# 夏目漱石の小説における文語表現について

土山　玄（一橋大学 森有礼高等教育国際流動化機構）

本研究では文学的文章において文体的特徴の継時的な変化が認められるのか，計量的な観点から検討を加える．本研究において採り上げたのは文豪として著名な夏目漱石の小説 22 作品である．夏目漱石の活動期間は 1905 年から 1916 年までであり，この期間において出現傾向が変化した文体的特徴を明らかにするために統計手法を用いて分析を行った．

分析の結果，「り」「き」「ごとし」「なり」「たり」などの文語表現と考えられる助動詞の出現率が 1905 年に発表された『幻影の盾』『一夜』『薤露行』において高く，以降の作品では出現傾向が低くなる．従って，計量的な判断に基づけば，夏目漱石の小説において文体の変化が認められたと考えられる．

# Quantitative Analysis of Written Japanese Language Styles in Soseki Natsume's Novels

Gen Tsuchiyama (Mori Arinori Institute for Higher Education and Global Mobility, Hitotsubashi University)

In this study, we investigate chronological change of writing style of Soseki Natsume. He is one of the masterful novelists in Modern Japan, and his litrary works are common subjects of the literary research. However, quantitative research for Natsume Soseki's works is not developed very well. In this study, we analyse the relative frequency of the words and appearance ratio of the parts of speech using statistical method. The results of the analysis indicate that "Mabaroshi no Tate," "Ichiya," and "Kairoko" which are published in 1905 have a high appearance ratio of words considered to be written Japanese language styles. Therefore, we conclude that they quantitatively differ from the other novels.

## １．はじめに

文学的な文章を対象とし，統計手法を用いた研究は広く行われている．一般に，このような研究分野は計量文献学と称される．古典文学などの歴史的な文献は，著者や成立時期について議論の対象となっている場合が少なからず存在する．このような問題を扱うとき，従来は記述内容の検討や成立に関する歴史的事実の考証という観点から研究をおこなうのが主たる方法であったが，計量文献学ではこのような方法とは一線を画し，文章から得られる計量的なデータを収集し，これを分析することによって当該文献の文体的特徴を把握し，結論を導き出すのである．このような文章の計量分析では，著者の識別や推定，あるいは文献の成立順序の推定などが研究されている．前者は著者間の文体的特徴の相違について検討を加え，後者は著者の習慣的，形式的な文体的特徴の変化について検討を加える．

本研究は執筆順序の推定に類する研究であり，日本における文学的文章においても継時的な文体の変化が計量的な手法を用いることで観察されるのか，検討を加える．このような文体の継時的な変化に注目し，執筆順序の推定を行った研究の嚆矢は古く 19 世紀まで遡る．L. Campbell は 1867 年にプラトンの 30 余りある『対話篇』の執筆年代を推定するために，語の出現頻度を用いて計量的な分析を行っている[1]. プラトンの執筆時期は 50 年から 60 年に及ぶと考えられているが，『法律』がプラトンの最後の著作であることをアリストテレスが言及していることを除き，『対話篇』の執筆順序は不明であった．プラトンの思想には発展がみられることから，プラトンの思想を体系的に理解するためには，著作の執筆順序を明らかにすることがきわめて重要であったとされる[2].

日本語で記述された文献を対象とした執筆順序の推定の研究として芥川龍之介の文章をについて分析を行った金 (2009) があげられるが[1]，著者識別を目的とした研究に比べると十分に展開しているとは言えない．金 (2009) では芥川龍之介の文章 309 編について分析を行っている．分

表 1　分析に用いた作品リスト

| タイトル | 掲載時期 | 地の文の延べ語数 | 会話文の延べ語数 |
|---|---|---|---|
| 吾輩は猫である（一〜六） | 1905年1月 | 51462 | 51199 |
| 幻影の盾 | 1905年4月 | 12910 | 2523 |
| 琴のそら音 | 1905年5月 | 7709 | 6303 |
| 一夜 | 1905年9月 | 2961 | 1215 |
| 薤露行 | 1905年11月 | 10156 | 3091 |
| 吾輩は猫である（七〜十一） | 1906年1月 | 59138 | 46671 |
| 趣味の遺伝 | 1906年1月 | 22550 | 2579 |
| 坊っちゃん | 1906年4月 | 47884 | 9802 |
| 草枕 | 1906年9月 | 43345 | 13158 |
| 二百十日 | 1906年10月 | 4478 | 15403 |
| 野分 | 1907年1月 | 32182 | 29582 |
| 虞美人草 | 1907年6月23日〜1907年10月29日 | 68249 | 65186 |
| 坑夫 | 1908年1月1日〜1908年4月6日 | 83307 | 9521 |
| 文鳥 | 1908年6月 | 7093 | 45 |
| 夢十夜 | 1908年7月 | 9988 | 934 |
| 三四郎 | 1908年9月1日〜1908年12月29日 | 79383 | 28417 |
| それから | 1909年5月31日〜1909年8月14日 | 95502 | 23855 |
| 門 | 1910年3月1日〜1910年6月12日 | 80178 | 14280 |
| 彼岸過迄 | 1912年1月1日〜1912年4月29日 | 108312 | 21307 |
| 行人 | 1912年12月6日〜1913年11月15日 | 92638 | 57169 |
| こころ | 1914年4月20日〜1914年8月11日 | 23744 | 9415 |
| 道草 | 1915年6月3日〜1915年9月14日 | 74206 | 23992 |
| 明暗 | 1916年5月26日〜1916年12月14日 | 140959 | 71432 |

