森鷗外の小説を対象とした文体の継時的な変化についての計量的な検討

Quantitative Analysis Of Chronological Changes Of The Writing Style In Ogai Mori's Novels

土山 玄
Gen Tsuchiyama
お茶の水女子大学 文理融合 AI・データサイエンスセンター, 東京都文京区 2-1-1
Ochanomizu University, 2-1-1, Ohtsuka, Bunkyo-ku, Tokyo

概要: 本研究では森鷗外の小説 47 作品を分析対象として、継時的に出現傾向が変化する文体的特徴の抽出を試みた。分析において、用いた特徴量は品詞の比率と助詞、および助動詞の出現率である。これらの特徴量に対して上掲の 47 作品の出版年を目的変数としてランダムフォレストを行い、変数重要度を求めることで継時的に出現傾向の変化する文体的特徴の抽出を行った。分析の結果、1890 年に出版された『うたかたの記』及び『舞姫』、1891 年に出版された『文づかひ』の 3 作品は他の小説と異なる傾向を有していることが明らかになった。これに加えて、形容詞の比率や助動詞の「ない」の出現率などにおいて 1912 年以降の作品ではそれ以前の作品と異なる出現傾向を有する可能性が認められた。

Abstract: In this study, we investigate chronological change of writing style of Ogai Mori. He is one of the masterful novelists in Modern Japan, and his litrary works are common subjects of the literary research. In this study, we analyse the relative frequency of the words and appearance ratio of the parts of speech using random forests. The results of the analysis indicate that "Utaka no Ki" and "Maihime" which are published in 1890 and "Fumi dukahi" which is published in 1891 are different in writing style form other works. In addition, we reveal that the tendency of word occurrence of auxiliary verb is different before and after 1912.

キーワード: 計量文献学、テキストマイニング、機械学習、ランダムフォレスト
Keywords: Stylometrics, Text mining, Machine learning, Random forests

1. はじめに

文学作品を対象とし、計量的な手法を用いて文章を分析する研究は計量文献学と称される。計量文献学は、著者の文体に関わる習慣的、形式的特徴を統計的に分析することで著者の識別や推定、文献の成立年代、あるいは成立の順序を推定する学問分野である。このような計量文献学では文体的特徴の出現傾向を調査することで著者の識別や推定を行うことが多く、また数多くの研究成果が報告されている。計量文献学では文体は著者の個性を映しており、文体的特徴は著者間において出現傾向が相違するという考えに基づいている。文体的特徴とはすなわち文章にあらわれる著者の形式的、あるいは習慣的な表現形式のことである。

また、1 人の作家が多数の作品を残した場合、このような文体的特徴には継時的に出現傾向が変化するものもあることが推測される。つまり、特定の著者の文体的特徴の継時的な変化に注目することで、文体の成長や発展について考察するための透明性の高い資料を提出できると考えられる。

そこで、本研究では日本における文豪として知られる森鷗外の小説を分析対象とし、計量的に分析を行う

ことで継時的に出現傾向が変化すると考えられる文体的特徴を指摘する。本研究では森鷗外の小説について分析を行うにあたって、文体を規定する要素、すなわち文体的特徴であると考えられる、各作品における品詞の比率と単語の出現率を特徴量として多変量解析を行った。特に単語の出現率では助詞及び助動詞を採り上げた。この 2 品詞を採り上げることについて、助詞及び助動詞は名詞や動詞などと異なり、文中において語彙的意味を担うのではなく文法的機能を担うからである。語彙的意味を担う単語の出現率は小説において描かれるストーリーによって影響されるものと考えられるが、助詞や助動詞などの文法的機能を担う単語の出現率はストーリーによる影響は強くないと考えられる。なお、助詞や助動詞のような語彙的機能を担う単語は機能語と称される。

2. 関連研究

日本の文学的文章を対象とし、計量的な手法を用いて著作の執筆順序の推定を目的とした研究では金 (2009) が著名である。金 (2009) では芥川龍之介の著作について分析を行っている。芥川龍之介の文章 309 編を分析対象として採り上げ、統計手法を用いた分析を行った結果、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。

次いで、土山 (2019a) では森鷗外と並び文豪と称される夏目漱石の小説 22 作品を採り上げ、金 (2009) と同様に統計手法を用いて継時的に出現傾向は変化すると考えられる文体的特徴について検討を加えている。土山 (2019a) では夏目漱石の『自然を寫す文章』において「今日では一番言文一致が行はれて居るけれども、句の終りに「である」「のだ」とかいふ言葉があるので言文一致で通つて居るけれども、「である」「のだ」を引き抜いたら立派な雅文になるのが澤山ある。」という指摘があることから、文末表現を採り上げ主成分分析を行っている。その結果、文末表現については 1908 年頃に量的な特徴の変化が認められ、1909 年に発表された『それから』以降の作品は文末に助動詞を用いることが増加し、特に文末に助動詞の「た」の使用の増加が顕著であることを指摘している。

