助動詞」、「サ変接続_動詞_動詞_て」、「動詞_助動詞_ば」、「動詞_動詞_助動詞_名詞_、」、「接頭詞_名詞_は」と「名詞_か_に」が現れた。横光利一作品では、このような変数が多く用いられている。

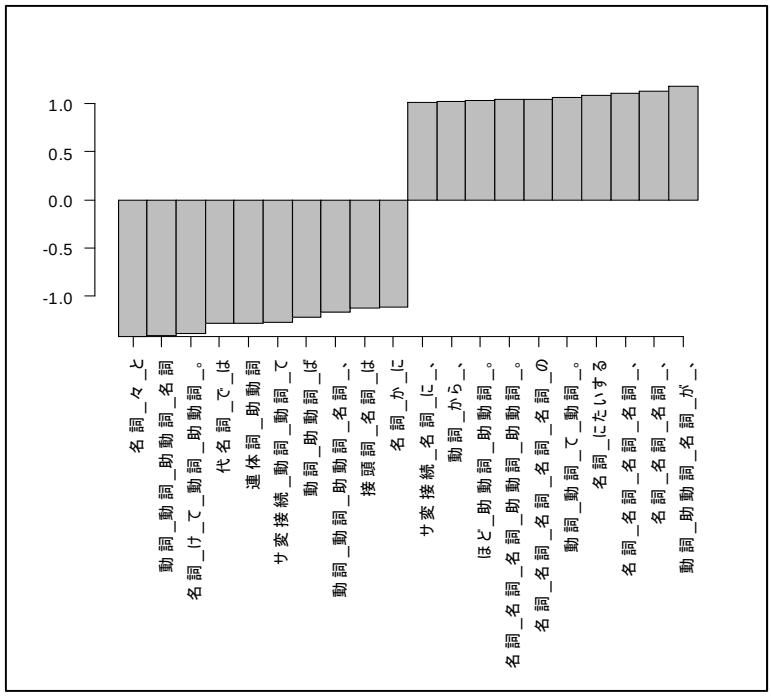

図 9.24 川端康成と横光利一の文節パターン変数の第 1 スコアの棒グラフ

文字記号 bi-gram、タグ付き形態素と文節パターンを用いたクラスター分析の結果を図 9.25~9.27 に示す。図 9.25~9.27 では、いずれも川端康成と横光利一で大きく二つのクラスターを形成し、川端康成の作品は左側のクラスター、横光利一の作品は右側のクラスターに入っている。

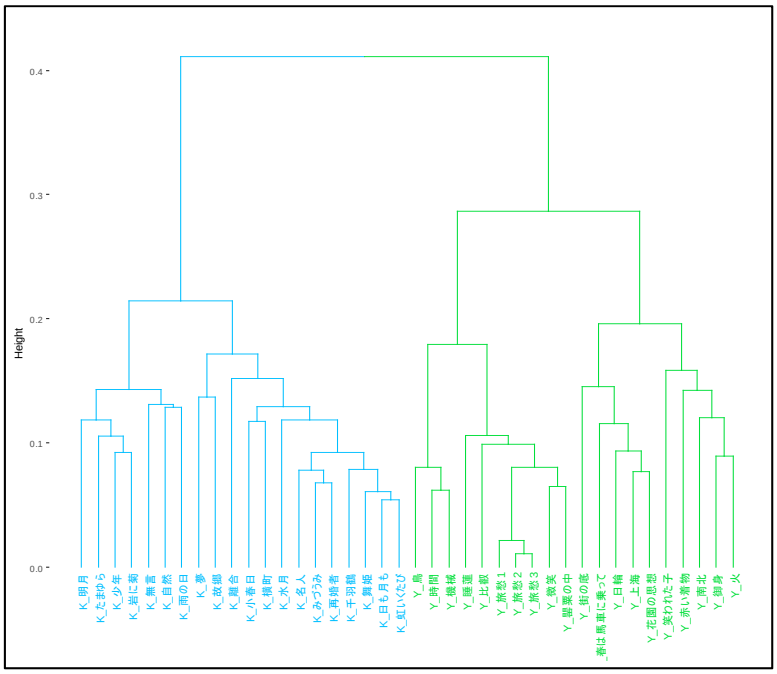

図 9.25 川端康成と横光利一の文字記号 bi-gram の階層的クラスター樹形図

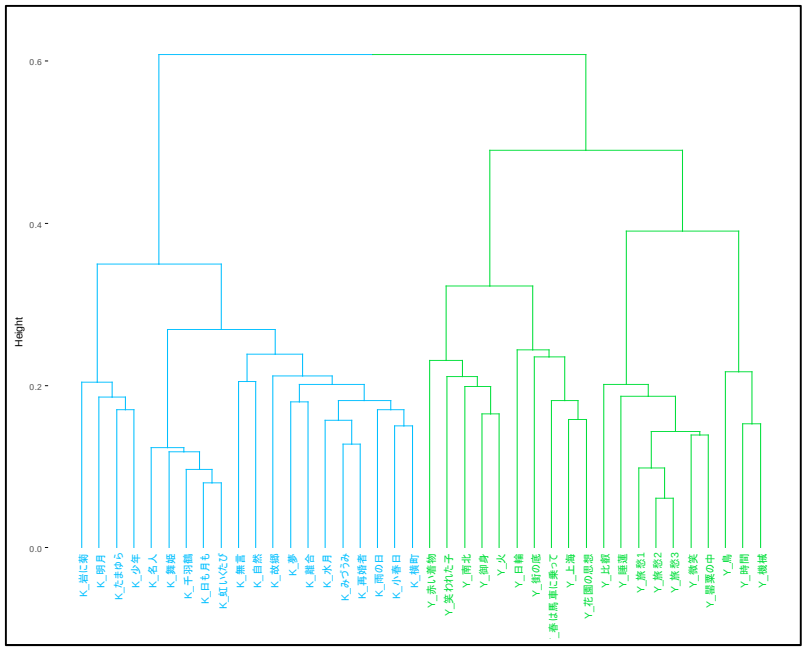

図 9.26 川端康成と横光利一のタグ付き形態素の階層的クラスター樹形図

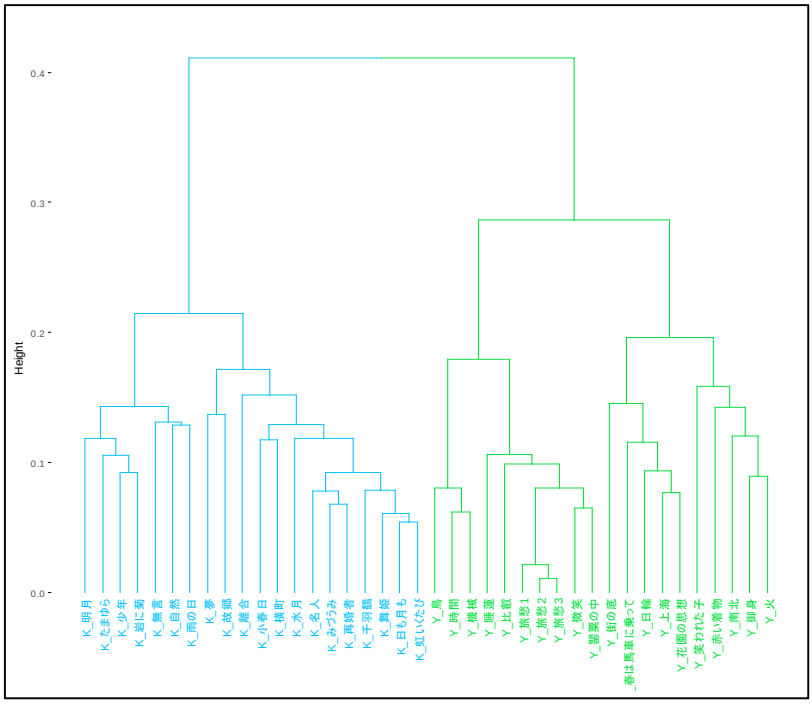

図 9.27 川端康成と横光利一の文節パターンの階層的クラスター樹形図

9.3 4 人の文体の階層的クラスター分析

9.2 節では、先行研究に従って 4 人の作品の一対比較を行った。本節では、この 4 人の作品を総合的に考察するために、文字記号 bigram、タグ付き形態素と文節パターンを特徴量として用いた場合の階層的クラスター分析の結果を紹介する。

文字記号 bigram を用いた階層的クラスター分析の樹形図を図 9.28 に示す。図 9.28 では、川端康成の作品は左側から 3 番目のクラスターに集まっている。