析の結果, 係助詞の「は」及び格助詞の「に」「を」「の」の出現率が継時的に増加し，反対に格助詞の「が」「と」や接続助詞の「て」の出現率が減少していることを明らかにした[3].

本研究では日本近代における文豪として著名な夏目漱石の小説を採り上げ, 継時的に出現傾向が変化する文体的特徴が認められるのか検討するために統計的な分析を行う.

夏目漱石の小説家としての活動期間は『吾輩は猫である』を発表した 1905 年から『明暗』が未完のまま発表された 1916 年までであり，本研究ではこの期間に発表された 22 作品について分析を行った. 本研究ではこれらの作品の品詞の比率や単語の出現傾向に変化が認められるのか計量的な観点から検討を加える．また，夏目漱石の小説を対象とした関連研究として，文体の継時的な変化についての研究とした研究があり[4], 動詞と補助動詞が隣接共起する頻度が継時的に減少する傾向が認められることが報告されている. そこで，本研究では品詞の共起関係だけではなく，単語の出現傾向についても考察を行う.

## 2．データ

本研究で採り上げた小説は表 1 に示した 22 作品である．これら 22 作品の発表年も表 1 に示した.『吾輩は猫である』は第一から第六までは 1905 年に発表されており，第七から第十一は 1906 年に発表されている．加えて，表 1 に示したように『吾輩は猫である』の第六の発表と第七の発表との間に『幻影の盾』『琴のそら音』『一夜』『薤露行』という 4 作品が発表されている．そのため, 本研究においては『吾輩は猫である』は第一から第六までと第七から第十一までの 2 つに分割した.『行人』も発表年が複数年にまたがっているが,『行人』は朝日新聞に連載された小説であり，連載期間中に他の小説が発表されていないため,『吾輩は猫である』のように分割しない.

また, 本研究で用いた小説のテキストデータは青空文庫 (http://www.aozora.gr.jp/) から入手した. 形態素解析は MeCab ver. 0.996 を，形態素解析に使用した辞書は UniDic ver. 2.0.1 を用いた.

## 3．分析

### 3.1　特徴量について

本研究では, 作品別の品詞の比率と単語の出現率を特徴量として分析に用いた. 品詞の比率は各作品の延べ語数に対する各品詞の出現頻度の割合である．次いで，単語の出現率については上記