タグ付き形態素を用いた階層的クラスター分析の樹形図を図 9.29 に示す。図 9.29 では、川端康成の作品は左側から 2 番目のクラスターに集まっている。

文節パターンを用いた階層的クラスター分析の樹形図を図 9.30 に示す。図 9.30 では、川端康成の作品は左側から 3 番目のクラスターに集まっている。

以上の 4 人の作品のクラスター分析では、川端康成の作品は単独で一つのクラスターに集まっているため、ほかの 3 人とははっきり分れたため文体が存在すると言える。

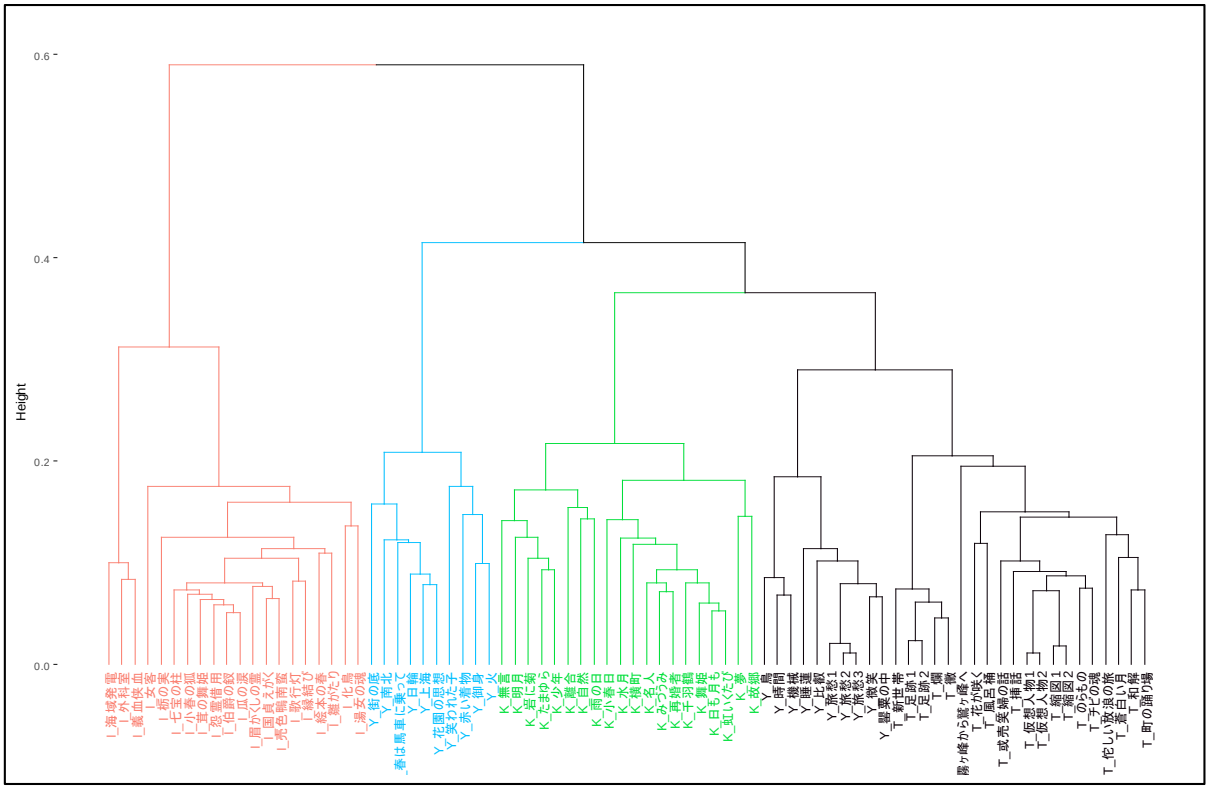

図 9.28 4 人の文字記号の bigram の階層的クラスター樹形図

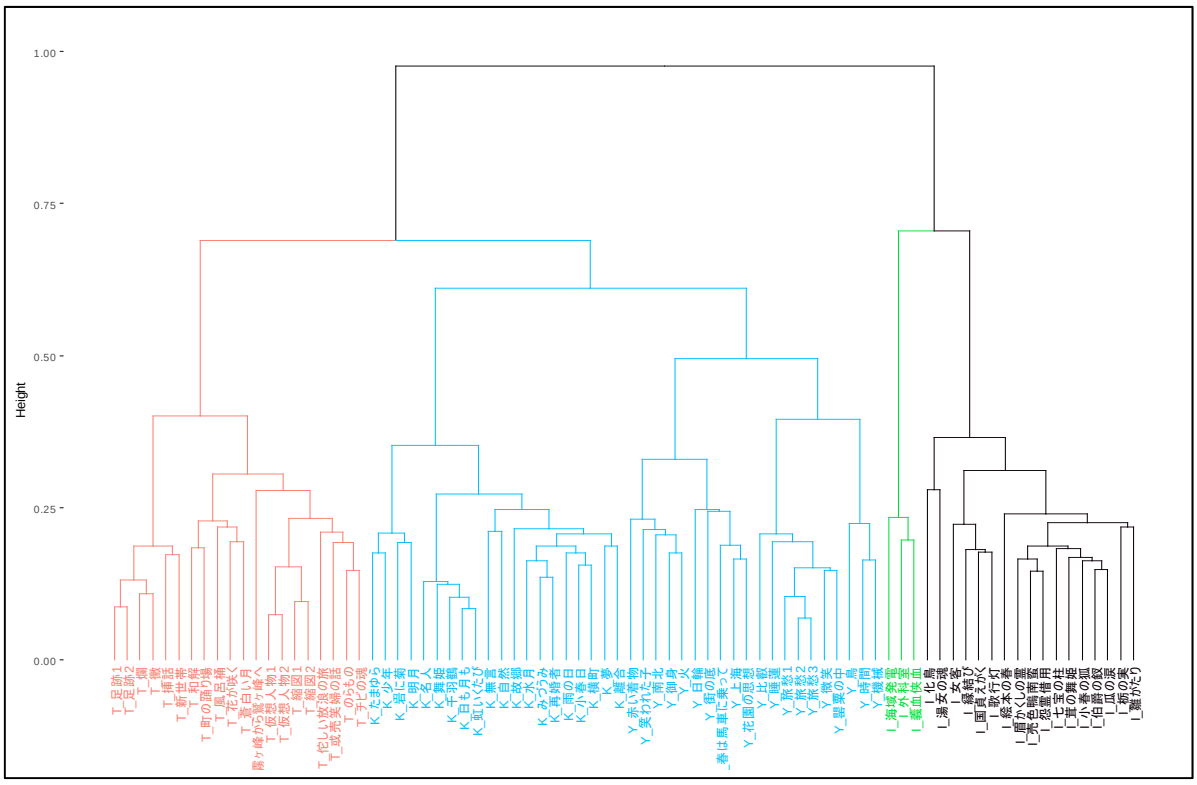

図 9.29 4 人のタグ付き形態素の階層的クラスター樹形図

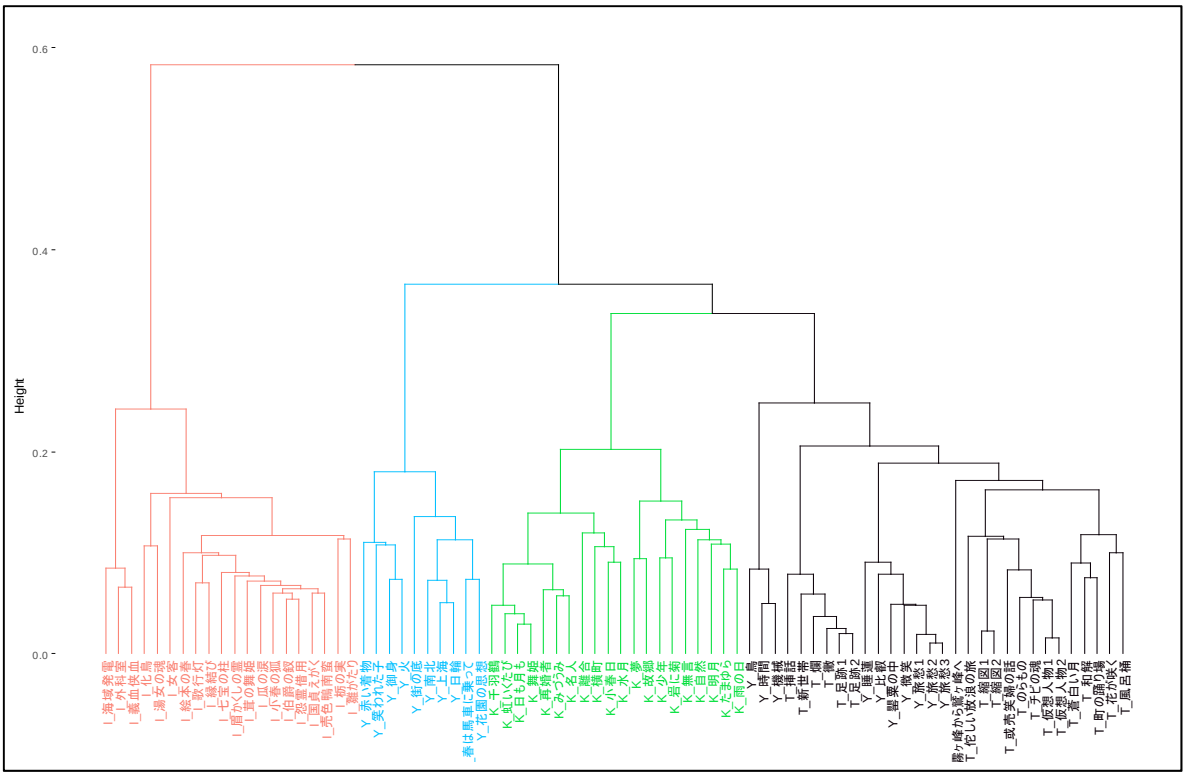

図 9.30 4 人の文節パターンの階層的クラスター樹形図

9.4 本章のまとめ

本章では、先行研究を踏まえ、計量的手法を用いて泉鏡花、徳田秋聲と横光利一と文体の比較を行った。文字記号 bi-gram、タグ付き形態素と文節パターンを文体特徴量とし、対応分析、クラスター分析を行った。その結果、川端康成と各作家とは異なるグループを形成し、川端康成が自分の文体を持つことを示した。また、4 人の文体を比較する場合でも、川端康成の文体の存在が確認された。以上の結果より、泉鏡花、徳田秋聲と横光利一の文体と比較する場合、川端康成の文体は存在すると言える。

第 10 章 川端康成の文体の変化問題

第二次世界大戦の敗戦と親友の相次いだ死は川端康成に精神的打撃を与え、川端文学も敗戦を境にして大きく変貌を遂げた。そこで、川端文学は大きく前期と後期で分かれている (山中, 1999)。特に、後期文学の特徴は川端文学の「魔界」として広く知られている (富岡, 2014; 李, 2013)。この「魔界」の定義に関して、原 (1984)は、以下のように述べ、「魔界」の思想は戦後作品に多く現れたことを示した。

> さてそれでは川端文学における「魔界」とは何か。一般的には頽廃的官能美の横溢する『眠れる美女』 (昭35.1-36.1)あたりをもってその典型とし、「魔界」なる語の初出を見る『舞姫』 (昭25.12-昭26.3)をその最初の発現としているが、作家論的視点から作品世界の「魔界」を形象する作者の肉聲を聞こうとすれば、『みづうみ』 (昭29.1-12)こそ欠くことのできない作品であり、その最初の発見は『みづうみ』とまさしく相似の世界が描かれた『反橋』 (昭23.10)、『しぐれ』 (昭24.1)、『住吉』 (昭24.4)の『反橋』三部作の中に求められるのだ。のちに『舞姫』を書く中で一休の偈頌＜仏界易入魔界難入＞に出会い、三部作において無自覚的に描かれた世界を「魔界」と名づけえた川端は、さらにそこに人間川端の肉聲のみならず芸術家川端の苦悩の精神をも盛って『みづうみ』の中に「魔性」の人間桃井銀平を登場させたのである。

敗戦をきっかけとして、川端康成の文体には変化が生じたかを明らかにするために、本章では、川端康成の 1969 年の全集から 90 編の小説を選んで研究の対象とした。選ばれた小説のリストを付録 1 に示す。その中に執筆期間が数年にわたる長編小説もあったが、執筆途中の校正の可能性を考え、分析では、小説の年度を執筆終了の年にした。文体変化の研究に用いた指標は文字記号 bi-gram、タグ付き形態素と文節パターン、語彙の豊富さ、平仮名の比率と主要品詞の比率である。

10.1 文体特徴量による分析

本節では、文体特徴量の文字記号 bi-gram、タグ付き形態素と文節パターンを用いて川端康成の文体変化を分析する。川端康成の文体は敗戦をきっかけとして変化が生じた場合、1945 年を境に川端康成の作品は二つのグループに分かれると考えられる。

文字記号 bi-gram、タグ付き形態素と文節パターンに基づいた対応分析の結果を図 10.1~10.3 に示す。図 10.1~10.3 から、いくつかの中心から離れた作品を除き、大多数の戦前と戦後の作品はグラフの原点を中心として集まり、はっきりとしたグループ分けが見られない。つまり、本節で用いた文体特徴量では、終戦を境とする川端康成の文体変化が見られなかった。