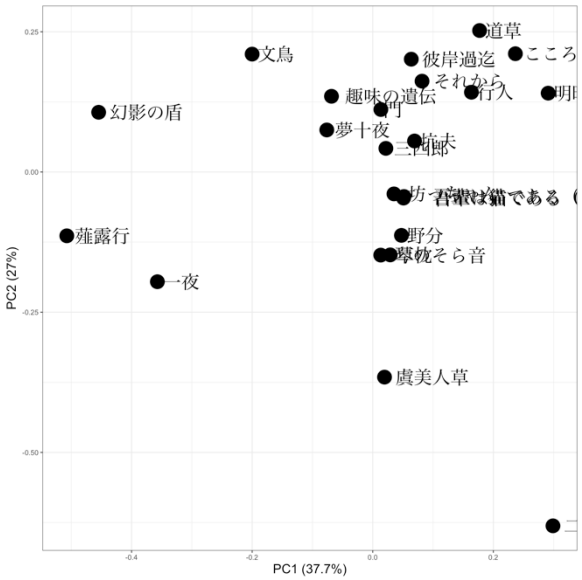


図 1　14 品詞の主成分分析の結果

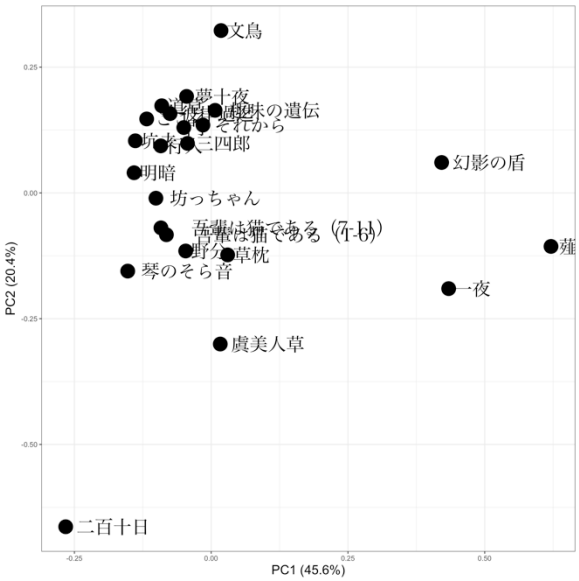


図 3　24 品詞タグの主成分分析

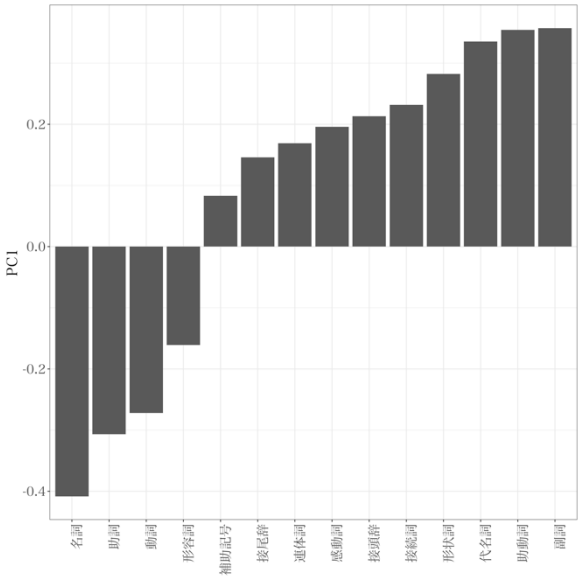


図 2　14 品詞についての主成分分析の第 1 主成分の主成分負荷量

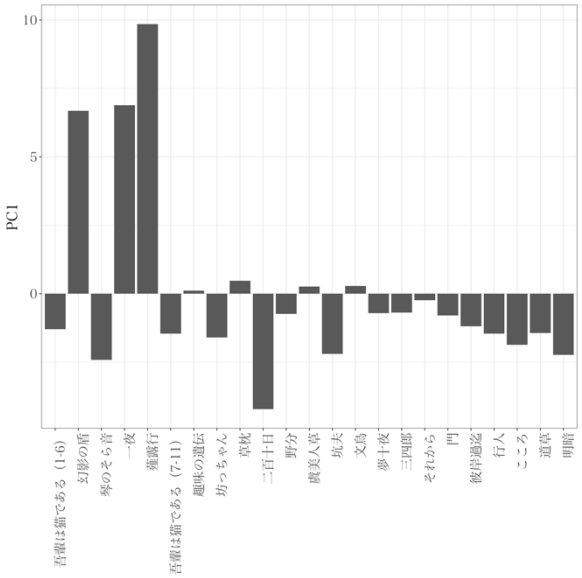


図 4　24 品詞タグを用いた主成分分析の第 1 主成分の主成分得点

の品詞タグ別に計算し，見出し語となる単語の出現頻度の総度数に対する割合を求めた．

### 3.2　品詞の比率

まず，本研究では形態素解析によって得られた 14 の品詞タグの作品別の比率について，相関係数行列を用いた主成分分析を行った．分析に用いた品詞タグは名詞，代名詞，動詞，形容詞，形状詞 (形容動詞の語根)，副詞，連体詞，接続詞，感動詞，接頭辞，接尾辞，助詞，助動詞，補助記号である．図 1 は主成分分析の結果であり，第 1 主成分の主成分得点と第 2 主成分の散布図である．図 1 より第 1 主成分の主成分得点に基づくと 1905 年に発表された『幻影の盾』『一夜』『薤露行』の 3 作品が他の小説から離れて付置されると判断される．図 2 は主成分分析によって求められた第 1 主成分の主成分負荷量であり，これより『幻影の盾』『一夜』『薤露行』の 3 作品は名詞や助詞の比率が他の作品より高く，反対に副詞や助動詞の比率が低いと考えられる．名詞や副詞と異なり，助詞や助動詞は文中において語彙的意味ではなく文法的機能を担う語である．従って，助詞や助