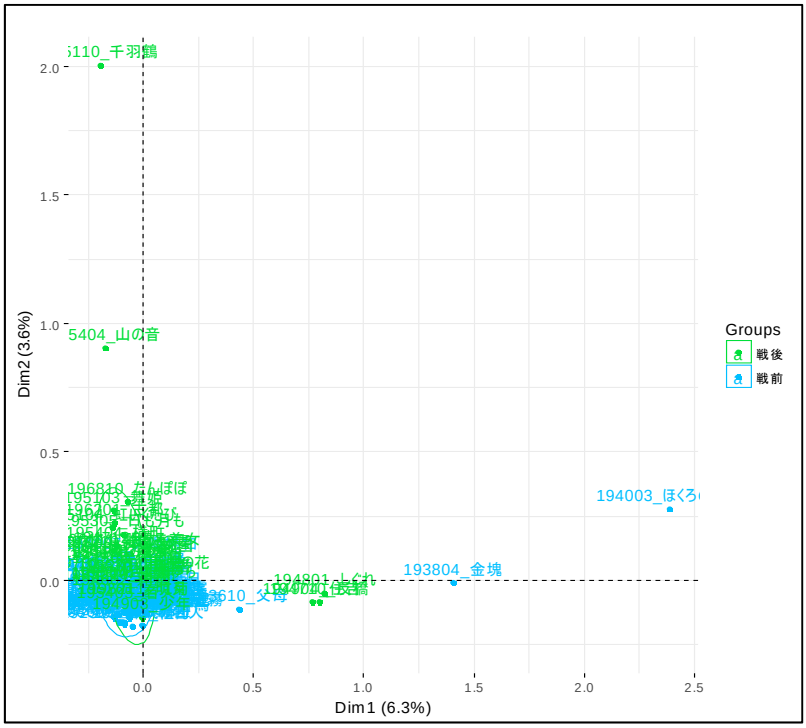

図 10.1 川端康成小説における文字記号 bi-gram 対応分析個体の第 1、2 スコアの散布図

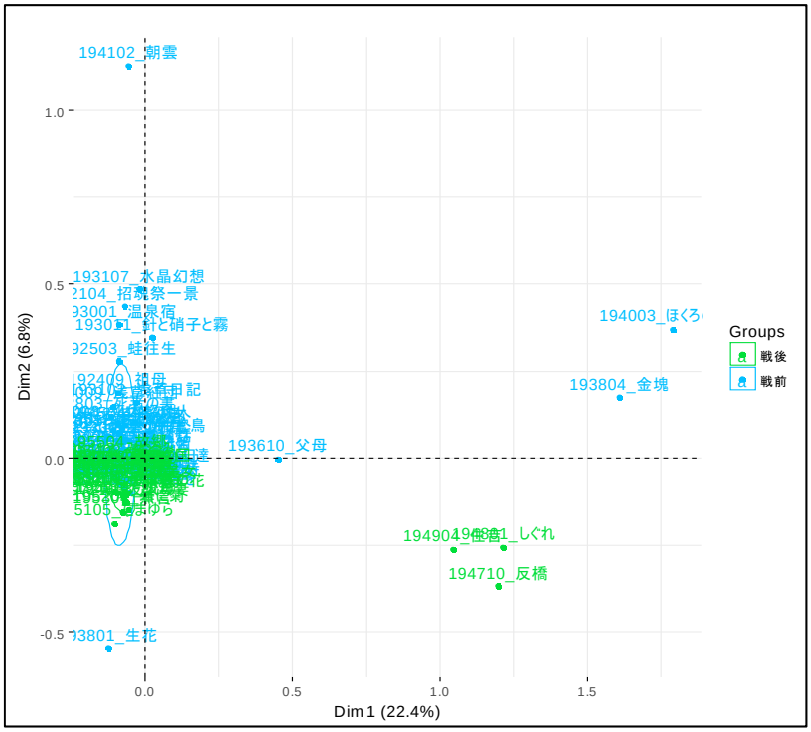

図 10.2 タグ付き形態素を用いた川端康成小説の個体の第 1、2 スコアの散布図

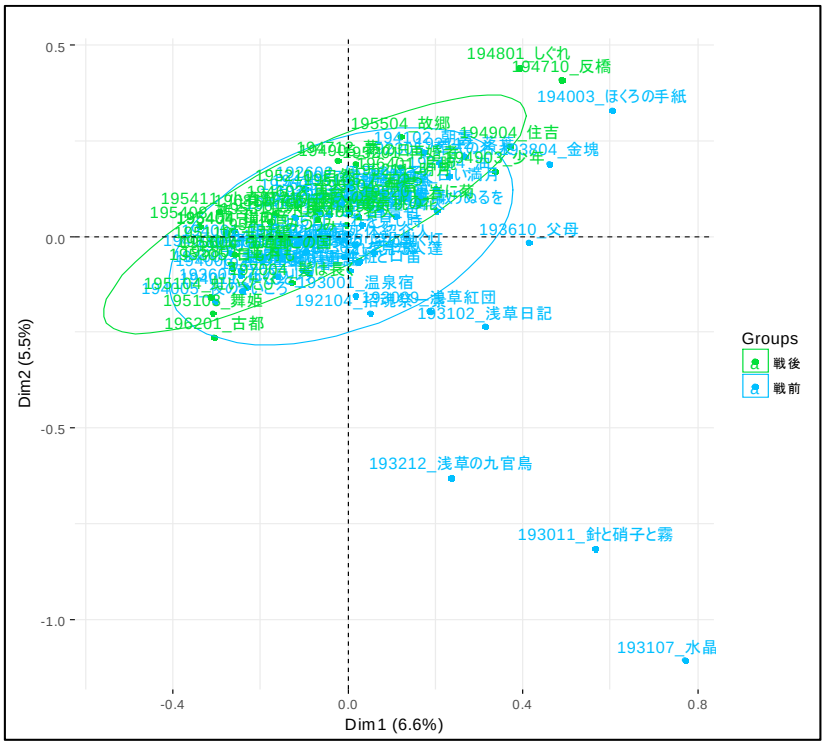

図 10.3 文節パターンを用いた川端康成小説の個体の第 1、2 スコアの散布図

10.2 語彙の豊富さ

語彙は文の基本単位である。作家の文体を表す一つの指標として語彙の豊富さが挙げられる。本節では、語彙の豊富さを表す指標の s 値の経年変化を図 10.4 に示す。この図から s 値の下がっている傾向が見られる。戦前作品の s 値の平均値は 0.91 で、戦後作品の s 値の平均値は 0.9 である。このデータに対して平均の差の検定を行った結果、p 値は 4.96×10^{-11} で、効果量の d は 1.64 で大である。この結果から敗戦を境に語彙の豊富さには差があることが確認された。

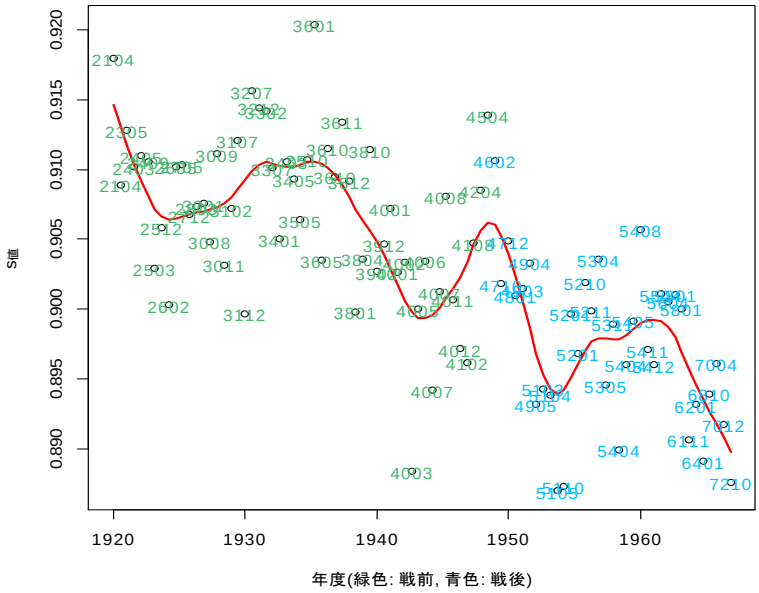

図 10.4 川端康成作品における語彙の豊富さの経年変化

10.3 主要品詞の比率の経時変化

日本語の現代小説文において、品詞の出現率順で並べると、助詞 (30%~40%)、名詞 (20%~30%)、動詞 (15%~20%)である。このような品詞は、日本語で文を書く以上だれでも使わなくてはいけないものであるため、文体に変化が生じた場合品詞の比率に反映される可能性は高いと思われる。本節ではこの 3 種類の品詞のほかに、形容詞、副詞、接続詞といった日本語主要品詞の比率について研究を行った。

日本語助詞使用率の経年変化を図 10.5 に示す。戦前作品の名詞使用率の平均値は 0.3 で、戦後作品の名詞使用率の平均値も 0.31 である。このデータに対して平均の差の検定を行った結果、p 値は 0.01 で、効果量 d の値は 0.51 である。図 10.5 と総合的に考えると、この p 値はサンプルサイズの影響によるもである可能性があり、敗戦を境に助詞の使用率に差はあるが、顕著であるとはいいがたい。

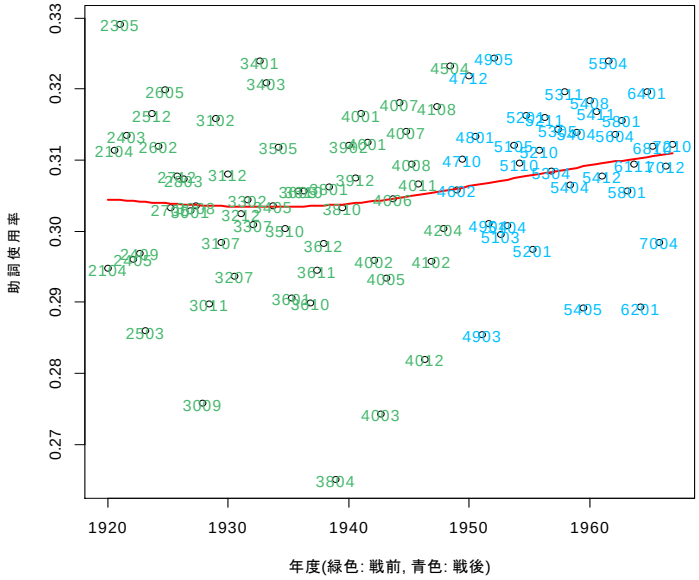

図 10.5 川端康成作品における助詞使用率の経年変化

日本語名詞使用率の経年変化を図 10.6 に示す。戦前作品の名詞使用率の平均値は 0.29 で、戦後作品の名詞使用率の平均値も 0.29 である。このデータに対して平均の差の検定を行った結果、p 値は 0.55 で有意の差がない。