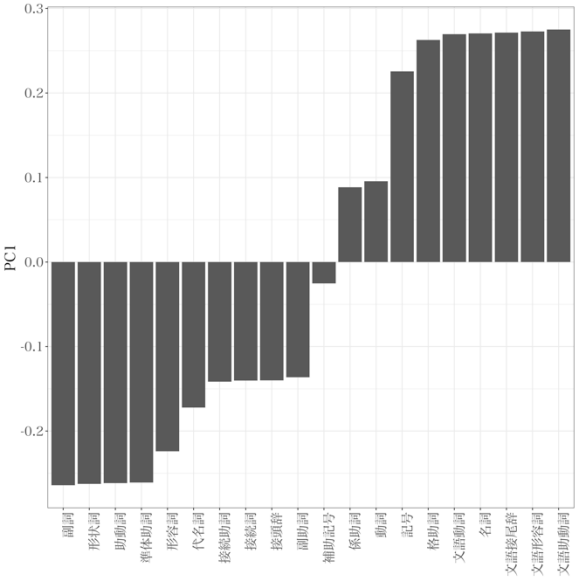


図 5　24 品詞タグを用いた主成分分析の第 1 主成分の主成分負荷量

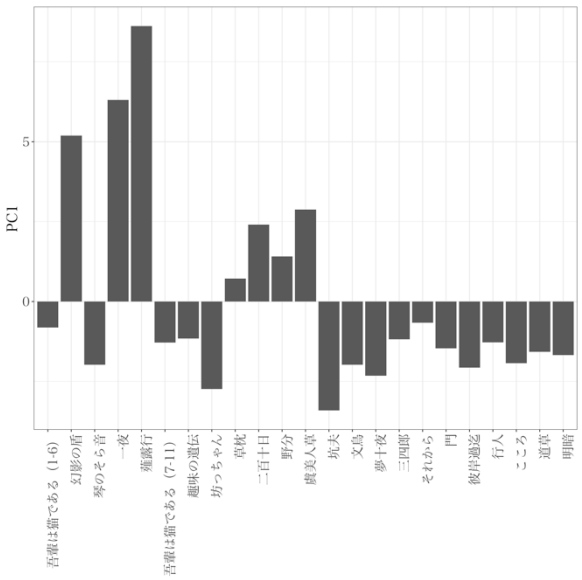


図 7　会話文を除外した 24 品詞タグを用いた主成分分析の第 1 主成分の主成分得点

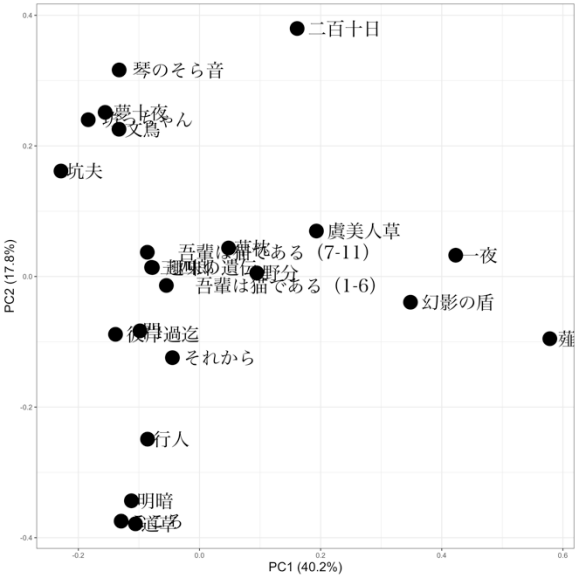


図 6　会話文を除外した 24 品詞タグの主成分分析

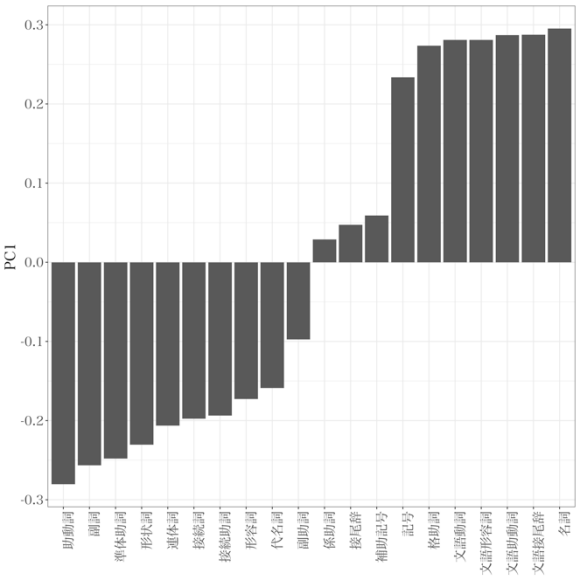


図 8　会話文を除外した 24 品詞タグを用いた主成分分析の第 1 主成分の主成分負荷量

動詞の出現傾向は小説の内容の相違の影響を受けにくいと考えられ，ここに小説のストーリーの相違に起因しない文体的特徴の相違が認められ得ると推測される．

次に，助詞を 1 つのタグとはせず，本研究では格助詞，副助詞，係助詞，接続助詞，終助詞，準体助詞というように 6 つのタグに分割した．これに加えて，形態素解析の際に語であるかの判定が可能であり，本研究ではこの情報を分析に使用し動詞，形容詞，助動詞，接尾辞の 4 つは形態素解析の際に文語と判定された語があったことから，文語動詞，文語形容詞，文語助動詞，文語接尾辞という 4 つのカテゴリを新たに作った．ただし，この文語という判定は正確ではなく，現代語用の活用に含まれない語形が文語として出力されたものである．従って，文語と口語が同形の単語は文語と判定されない．しかし，助動詞の「べし」や「ごとし」などは文語と判定されることから，