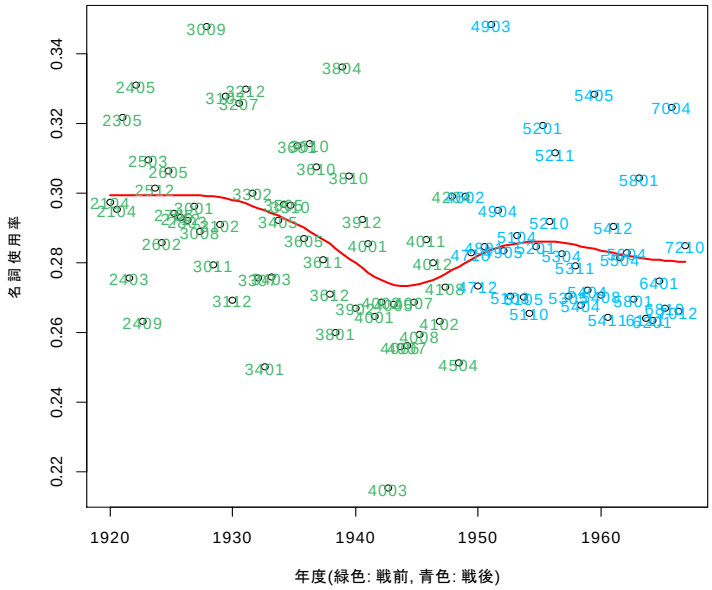

図 10.6 川端康成作品における名詞使用率の経年変化

日本語動詞使用率の経年変化を図 10.7 に示す。戦前作品の動詞使用率の平均値は 0.14 で、戦後作品の動詞使用率の平均値も 0.14 である。このデータに対して平均の差の検定を行った結果、p 値は 0.85 で有意の差がない。

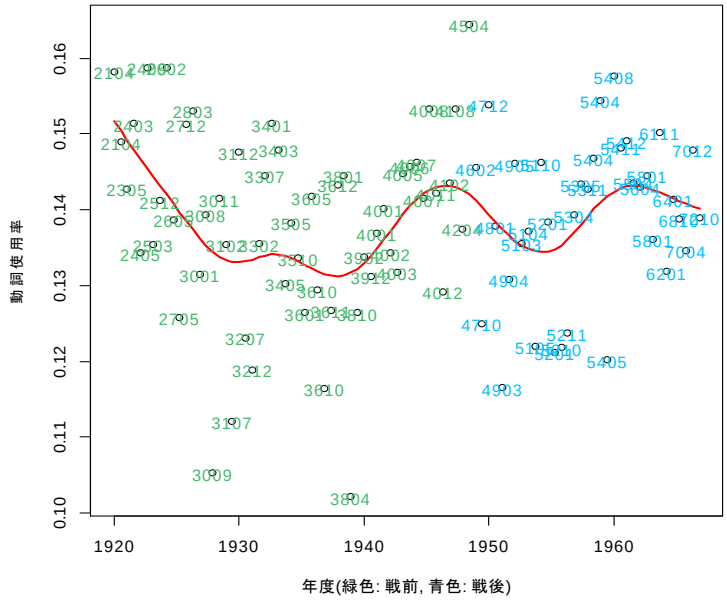

図 10.7 川端康成作品における動詞使用率の経年変化

日本語形容詞使用率の経年変化を図 10.8 に示す。戦前作品の形容詞使用率の平均値は 0.019 で、戦後作品の形容詞使用率の平均値も 0.019 である。このデータに対して平均の差の検定を行った結果、p 値は 0.98 で有意の差がない。

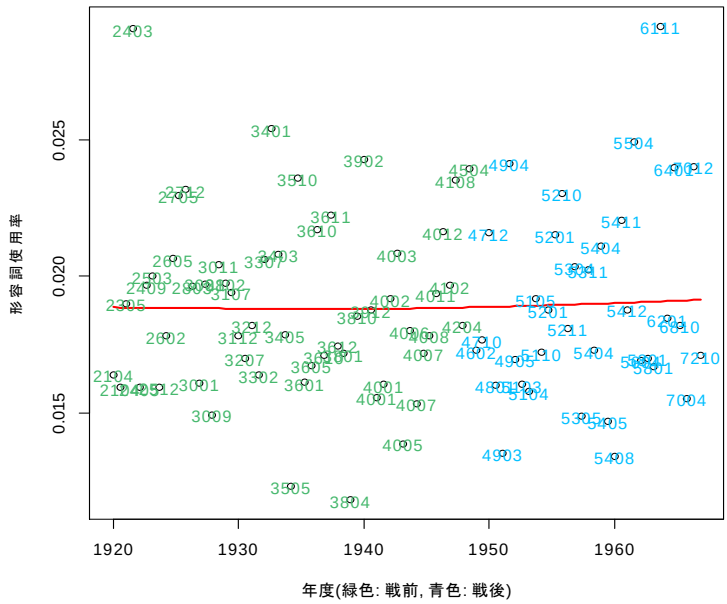

図 10.8 川端康成作品における形容詞使用率の経年変化

日本語副詞使用率の経年変化を図 10.9 に示す。戦前作品の形容詞使用率の平均値は 0.02 で、戦後作品の形容詞使用率の平均値も 0.016 である。このデータに対して平均の差の検定を行

った結果、p 値は 8.18×10^{-8} で、効果量 d の値は 1.25 で大である。この結果から、敗戦を境に副詞の使用率には差があることが確認された。

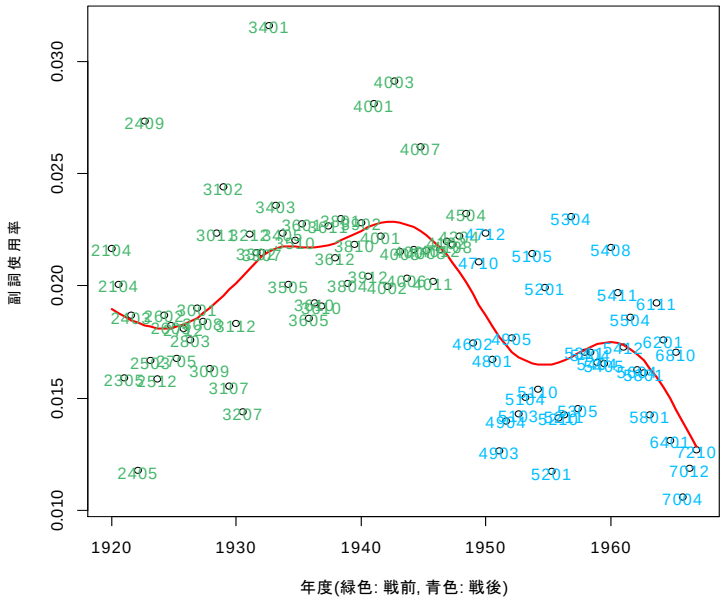

図 10.9 川端康成作品における副詞使用率の経年変化

日本語接続詞使用率の経年変化を図 10.10 に示す。戦前作品の接続詞使用率の平均値は 0.007 で、戦後作品の接続詞使用率の平均値も 0.006 である。このデータに対して平均の差の検定を行った結果、p 値は 0.006 で、効果量 d の値は 0.59 で中である。この結果から敗戦を境に接続詞の使用率に差はあるが、副詞ほど顕著ではない。

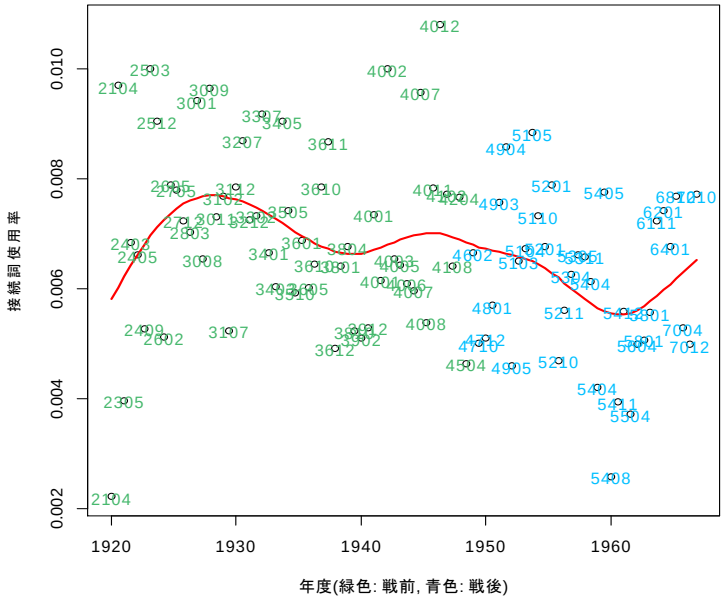

図 10.10 川端康成作品における接続詞使用率の経年変化

10.4 考察

川端康成の文体変化問題について、川端康成 1969 年の全集に収録される 90 編の小説を研究対象とし、川端康成文体の変化時期とされた終戦説の検証を行った。その結果、終戦の 1945 年を境に語彙の豊富さと機能語の助詞、副詞、接続詞に変化があり、文字記号 bigram、タグ付き形態素、文節パターンと内容語の名詞、動詞、形容詞に変化がなかった。

第 11 章 川端康成の語彙問題

五味 (1981)は、著作『魔界』の中に、「川端康成は語彙が乏しく、それをごまかすために平仮名を多用している」と批判しているが、これは誤解であると思われる。なぜなら川端康成は極力平易な言葉のつながりで豊かな表現をしようとしていた。そこで川端康成の作品を読む際に語彙が少ないように感じる。五味の説に関しては、小谷野 (2013)は、著作『川端康成伝－双面の人』の中で、次のように解釈している。川端康成の『眠れる美女』の一段落を次に示す。

> 江口老人も六十七年の生涯のうちには、女とのみにくい夜はもちろんあった。しかもさういふみにくいことの方がかへって忘れられないものである。それはみめかたちのみにくさといふのではなく、女の生のふしあはせなゆがみから来るものであった。江口はこの年になって、女とみにくい出合ひをまた一つ加へたくはない。この家に来ていざとなって、さう思ふのだった。しかし眠らされ通じで目覚めない娘のそばに一夜横たはらうとする老人ほどみにくいものがあらうか。江口はその老いのみにくさの極みをもとめて、この家にきたのではなかったか。

小谷野 (2013)は、この短い文章に「みにくい (さ)」が出てくることに気づき、しかも川端康成の『山の音』でも、「やさしい」が繰り返し出てくると述べた。ここの「みにくい (さ)」は、漢字ではなくて、平仮名で書かれていたので、五味が「平仮名」を多用していると指摘したのは、こういったところであろう。本章では、川端康成の少ない語彙で豊かな表現をするという語彙問題と平仮名多用の問題の解明を試みる。

11.1 語彙問題

本節では、川端康成作品における語彙の豊富さを泉鏡花、徳田秋聲、横光利一と比較する。そのために用いた指標は s 値である。川端康成と泉鏡花、芥川龍之介、徳田秋聲、横光利一の 20 編の小説における s 値を計算し、その箱ひげ図を図 11.1 に示す。

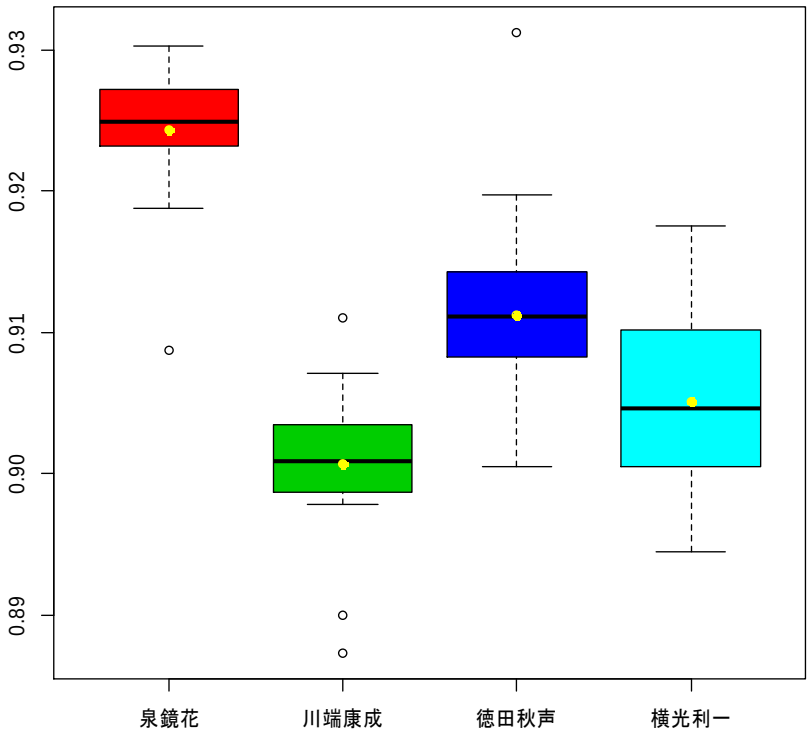

図 11.1 語彙の豊富さ (s 値)の箱ひげ図

図 11.1 から、s 値は泉鏡花、徳田秋聲、横光利一と川端康成の順で小さくなり、ほかの 3 人と比べ、川端康成の語彙がやや少ない。しかし、図 11.1 に示した結果は、川端康成と泉鏡花、徳田秋聲、横光利一の全集からそれぞれ 20 編の小説を抽出して作成したものである。この 20 編はあくまで各作家の全集 (母集団)から得たサンプルで、川端康成と比べて各作家の語彙の豊富さの平均に差があるかを調べるために一元配置分散分析を行った。

ここでの帰無仮説 (H0)を川端康成と泉鏡花、徳田秋聲、横光利一の全集 (母集団)の間に語彙の豊富さを表すs値の差がないとし、対立仮説 (H1)をその差があるとする。分散分析の結果、p値は2×10^{-16}は有意水準を大きく下回っているため、帰無仮説は棄却され、対立仮説が採択される。この分散分析の効果量η^2は0.70で、大である。帰無仮説は棄却されたことから、泉鏡花、徳田秋聲、横光利一の全集 (母集団)の間にs値には差があることが分かった。具体的にどの二群間に差があるかを調べるために、Tukeyの方法による多重比較を行った。その結果を図11.2に示す。

図11.2に示したように、川端康成と横光利一の間には95%信頼空間に0が含まれるため、有意の差がないが、残りのペアの間には間には有意の差があることが分かった。つまり、4人の中で川端康成の語彙は横光利一とほぼ同じで、泉鏡花と徳田秋聲より語彙量はすくないという結論になる。以上の分析により、泉鏡花、徳田秋聲と比べて川端康成の語彙量が少ない方で、少ない語彙で豊かな表現をしたことが実証された。

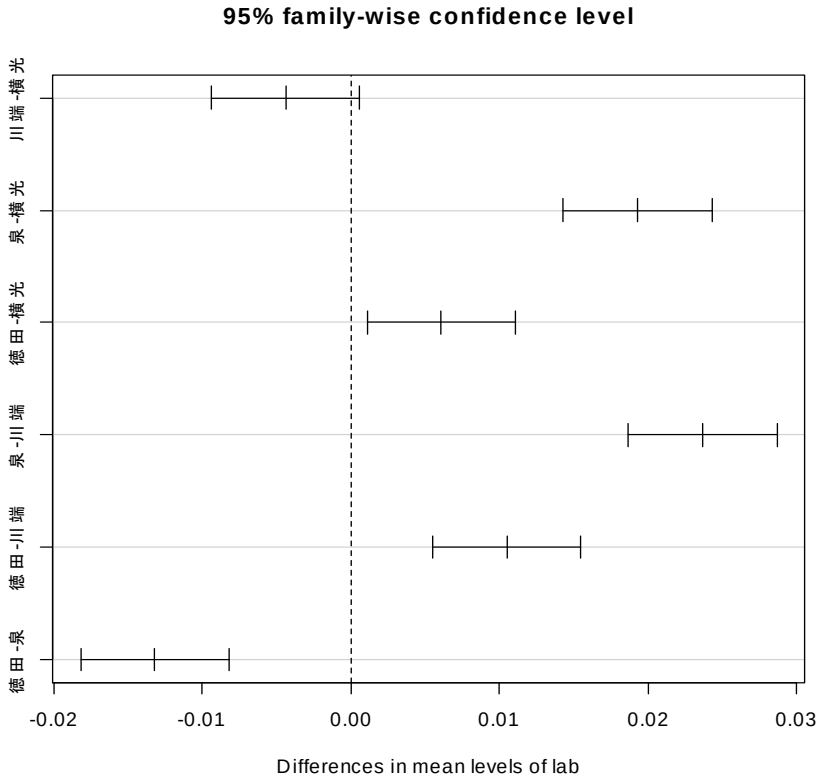

図 11.2 語彙の豊富さ (s 値)の多重比較

11.2 平仮名の多用問題

日本語文の構成要素は漢字、平仮名、片仮名と記号がある。そのうちの記号は文の区切りを表すものであるため、平仮名使用率との関連性がない。平仮名の多用問題のため、指標として平仮名使用率を用いる。

泉鏡花、川端康成、徳田秋聲と横光利一の20編の作品における平仮名使用率の箱ひげ図を図11.3に示す。図11.3の平仮名使用率の値では、川端康成は泉鏡花、徳田秋聲と横光利一より大きい。直観では、確かに川端康成の作品において平仮名の使用は多い。

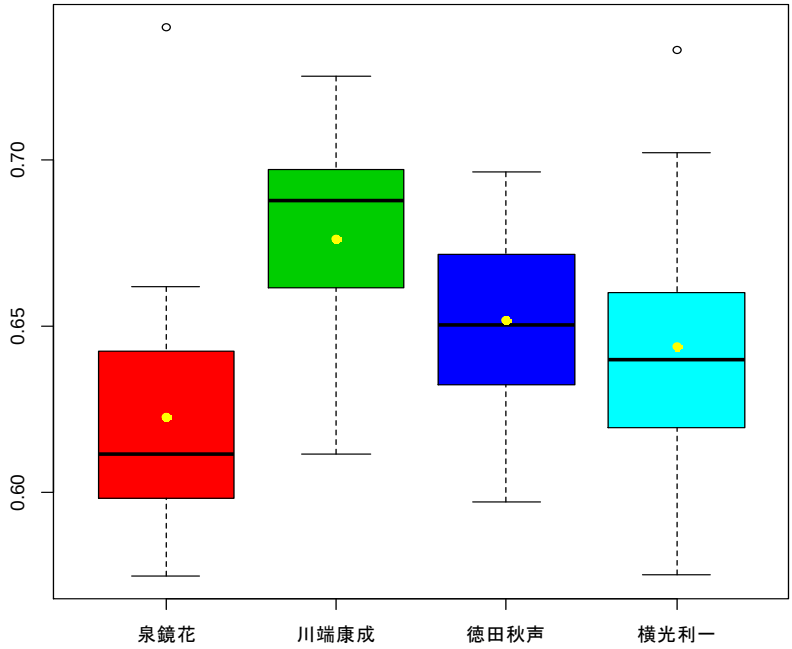

図 11.3 平仮名使用率の箱ひげ図

図11.3に示した結果は、川端康成と泉鏡花、徳田秋聲、横光利一の全集からそれぞれ20編を抽出して作成したものである。この20編はあくまで各作家の全集 (母集団)から得たサンプルで、川端康成と比べて各作家の平仮名使用率に平均の差があるかを調べるために一元配置分散分析を行った。

ここでの帰無仮説 (H0)を川端康成と泉鏡花、徳田秋聲、横光利一の全集 (母集団)の間に平仮名使用率の差がないとし、対立仮説 (H1)をその差があるとする。分散分析の結果、p値は1.29×10^{-5}は有意水準を大きく下回っているため、帰無仮説は棄却され、泉鏡花、徳田秋聲、横光利一の全集 (母集団)の間に平仮名使用率には差がある。この分散分析の効果量η^2は0.27で、大である。Tukeyの方法による多重比較を行った結果を図11.4に示す。

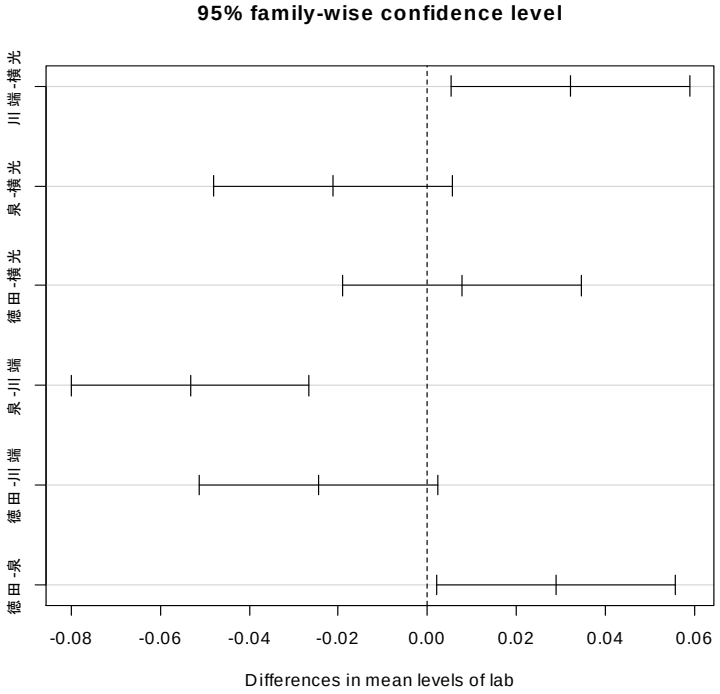

図 11.4 平仮名使用率の多重比較

図11.4に示したように、川端康成と徳田秋聲間には95%信頼区間に0が含まれるため、有意の差がないが、川端康成と泉鏡花、川端康成と横光利一の間には有意の差がある。この結果から、泉鏡花、横光利一と比べ、川端康成の作品には平仮名使用率が高いことが明らかになった。

11.3 本章のまとめ

本章では、川端康成の語彙問題と平仮名多用の問題についての検証を行った。語彙の豊富さ指標を表すs値では、泉鏡花、徳田秋聲と比較する場合、川端康成のs値が小さい。平仮名使用率では、泉鏡花、横光利一と比べ、川端康成の値が大きいことが分かった。

第 12 章 結論と課題

12.1 結論

本論文は、初めて川端康成の代筆問題を体系的に論じたものである。長年にわたって未解決とされた川端康成の代筆問題と文体問題 (存在問題と変化問題)の解明に計量的手法を導入したところに新規性がある。また、本研究の結果は従来の内省法による代筆判断に客観的な証拠を提供している。