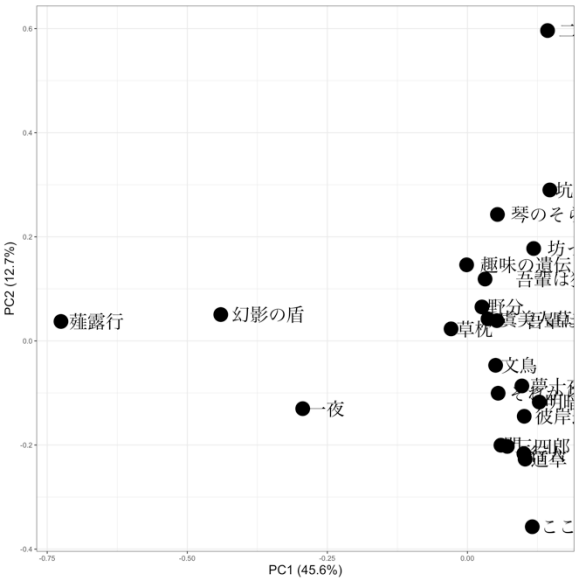


図 9　助動詞上位 20 語の出現率についての主成分分析の結果

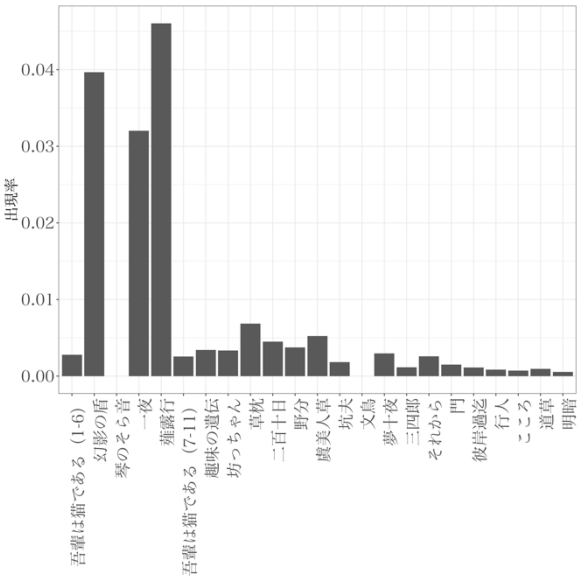


図 11　全文における「り」の出現率

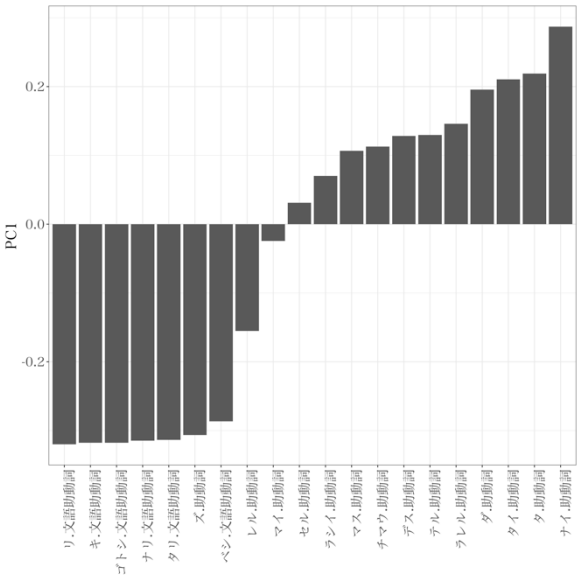


図 10　助動詞上位 20 語の主成分分析の第 1 主成分の主成分負荷量

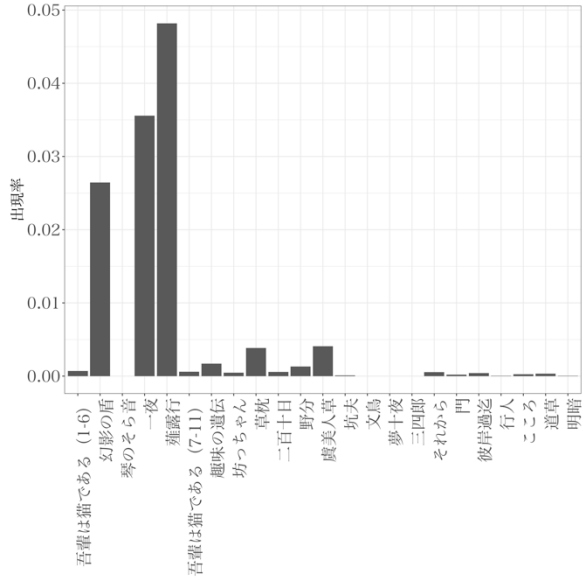


図 12　全文における「き」の出現率

これらの 4 つのカテゴリを用い主成分分析を行った．新たに分析に用いた品詞タグの総数は 24 である．

図 3 は 24 の品詞タグを用いた主成分分析によって求められた第 1 主成分と第 2 主成分の主成分得点の散布図である．図 3 より第 1 主成分の主成分得点に基づくと図 1 と同様に『幻影の盾』『一夜』『薤露行』の 3 作品が他の小説から離れて付置されると判断される．図 4 は第 1 主成分の主成分得点を棒グラフとして可視化したグラフであるが，『幻影の盾』『一夜』『薤露行』の 3 作品が他の作品とは異なる量的特徴を有していることは明らかである．また，図 5 は第 1 主成分の主成分負荷量である．なお，図 5 に示した変数は主成分負荷量の大きい 10 変数と小さい 10 変数である．図 4 及び図 5 より，1905 年発表された『幻影の盾』『一夜』『薤露行』の 3 作品は文語助動詞や文語形容詞などの比率が他の作品より高いと考