川端康成の代筆問題は主に少女小説の『乙女の港』、『花日記』と『コスモスの友』、睡眠薬中毒時期の『古都』と『眠れる美女』、その他に『山の音』が挙げられる。本論文では、文字記号bi-gram、内容語を除いたタグ付き形態素と文節のパターンを文体特徴量とし、対応分析、クラスター分析、AdaBoost、HDDA、LMT、RFとSVMを計量的手法として代筆問題の検証を行った。その結果、少女小説の『乙女の港』、『花日記』は川端康成と中里恒子の共同執筆であることを判明した。『コスモスの友』、『古都』、『眠れる美女』と『山の音』は代筆者が書いた可能性が低いと考えられる。

川端康成の文体問題は、主に、文体の存在問題と変化問題が挙げられる。文体の存在問題では、先行研究を踏まえて泉鏡花、徳田秋聲と横光利一との比較を行い、川端康成作品における文体の存在が確認された。文体の変化問題では、終戦の 1945 年を境に、川端康成の語彙の豊富さと、機能語の助詞、副詞、接続詞に変化が現れた。川端康成は極力平易な言葉のつながりで豊かな表現をしようとしたため、泉鏡花、徳田秋聲と比較する場合、川端康成の文章における語彙の豊富さを表す s 値が小さい。また、平仮名使用率に関しては、泉鏡花、横光利一と比べ、川端康成の値が大きい。

12.2 課題

本論文で取り上げなかった川端康成の代筆疑惑小説として、『空の片仮名』、『歌劇学校』、『万葉姉妹』などが挙げられる。その内『空の片仮名』の代筆者は内田憲太郎、『歌劇学校』は平山宮子、『万葉姉妹』は佐藤碧子とされている。しかし、代筆者の存在が指摘される作品は少なく、代筆問題を解明するための学習データの確保が難しいため、本論文で用いた教師ありの手法は適用できなくなる。このような問題は特異値検出の手法を用いるべきと考えられ、今後の課題として残す。

また、本論文では、すでに川端康成の 1984 年に完結した全集で削除されたものを対象としなかった。しかし、先行研究を調べた限りでは、削除された理由がそれほど記されていないため、削除されたものの中に川端康成本人の作品がある可能性もあると考えられ、今後の課題として取り組んでいきたい。

共著の可能性がある代筆問題を議論するために長い文章を小分けにして議論するのは一般的である。既存の研究では、劉 (2016)は文章を 5000 字ごとに分け、上阪 (2016)は本論文と同様に章ごとに分けた。このような分け方は恣意性を帯び、文章を分割するための境界線は必

ずしも理想的であるとはいえない。本論文でも、分類器による分析では同じ章が異なる著者に分類されるケースが現れた。これは恣意的に分割された各章には複数著者の文体特徴が入っていたことが一因だと考えられる。しかし、日本語ではまだ著者識別の手法に耐える適切な文の長さについての基礎研究がない。著者識別のために文章分割基準と文体の変わり目検出研究も今後の研究課題としたい。

謝辞

本論文を作成するにあたり、指導教員の金明哲先生に計り知れないご指導・サポートを賜りました。金先生の叱咤激励がなければ、修了というゴールにたどりつくことができなかったに違いありません。私はいきなり金先生の研究室に入ってきましたが、これほどの学識、指導力と包容力をお持ちの先生に恵まれて心より幸運だと存じます。最後の最後まで貴重なお時間を割いていただき、長期間に亘りご指導・励まし・見守ってくださり、本当にありがとうございました。

続いて、本論文の審査委員を引き受けてくださいました矢野環先生、山内信幸先生、浦部治一郎先生と山中正樹先生には、ご多忙の中数多くの貴重なご指導・コメントをお寄せいただきましたこと、心より御礼を申し上げます。

さらに、博士後期課程に入学以来、研究の面でずっとサポートしていただいた副指導教員の沈力先生と川崎廣吉先生に厚く御礼を申し上げます。

そして、いつも温かく見守り、さまざまな面でサポートしていただいた同志社大学文化情報学研究科の鄭躍軍先生と同志社大学日本語・日本文化教育センターの李長波先生をはじめとする諸先生方に心からの感謝の意を表します。

統計的手法について教えてくださった滋賀大学特任講師の李鍾賛先生、本論文を読んで頂き、日本語の表現をご指摘いただいた柴田麟太郎さん、尾城奈緒子さんと奥島美葵さんに深く御礼を申し上げます。

いつも研究室のゼミで有意義な議論をしていただき、論文のデータから分析手法に至る数々の改善点を助言していただいたデータサイエンス研究室の入江さやか先生、劉雪琴さん、李広微さん、鄭弯弯さん、尾城奈緒子さん、井口慎也さん、三船正暁さん、柳燁佳さん、黄善玉さん、行村隆平さん、袁徐晟さんに感謝の意を表します。

大阪大学の上阪彩香先生、同志社大学文化情報学研究科元助教の松森智彦先生、元助手の宮武慶之先生、外国人特別助手の陳艷艷先生には、入学以来、研究と教育関連の仕事で大変お世話になりました。記して感謝の意を述べたいと思います。

本論文で用いたデータの一部は、平成 28 年度笹川科学助成金 (研究助成部門)を受けて作成したものです。ここで助成金を交付してくださった笹川財団と、データ作成にご協力いただいた鄭弯弯さんと黄暉さんに感謝の意を申し上げます。

最後に、長期間にわたる日本での留学生活を支えてくれた家族に最高の感謝を捧げます。

参考文献

Abbasi, A., & Chen, H. (2005). Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems, 20*(5), 67-75.

Aljumily, R. (2015). Hierarchical and non-hierarchical linear and non-linear clustering methods to "Shakespeare authorship question". *Social Sciences, 4*(3), 758-799.

Argamon, S., Saric, M., & Stein, S. (2003). Style mining of electronic messages for multiple authorship discrimination: First results. *Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining, USA,* 475-780.

Argamon, S., Whitelaw, C., Chase, P., Hota, S. R., Grag, N., & Levitan, S. (2007). Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology, 58*(6), 802-822.

Baayen, R., van Halteren, H., & Tweedie, F. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing, 11*(3), 121-131.

Bogdanova, D. & Lazaridou, A. (2014). Cross-language authorship attribution. *Proceedings of the Ninth International Conference on Language Resources and Evaluation, Iceland,* 26-31.

Bouveyron, C., Girard, S. & Schmid, C. (2007). High-dimensional discriminant analysis. *Communications in Statistics: Theory and Methods, 36*(14), 2607-2623.

Breiman, L. (2001). Random forests. *Machine Learning, 45,* 5-32.

Bringra, C. S. (1963). Mark Twain and the Quintus Curtius Snodgrass letters: A statistical test of authorship. *Journal of the American Statistical Association, 58,* 85-96.

Burrows, J. F. (1987). Word patterns and story shapes: The statistical analysis of narrative style. *Literature and Linguistic Computing, 2,* 61-70.

Burrows, J. F. (1992). Not unless you ask nicely: The interpretative nexus between analysis and information. *Literary and Linguistic Computing, 7*(2), 91-109.

Delgado, M. F., Cernadas, E., Barro, S., & Amorim, D. G. (2014). Do we need hundreds of classifiers to solve real world classification problems?. *The Journal of Machine Learning Research, 15*(1), 3133-3181.

De Morgan, S. E. (1882). *Memoir of Augustus de Morgan*. London: Longmans, Green, and Co.

De Vel, O., Anderson, A., Corney, M., & Mohay, G. M. (2001). Mining e-mail content for author idenfication forenics. *SIGMOD Record, 30*(4), 55-64.

Eder, M. (2011). Style-markers in authorship attribution a cross-language study of the authorial fingerprint. *Studies in Polish Linguistics, 6,* 99-114.

Eder, M. (2015). Visualization in stylometry: Cluster analysis using networks. *Digital Scholarship in the Humanities, 32*(1), 50-64.

Forsyth, R., & Holmes, D. (1996). Feature-finding for text classification. *Literary and Linguistic Computing, 11*(4), 163-174.

Frantzeskou, G., Stamatatos, E., Gritzalis, S., & Katsikas, S. (2006). Source code author identification based on n-gram author profiles. In Maglogiannis, I., Karpouzis, K., & Bramer, M. (eds), *Artificial Intelligence Applications and Innovati*ons, AIAI 2006. *IFIP International Federation for Information Processing*, Vol. 204, Boston, MA: Springer, 508-515.

Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Informaton and Computation, 121*(2), 256-285.

Gamon, M. (2004). Linguistic correlates of style: Authorship classification with deep linguistic analyais features. *Proceedings of the 20th International Conference on Computational Linguistics, University of Geneva, Switzerland,* 611-617.

Grant, T. D. (2007). Quantifying evidence for forensic authorship analysis. *International Journal of Speech Language and the Law, 14*(1), 1-25.

Grieve, J. (2007). Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing, 22*(3), 251-270.

Habash, A., Guinn, C., Kline, D., & Patterson, L. (2012). Language analysis of speakers with dementia of the Alzheimer's type. *Annals of the Master of Science in Computer Science and Information Systems at UNC Wilmington, 6*(1), paper 11.

Holmes, D. I. (1998). The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing, 13*(3), 111-117.

Hoorn, F. J., Kowalczyk, W., Frank, L. S., & Ham, F. (1999). Neural network identification of poets using letter sequences. *Literary and Linguistic Computing, 14*(3), 311-338.

Jin, M. Z. (1998). Latest developments of stylometry in Japanese. *INFORMATION, 1*(2), 57-64.

Juola, P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieval, 1*(3), 233-334.

Kjell, B. (1994). Discrimination of authorship using visualization. *Information Processing and Management, 30*(1), 141-150.

Koppel, M., & Schler, J. (2003). Exploiting stylistic idiosyncrasies for authorship attribution. *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis, Mexico,* 69-72.

Landwehr, N., Hall, M., & Frank, E. (2005). Logistic model trees. *Machine Learning, 59*(1-2), 161-205.

Lee, J. C., Choe, J. W., & Jin, M. Z. (2017). Authorship attribution of Korean texts by using phrase patterns. *INFORMATION, 20*(1B), 417-428.

Mendenhall, C. T. (1887). The characteristic curves of composition. *Science, 214 S,* 234-246.

Mendenhall, C. T. (1901). A mechanical solution of a literary problem. *Popular Science Monthly, 60*(2), 97-105.

Mosteller, F., & Wallace, D. L. (1964). *Inference and disputed authorship: The Federalist.* Boston, MA: Addison Wesley.

Peng, F. C., Schuurmans, D., & Wang, S. J. (2004). Augumenting naive bayes text classifier with statistical language models. *Information Retrieval, 7*(3-4), 317-345.

Rudman, J. (1998). The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities, 31,* 351-365.

Sanderson, C., & Guenter, S. (2006). Short text authorship attribution via sequence kernels. Markov chains and author unmasking: An investigation. *Proceedings of the International Conference on Empirical Methods in Natural Language Engineering, Australia,* 482-491.

Seroussi, Y., Zukerman, I., & Bohnert, F. (2014). Authorship attribution with topic models. *Computational Linguistics, 40*(2), 269-310.

Stamatatos, E. (2006). Authorship attribution based on feature set subspacing ensembles. *International Journal on Artifical Intelligence Tools, 15*(5), 823-838.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology, 60*(3), 538-556.

Sun, H., & Jin, M. Z. (2017). Verifying the authorship of the Yasunari Kawabata novel The Sound of the Mountain. *Journal of Mathematics and System Science, 7,* 127-141.

Sun, H., & Jin, M. Z. (2016). The relation between stylometry and neuroscience. *Neuroscience and Biomedical Engineering, 4*(3), 174-180.

Tabata, T. (2012). Approaching Dicken's style through random forests. *Digital Humanities 2012 Conference Proceedings, Germany,* 389-391.

Tsuchiyama, G., & Murakami, M. (2013). Authorship identification of classical Japanese literature using quantitative analysis. *Journal of Mathematics and System Science, 3*(12), 631-640.

Uesaka, A., & Murakami, M. (2015). Verifying the authorship of Saikaku Ihara's work in early modern Japanese literature: A quantitative approach. *Digital Scholarship in the Humanities, 30*(4), 599-607.

Upendra, S., Thamar, S., Manuel, M. Y. G., Optica, Y. E., Steven, B., & Paolo, R. (2014). Cross-topic authorship attribution: Will out-of-topic data help?. *Proceedings of the 24th International Conference on Computational Linguistics, Ireland,* 1228-1237.

Vapnik, N. V. (1998). *Statistical Learning Theory*. New York：Wiley.

Yule, U. G. (1938). On sentence-length as a statistical characteristic of style in prose: With application to two cases of disupted authorship. *Biometrika, 30*(3-4), 363-390.

Yu, B. (2012). Function words for Chinese authorship attribution. *Workshop on Computational Linguistics for Literature, Canda,* 45-53.

Zaitsu, W., & Jin, M. Z. (2015). Identifying the author of illegal documents through text mining [テキストマイニングを用いた犯罪に関わる文書の筆者識別]. *Japanese Journal of Forensic Science and Technology, 20*(1), 1-14.

Zhao, Y., & Zobel, J. (2005). Effective and scalable authorship attribution using function words. *Proceedings of the 2nd Asia Information Retrieval Symposium, Korea*, 174-189.

Zhao, Y., & Zobel, J. (2007). Searching with style: Authorship attribution in classic literature. *Proceedings of the Thirtieth Australasian Computer Science Conference, Australia*, 59-68.

Zheng, R., Li, J., Chen, H., & Huang, Z. (2006). A framework for authorship attribution. *Proceedings*

of the 2nd Asia Information Retrieval Symposium, Korea, 92-105.
Zipf, K. G. (1932). *Selected studies of the principle of relative frequency in language*. Cambridge, MA: Harvard University Press.
揚妻祐樹, 木村義之, 小林明子, 小林肇, 森雄一, 渡邊ゆかり (2015). 『文章と文体』, 朝倉書店.
板坂剛 (1997). 『極説三島由紀夫―切腹とフラメンコ』, 夏目書房.
板坂剛, 鈴木邦男 (2010). 『三島由紀夫と一九七〇年』, 鹿砦社.
今村潤子 (1988). 『川端康成研究』, 審美社.
上阪彩香 (2016). 「西鶴浮世草子の文章に関する数量的研究―遺稿集を中心とした著者の検討―」, 同志社大学博士論文.
臼井吉見 (1952). 『川端康成の文体』, 文学会.
内田静枝 (2009). 「解題『乙女の港』と『少女の友』」, 川端康成著『少女の友』, 実業の日本社.
梅原猛 (1985). 『神々の流竄』, 集英社.
大森郁之助 (1991). 「『乙女の港』・その地位の検証 : lesbianism の視点ほか、または、八木洋子頌」, 『札幌大学女子短期大学部紀要』, 17, A1-A18.
尾城奈緒子 (2016). 「太宰治の文体の計量分析―助詞を中心として―」, 『行動計量学会第 44 回大会抄録集』, 164-165.
小野洋平 (2015).「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―」,『計量国語学』, 29 (8), 296-312.
樺島忠夫 (1954). 「現代文における品詞の比率とその増減の要因について」, 『国語学』, 18, 15-20.
樺島忠夫 (1955). 「分類した品詞の比率に見られる規則性」, 『国語国文』, 385-387.
樺島忠夫 (1963). 『表現論―言葉と言語行動』, 総芸舎.
樺島忠夫, 寿岳章子 (1965). 『文体の科学』, 総芸舎.
川勝麻里 (2009). 「川端康成『コスモスの友』は中里恒子代作か: 川端『純粋の聲』の感想文草稿を手掛かりに」, 『明海大学教養論文集』, 20, 55-64.
川端康成記念会 (1981). 『川端康成全集 補巻 2』, 新潮社.
川端康成記念会 (1982). 『川端康成全集 第 33 巻 評論 5』, 新潮社.
金明哲 (1997). 「助詞の分布に基づいた日記の書き手識別」, 『計量国語学』, 20 (8), 357-367.
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』, 33 (5), 225-240.
金明哲 (2003). 「自己組織化マップと助詞分布を用いた書き手の同定およびその特徴分析」, 『計量国語学』, 23 (8), 369-386.
金明哲 (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.
金明哲 (2009). 「文書の執筆時期の推定―芥川龍之介の作品を例として―」, 『行動計量学』, 36 (2), 89-103.

金明哲 (2013). 「文節パターンに基づいた文章の書き手の識別」, 『行動計量学』, 40 (1), 17-28.
金明哲 (2014). 「統合的分類アルゴリズムを用いた文章の書き手識別」, 『行動計量学』, 41 (1), 35-46.
金明哲 (2016). 『定性的データ分析』, 共立出版.
栗原雅直 (1982). 『川端康成精神医学者による作品分析』, 中央公論社.
河野仁昭 (1995). 『川端康成—内なる古都』, 京都新聞社.
小関有希 (2012). 「中里恒子著作目録—『まりあんぬもの』の可能性—」, 『リテラシー史研究』, 5, 13-23.
小林一郎 (1982). 『川端康成研究—東洋的な世界—』, 明治書院.
小林雄一郎, 小木曽智信 (2013).「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—」, 『国立国語研究所論集』, 29-43.
五味康祐 (1981). 「魔界」,『五味康祐代表作集』, 第 10 巻, 新潮社, 213-226.
小谷野敦 (2013). 『川端康成伝—双面の人』, 中央公論新社.
小谷野敦, 深澤晴美 (編) (2016). 『川端康成詳細年譜』, 勉誠出版.
木幡瑞枝 (1992). 『川端康成作品論』, 勁草書房.
佐伯彰一 (1959). 『川端康成の文体』, 角川書店.
下條正純 (2009). 「川端康成『乙女の港』の人物関係と女学生ことば (レトリックの眼で見た世界-虚偽・悪文・映画・判決)」, 『表現研究』, 40-49.
孫昊, 李鍾賛, 金明哲 (2015). 「データから見る川端康成のゴーストライダー問題—『乙女の港』は誰の作品なのか」, 『日本行動計量学会第 43 回大会抄録集』, 216-217.
孫昊, 李鍾賛, 金明哲 (2015). 「データに基づいた『コスモスの友』の著者推定」, 『計量国語学会第 59 回大会予稿集』, 1-6.
孫昊, 李鍾賛, 金明哲 (2015). 「データ解析に基づいた『花日記』の代作問題検証」, 『情報処理学会研究報告』, 1-6.
土山玄 (2015). 「計量文献学による『源氏物語』の成立に関する研究」, 同志社大学博士論文.
寺田透 (1949). 『作家私論』, 改造社.
富岡幸一郎 (2014). 『川端康成 魔界の文学』, 岩波書店.
鳥羽一英 (1969). 「出発まで (1): 川端康成論のうち・孤児根性について」, 『愛知教育大学研究報告, 人文科学』, 18, 19-31.
羽鳥徹哉, 原善 (1998). 『川端康成全作品研究事典』, 勉誠出版.
中嶋展子 (2010). 「川端康成『乙女の港』論—『魔法』から『愛』へ—中里恒子草稿との比較から」, 『岡山大学大学院社会文化科学研究科紀要』, 29, 1-14.
中村明 (1993). 『日本語の文体』, 岩波書店.
中村明 (2009). 『文体論の展開—文藝への言語的アプローチ』, 明治書院.
韮沢正 (1965). 「由良物語の著者の統計的判別」, 『計量国語学』, 33, 21-38.
波多野完治 (1950). 『現代文章心理学—小説・新聞・論文のスタイル』, 新潮社.
原善 (1987). 『川端康成の魔界』, 有精堂.

馬場重行 (1981). 「川端康成の少女小説－『乙女の港』をめぐって」, 『川端康成研究』.
林武志 (1982). 『鑑賞日本現代文学第 15 巻川端康成』, 角川書店.
松浦司, 金田康正 (2000).「n-gram の分布を利用した近代日本語の著者推定」,『計量国語学』, 22 (6), 255-238.
松村明編 (2006). 『大辞林 (第三版)』, 三省堂.
三島由紀夫 (1956). 「永遠の旅人—川端康成の人と作品」, 『別册文藝春秋』, 51.
三田英彬 (1994). 「『抒情歌』と川端の世界観」, 『学海』, 10, 71-79.
ミルカイヴィッチ著, 早上輝洋, 井上史雄訳 (1974). 『言語学の流れ』, みすず書房.
村上征勝, 伊藤瑞叡 (1991). 「日蓮遺文の数理研究」, 『東洋の思想と宗教』, 8, 27-35.
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, 金明哲, 土山玄, 上阪彩香 (2016). 『計量文献学の射程』, 勉誠出版.
村上征勝, 古瀬順一 (2001). 「川端作品の文章の計量分析」, 『日本行動計量学会大会発表論文抄録集』29, 306-307.
ラボック P. 著, 佐伯彰一訳 (1957), 『小説の技術』, 株式会社ダヴィッド社.
李聖傑 (2014). 『川端康成の「魔界」に関する研究ーその生成を中心に』, 早稲田大学出版会.
李賢平 (1987). 「『紅楼夢』成書新説」, 『Fudan 学報 (社会科学版)』, 第 5 期, 3-16.
劉雪琴 (2016). 「計算機統計学のアプローチによる宇野浩二の文体分析」, 『日本計算機統計学会大会論文集』, 30, 17-20.
山田吉郎 (1980). 『「古都」の精神構造』, 教育出版センター.
山中正樹 (1999). 「川端康成の戦後・序説：川端康成と敗戦」, 『桜花学園大学研究紀要』, 1, A41-A51.
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, 147-156.
安本美典 (1959). 「文章の性格学への基礎研究—因子分析法による現代作家の分類」, 『国語国文』, 19-41.
安本美典 (2009). 「計量文体論・文章心理学」, 『計量国語学事典』, 朝倉出版, 253-563.