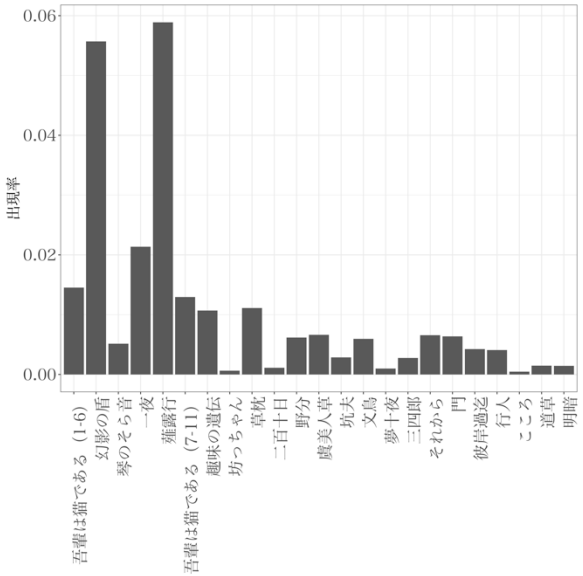


図 13　全文における「ごとし」の出現率

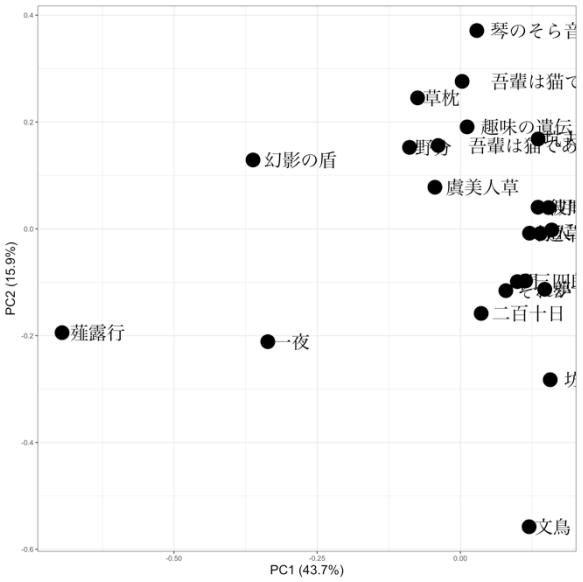


図 14　会話文を除外した助動詞上位 20 語の出現率についての主成分分析の結果

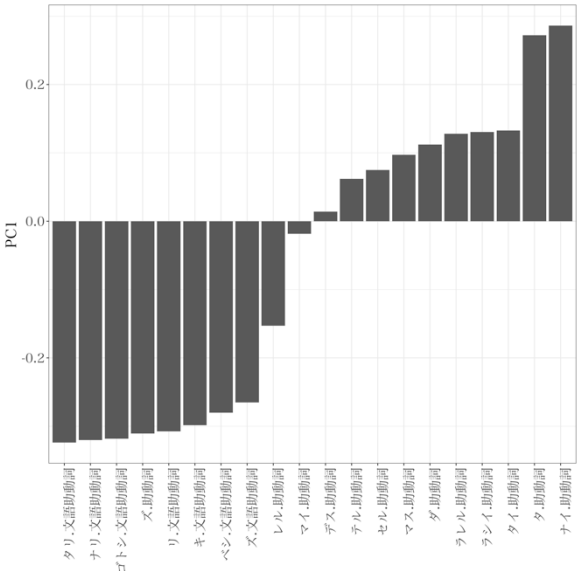


図 15　会話文を除外した助動詞上位 20 語の主成分分析の第 1 主成分の主成分負荷量

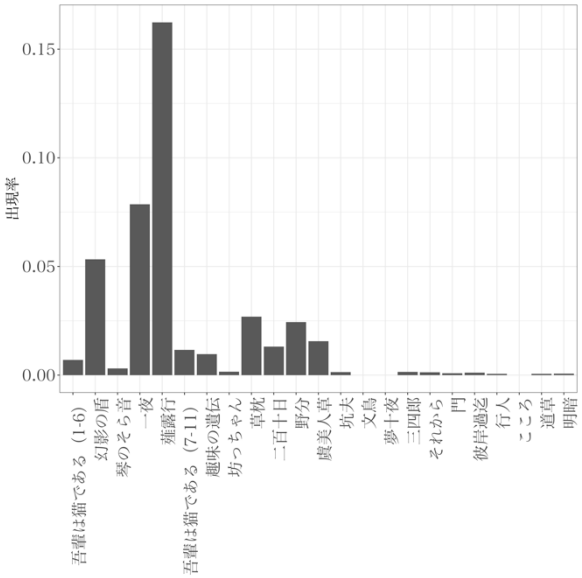


図 16　地の文における「たり」の出現率

えられる．なお，図 5 において，文語助動詞の他に助動詞という変数があるが，助動詞は文語助動詞と判定された単語を除いた品詞タグである．

以上の分析は分析対象とした 22 作品の全文を用いた．しかし，小説内部における会話文は描かれるストーリーや舞台に応じて，地の文とは文体が同様ではないことが考えれる．上掲の金 (2009) においても会話文を分析から除外している[3]ことから，本研究においても会話文を削除し，同様に 24 の品詞タグの比率について改めて分析を行った．図 6 は第 1 主成分及び第 2 主成分の主成分得点の散布図である．