付録 A

本論文で用いた川端康成1969年全集の小説リストを次の表に示す。

表A1 本論文の対象となる川端康成小説リスト

発表時期	作品名	雑誌名
大正 10 (1921)年 4 月	招魂祭一景	新思潮
大正 10 (1921)年 7 月	油	新思潮
大正 12 (1923)年 5 月	葬式の名人	文芸春秋
大正 13 (1924)年 3 月	篝火	新小説
大正 13 (1924)年 5 月	空に動く灯	我観
大正 14 (1925)年 3 月	蛙往生	文芸時代
大正 14 (1925)年 12 月	白い満月	新小説
大正 15 (1926)年 1 月〜2 月	伊豆の踊子	文芸時代
大正 15 (1926)年 5 月	文科大学挿話	女性
昭和 2 (1927)年 4 月〜5 月	春景色	文芸時代
昭和 3 (1928)年 3 月	死者の書	文芸春秋
昭和 4 (1929)年 4 月〜 昭和 5 (1930)年 8 月	死体紹介人	文芸春秋
昭和 4 (1929)年 10 月〜 昭和 5 (1930)年 1 月	温泉宿	改造
昭和 5 年 (1930)9 月 昭和 5 年 (1930)9 月	浅草紅団	新潮 改造
昭和 5 年 (1930)11 月	針と硝子と霧	文学時代
昭和 6 (1931)年 1 月 昭和 6 (1931)年 2 月	浅草日記	週刊朝日 新潮
大正 15 (1924)年 9 月	祖母	文芸時代
昭和 2 (1927)年 8 月~12 月	南方の火	中外商業新報
昭和 6 (1931)年 12 月	落葉	改造
昭和 6 年 (1931)1 月~7 月	水晶幻想	改造
昭和 7 (1932)年 7 月	それを見た人達	改造
昭和 7 (1932)年 6 月〜12 月	浅草の九官鳥	モダン日本
昭和 7 (1932)年 9 月〜 昭和 11 (1936)年 10 月	化粧と口笛	東京朝日新聞
昭和 8 (1933)年 2 月	二十歳	改造
昭和 8 (1933)年 7 月	禽獣	改造
昭和 8 (1933)年 11 月〜	散りぬるを	改造

昭和 9 (1934)年 5 月		
昭和 9 (1934)年 3 月	虹	中央公論
昭和 9 (1934)新年特別号	夢の姉	週刊朝日
昭和 10 (1935 年)5 月	田舎芝居	中央公論
昭和 10 (1935 年)10 月	童謡	改造
昭和 11 (1936 年)1 月	イタリアの歌	改造
昭和 11 (1936 年)11 月	これを見し時	文芸春秋
昭和 11 (1936 年)4 月 昭和 11 (1936 年)5 月	花のワルツ	改造
昭和 11 (1936 年)10 月	父母	改造
昭和 11 (1936 年)12 月	夕映少女	333
昭和 13 (1938 年)1 月	生花	中央公論
昭和 13 (1938 年)4 月	金塊	改造
昭和 13 (1938 年)10 月	百日堂先生	文芸春秋
昭和 13 (1938 年)12 月 昭和 14 (1939 年)12 月	高原	日本評論 (高原) 公論 (樅の家)
昭和 14 (1939 年)2 月	故人の園	大陸
昭和 15 (1940 年)1 月	正月三ヶ日	中央公論
昭和 15 (1940 年)1 月	母の初恋	婦人公論
昭和 15 (1940 年)2 月	女の夢	婦人公論
昭和 15 (1940 年)3 月	ほくろの手紙	婦人公論 (悪妻の手紙)
昭和 15 (1940 年)5 月	夜のさいころ	婦人公論
昭和 15 (1940 年)6 月	燕の童女	婦人公論
昭和 15 (1940 年)7 月	夫唱婦和	婦人公論
昭和 15 (1940 年)7 月	日雀	文芸春秋
昭和 15 (1940 年)8 月	子供一人	婦人公論
昭和 15 (1940 年)11 月	ゆくひと	婦人公論
昭和 15 (1940 年)12 月	年の暮	婦人公論
昭和 16 (1941 年)1 月 昭和 16 (1941 年)2 月 昭和 17 (1942 年)4 月	寒風	日本評論 (寒風) 改造 (冬の事) 改造 (赤い足)
昭和 16 (1941 年)2 月	朝雲	新女苑
昭和 20 (1945 年)4 月	冬の曲	文芸
昭和 10 (1935 年)1 月 昭和 10 (1935 年)1 月 昭和 10 (1935 年)11 月 昭和 10 (1935 年)12 月	雪国	文芸春秋 (夕景色の鏡) 改造 (白い朝の鏡) 日本評論 (物語) 日本評論 (徒労)

昭和 11 (1936 年)8 月 昭和 11 (1936 年)10 月 昭和 12 (1937 年)5 月 昭和 15 (1940 年)12 月 昭和 16 (1941 年)8 月 昭和 21 (1946 年)5 月 昭和 22 (1947 年)10 月		中央公論 (萱の花) 文芸春秋 (火の枕) 改造 (手毬歌) 公論 (雪中火事) 文芸春秋 (天の河) 暁鐘 (雪国抄) 小説新潮 (続雪国)
昭和 21 (1946 年)2 月	再会	世界
昭和 22 (1947 年)10 月	反橋	風雪別冊
昭和 22 (1947 年)12 月	夢	婦人文庫
昭和 23 (1948 年)1 月	しぐれ	文芸往来
昭和 24 (1949 年)5 月	雨の日	素直
昭和 24 (1949 年)4 月	住吉	個性
昭和 25 (1950 年)3 月~ 昭和 26 (1951 年)4 月	虹いくたび	婦人生活
昭和 23 (1948 年)1 月~5 月、8 月 昭和 27 (1952 年)1 月	再婚者	新潮 「再婚者の手記」に改題
昭和 24 (1949 年)5 月 昭和 24 (1949 年)8 月 昭和 25 (1950 年)1 月 昭和 25 (1950 年)11 月 昭和 25 (1950 年)12 月 昭和 26 (1951 年)10 月	千羽鶴	読物時事別冊 (千羽鶴) 別冊文芸春秋 (森の夕日) 小説公園 (絵志野) 小説公園 (母の口紅) 小説公園 (母の口紅続) 別冊文芸春秋 (二重星)
昭和 24 (1949 年)9 月 昭和 24 (1949 年)10 月 昭和 24 (1949 年)10 月 昭和 24 (1949 年)12 月 昭和 25 (1950 年)1 月 昭和 25 (1950 年)5 月 昭和 25 (1950 年)10 月 昭和 26 (1951 年)10 月 昭和 26 (1951 年)3 月 昭和 26 (1951 年)10 月 昭和 27 (1952 年)10 月 昭和 27 (1952 年)12 月 昭和 28 (1953 年)4 月 昭和 28 (1953 年)4 月	山の音	改造文芸 (山の音) 群像 (蝉の羽 (日まはり)) 新潮 (雲の炎) 世界春秋 (栗の実) 世界春秋 (続き (女の家)) 改造 (島の夢) 新潮 (冬の桜) 文学界 (朝の水) 群像 (夜の聲) 別冊文芸春秋 (春の鐘) 新潮 (鳥の家) 新潮 (都の苑) 別冊文芸春秋 (傷の後) 改造 (雨の中)

昭和 28 (1953 年)10 月 昭和 28 (1953 年)10 月 昭和 29 (1954 年)4 月		別冊文芸春秋 (蚊の群 (蚊の夢)) 別冊文芸春秋 (蛇の卵) オール読物 (秋の魚 (鳩の音))
昭和 25 (1950 年)12 月~ 昭和 26 (1951 年)3 月	舞姫	朝日新聞
昭和 26 (1951 年)5 月	たまゆら	別冊文芸春秋
昭和 23 (1948 年)5 月、8 月、9 月、10 月、12 月 昭和 24 (1949 年)3 月	少年	人間
昭和 27 (1952 年)1 月	岩に菊	文芸
昭和 26 (1951 年)8 月 昭和 27 (1952 年)1 月 昭和 27 (1952 年)5 月 昭和 29 (1954 年)5 月	名人	新潮 (名人) 世界 (名人生涯) 世界 (名人供養) 世界 (名人余香)
昭和 27 (1952 年)1 月~11 月 昭和 28 (1953 年)1 月~5 月	日も月も	婦人公論
昭和 27 (1952 年)10 月	自然	文芸春秋
昭和 28 (1953 年)4 月	無言	中央公論
昭和 27 (1952 年)11 月	明月	文芸
昭和 28 (1953 年)11 月	水月	文芸春秋
昭和 29 (1954 年)1 月	小春日	文芸
昭和 29 (1954 年)4 月	横町	別冊文芸春秋
昭和 29 (1954 年)8 月	離合	知性
昭和 29 (1954 年)1 月~12 月	みづうみ	新潮
昭和 30 (1955 年)4 月	故郷	新潮
昭和 31 (1956 年)1 月 昭和 31 (1956 年)4 月	あの国この国	小説新潮 (あの国この国) 小説新潮 (隣の人)
昭和 33 (1958 年)1 月	弓浦市	新潮
昭和 33 (1958 年)1 月	並木	文芸春秋
昭和 35 (1960 年)1 月~6 月 昭和 36 (1961 年)1 月~11 月	眠れる美女	新潮
昭和 36 (1961 年)10 月~ 昭和 37 (1962 年)1 月~	古都	朝日新聞
昭和 38 (1963 年)8 月~ 昭和 39 (1964 年)1 月	片腕	新潮

昭和 39 (1964 年)6 月 昭和 40 (1965 年)2 月~ 昭和 41 (1966 年)2 月 (休載 2 回) 昭和 42 (1967 年)11 月~ 昭和 43 (1968 年)10 月(休載 2 回)	たんぽぽ	新潮
昭和 45 (1970 年)12 月	竹の聲桃の花	中央公論
昭和 45 (1970 年)4 月	髪は長く	新潮
昭和 47 (1972 年)10 月	友人の妻	新潮

付録 B

本論文で用いた各文体特徴量の次元数、cutoff値、othersを含むと含まない場合のCarmer's Vを次の表に示す。

表B1 代筆問題における文体特徴量の情報

小説名	文体特徴量	次元数	Cutoff値	Carmer's V (Others含む)	Carmer's V (Others含まない)
乙女の港	文字記号bigram	2117	30	0.16	0.2
	タグ付き形態素	218	20	0.07	0.07
	文節パターン	351	20	0.08	0.09
花日記	文字記号bigram	2111	30	0.16	0.2
	タグ付き形態素	218	20	0.07	0.07
	文節パターン	352	20	0.08	0.09
コスモスの友	文字記号bigram	1915	30	0.17	0.21
	タグ付き形態素	202	20	0.07	0.07
	文節パターン	324	20	0.09	0.09
古都	文字記号bigram	2096	60	0.12	0.15
	タグ付き形態素	318	20	0.06	0.06
	文節パターン	600	20	0.07	0.08
眠れる美女	文字記号bigram	2037	40	0.15	0.18
	タグ付き形態素	222	20	0.06	0.06
	文節パターン	424	20	0.08	0.08
山の音	文字記号bigram	2004	50	0.16	0.2
	タグ付き形態素	244	20	0.07	0.07
	文節パターン	485	20	0.08	0.09