加えて，図 7 は第 1 主成分の主成分得点を棒グラフとして表したものである．図 7 において，1906 年に発表された『草枕』『二百十日』，1907 年に発表された『野分』『虞美人草』の 4 作品は第 1 主成分の主成分得点が正となるが，ここにおいても第 1 主成分において『幻影の盾』『一夜』『薤露行』の 3 作品は他の作品と

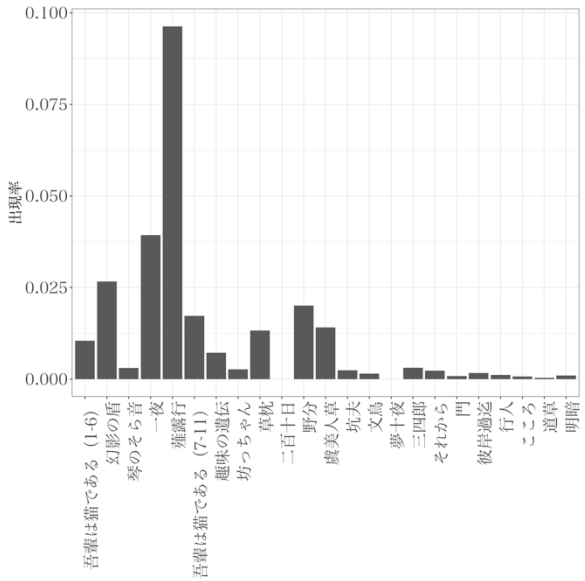


図 17　地の文における「なり」の出現率

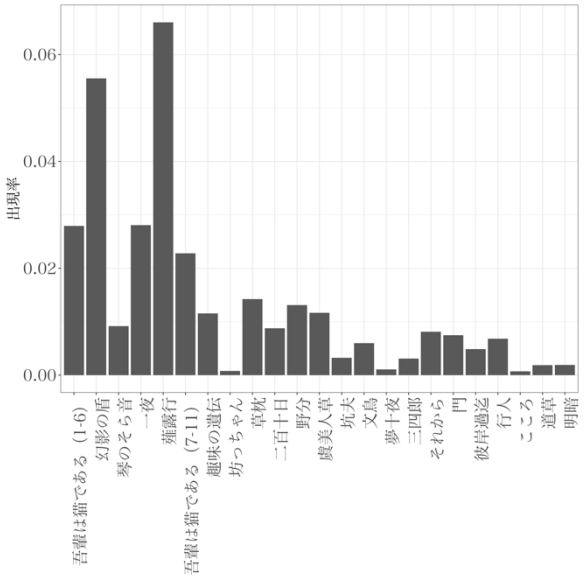


図 18　地の文における「ごとし」の出現率

異なる文体的特徴を有していると考えられる．また，図 8 は第 1 主成分の主成分負荷量の棒グラフであり，図 5 と同様に主成分負荷量の大きい 10 変数と小さい 10 変数である．図 8 より『幻影の盾』『一夜』『薤露行』の 3 作品は他の作品に比べ文語助動詞の比率が高く，文語助動詞を除く助動詞の比率が低いと考えられる．

このように，22 作品の全文を対象とした場合，22 作品の会話文除外した地の文のみを対象とした場合，どちらの場合においても文語表現と考えられる助動詞の比率が『幻影の盾』『一夜』『薤露行』の 3 作品において相対的に高いと考えられる．

### 3.3　単語の出現率

品詞の比率についての分析から，1905 年に発表された『幻影の盾』『一夜』『薤露行』の 3 作品は他の作品と異なる文体的特徴を有していると考えられる．そして，以上の分析結果からこれら 3 作品は機能語である文語助動詞と判定された助動詞を含む助動詞の出現傾向が他の小説と相違していると推測される．そこで，本研究ではすべての助動詞の出現傾向について分析を加えた．また，ここでも 22 作品の全文を対象とした分析と会話文を除外した分析を行った．

まず，22 作品の全文を対象とし主成分分析を行った．分析に用いたのは出現頻度上位 20 語である．上位 20 語までの累積度数は全体の 99.2%を占める．図 9 は第 1 主成分と第 2 主成分の主成分得点の散布図である．図 9 より第 1 主成分の主成分得点に基づき判断すると，『幻影の盾』『一夜』『薤露行』の 3 作品は他の小説とは助動詞の出現傾向が相違していると考えられる．また，図 10 に示した第 1 主成分の主成分負荷量から，上掲の 3 作品は「り」「き」「ごとし」「なり」「たり」などの出現傾向が高いと考えられる．

図 11 は助動詞「り」の出現率，図 12 は助動詞「き」，図 13 は助動詞「ごとし」の出現率を可視化したグラフである．図 13 に示したように，『一夜』における「ごとし」の出現率は『幻影の盾』及び『薤露行』に比べて低いが，それでもこれら 3 作品は「り」「き」「ごとし」と言った助動詞の出現率が他の小説における出現率よりも高いことが分かる．

次に，22 作品の会話文を除外した地の文について主成分分析を行った．全文を対象とした分析と同様に出現頻度上位 20 語を分析に用いた．上位 20 語までの累積度数は全体の 99.4%に該当する．図 14 は第 1 主成分と第 2 主成分の主成分得点の散布図であり，『幻影の盾』『一夜』『薤露行』の 3 作品の第 1 主成分の主成分得点は負となり，ここでも他の作品とは異なる助動詞の出現傾向を有していると考えられる．図 15 は第 1 主成分の主成分負荷量であり，『幻影の盾』『一夜』『薤露行』の 3 作品は他の小説と比較し，「たり」「なり」「ごとし」「ず」「り」「き」などの出現傾向が高いと考えられる．

図 16 は助動詞「たり」の出現率，図 17 は助動詞「なり」の出現率，図 18 は助動詞「ごとし」の出現率を可視化したグラフである．図 16 に示したようにこれら 3 作品の「たり」の出現率は他の作品の出現率を大きく上回っており，図 17 より『薤露行』における「なり」の出現率は非常に高いと言える．また，図 18 は図 13 と同様に『一夜』における「ごとし」の出現率は『幻影の盾』

や『薤露行』に比べると低く,『吾輩は猫である』における出現率と大きく変わらない.

先にふれたように, 形態素解析において文語と判定された助動詞は現代語用の活用に含まれない語形が文語として出力されたものである. しかし, 品詞の比率ではなく助動詞の単語の出現率を分析に用いたことで,『幻影の盾』『一夜』『薤露行』の 3 作品において「り」「き」「ごとし」「なり」「たり」などの助動詞の出現率が, 全文を対象とした場合, 地の文のみを対象とした場合, どちらにおいても他の作品に比べて出現率が高いことが明らかになった.

また, 芥川龍之介の著作を対象とした研究では助詞の出現率に継時的な文体変化が認められたと報告されている[3]が, 本研究の分析では, 助動詞の出現傾向において文体の変化が認められた.

## 4. 考察

本研究では, 近代日本の文豪として著名な夏目漱石の小説 22 作品を採り上げ, そのテキストデータに対して統計手法を用いて分析を行った. 品詞の比率及び単語の出現率について相関係数行列を用いた主成分分析を行った結果, 1905 年に発表された『幻影の盾』『一夜』『薤露行』の 3 作品と他の小説との間に文体的特徴の相違が認められた. これら 3 作品は助動詞の出現率の分析から,「り」「き」「ごとし」「なり」「たり」などの文語表現と考えられる助動詞の出現傾向が高いと考えられる.

従って, 夏目漱石は作家として活動を始めた 1905 年に発表した小説では文語表現を用いていたが, その後は文語表現の使用が少なくなっていることが明らかになった. よって, 計量的な判断に基づけば, 夏目漱石の文体的特徴に量的な変化が本研究の分析を通じて認められたと考えられる.

## 参考文献

[1] Campbell, Lewis. The SOPHISTES and POLITICUS of Plato. Vol. 1. Clarendon Press, 1867.
[2] 土山玄. 文学作品の計量分析: その方法と歴史. 研究報告人文科学とコンピュータ, 2015, Vol. 2015, No. 7, pp. 1-6.
[3] 金明哲. 文章の執筆時期の推定―芥川龍之介の作品を例として―. 行動計量学, 2009, Vol. 36, No. 2, pp. 89-103.
[4] 土山玄. 夏目漱石の文体の計量的な変化について. 日本行動計量学会第 45 回大会抄録集, 2017, pp. 180-181.