また、本研究と同様に森鷗外の小説を対象とした研究も報告されている。森鷗外は 1890 年から 1917 年まで作家として活動しているが、表 1 に示すように『うたかたの記』及び『舞姫』は 1890 年に、『文づかひ』は 1891 年に、『そめちがへ』は 1897 年に発表されている。その後の作品は 1909 年に発表されていることを考えると森鷗外の初期 4 作品が発表されてから 10 年以上の間隔を空けてから他の作品が発表されている。土山 (2019b) では森鷗外の小説 47 作品を対象とし、単語の出現率を特徴量とし計量的な分析を行っている。分析の結果、上掲の『うたかたの記』『舞姫』『文づかひ』の 3 作品は初期 4 作品を除く 43 作品に比べて文語表現、特に文語的な助動詞の出現率が顕著に高く、加えて上掲の 3 作品ほど顕著ではないが『そめちがへ』も 43 作品に比べると文語助動詞の出現率が高いことを指摘している。

表 1　森鷗外の小説と発表年

タイトル	発表年	タイトル	発表年
うたかたの記	1890	かのように	1912
舞姫	1890	興津弥五右衛門の遺書	1912
文づかい	1891	鼠坂	1912
そめちがへ	1897	佐橋甚五郎	1913
キタ・セクスアリス	1909	護持院原の敵討	1913
半日	1909	阿部一族	1913
鶏	1909	堺事件	1914
あそび	1910	大塩平八郎	1914
普請中	1910	安井夫人	1914
木精	1910	栗山大膳	1914
杯	1910	じいさんばあさん	1915
沈黙の塔	1910	二人の友	1915
牛鍋	1910	余興	1915
独身	1910	山椒大夫	1915
花子	1910	最後の一句	1915
里芋の芽と不動の目	1910	津下四郎左衛門	1915
青年	1910	魚玄機	1915
食堂	1910	伊沢蘭軒	1916
カズイスチカ	1911	壽阿彌の手紙	1916
妄想	1911	寒山拾得	1916
心中	1911	椙原品	1916
百物語	1911	渋江抽斎	1916
雁	1911	高瀬舟	1916
		細木香以	1917

3. データ

本研究に用いた森鴎外の小説は表 1 に示した 1890 年から 1917 年までに発表された 47 作品である。なお、これら 47 作品は上掲の土山 (2019b) と同じである。

また、これらの小説のテキストデータは web サイトの青空文庫 (http://www.aozora.gr.jp/) から入手した。

次に、それらのテキストデータに対し、形態素解析によって単語に品詞のタグ付けを行った。形態素解析は MeCab ver. 0.996 を、形態素解析の際に用いる辞書は UniDic ver. 2.0.1 を用いた。

このような処理によって作成されたテキストデータを対象に統計的な分析を行った。分析に際して、先にふれたように品詞の比率と助詞及び助動詞の出現率を特徴量として用いた。品詞の比率は作品別に各品詞の頻度を集計し、各作品の延べ語数に対する割合を求めた。次に、単語の出現率は品詞の比率と同様に作品別に各単語の頻度を集計し、各作品における品詞別の総度数に対する割合を求めた。

4. 分析

4.1 分析手法

本研究では分析において、主にランダムフォレストを用いた。ランダムフォレストを用いることで森鷗外の小説 47 作品において出現傾向が継時的に変化する文体的特徴を抽出した。ランダムフォレストとは機械学習の手法の 1 つであり、決定木あるいは回帰木のアンサンブル学習とも言える分析手法である。ランダムフォレストでは、まず分析対象の個体数の 2/3 にあたるブートストラップサンプルを抽出し、そのブートストラップサンプルを対象とし未剪定の決定木あるいは回帰木を生成する。また、未剪定の木を生成する際に、すべての変数を用いず、一般的に変数の数の平方根にあたる数の変数を用いる。ランダムフォレストはこのような未剪定の決定木あるいは回帰木を大量に生成し、分析結果を統合することで最終的な結果を得る。従って、ランダムフォレストを繰り返すと、同一の結果が得られることはおよそない。本研究では表 1 に示した出版年を目的変数としてランダムフォレストを行った。

また、ランダムフォレストでは分析を行う上で変数の重要度を推定する。本研究では森鷗外の小説の出版年を目的変数としているため、出版年の推定における変数重要度が求められる。従って、この変数重要度が高い変数が森鷗外の 47 作品の小説において出現傾向が継時的に変化している文体的特徴であると考えられる。

表 2　47 作品を対象とした品詞の変数重要度

品詞	重要度
感動詞	656.203
連体詞	223.859
助動詞	211.591
記号	147.152
補助記号	97.573
形容詞	71.985
接続詞	65.278
形状詞	62.134
助詞	44.009
代名詞	42.878
名詞	32.190
副詞	30.031
接尾辞	29.360
動詞	27.646
接頭辞	11.721

4.2 分析結果

本研究ではまず品詞の比率について分析を行った。分析では先にふれた Mecab 及び Unidic を用いた形態素解析においてタグ付けされた品詞のタグを用いた。具体的には名詞、代名詞、動詞、形容詞、形状詞、副詞、連体詞、接続詞、感動詞、助詞、助動詞、接頭辞、接尾辞、補助記号、記号の 15 のタグである。なお、形状詞は形容動詞の名詞語根に相当する。補

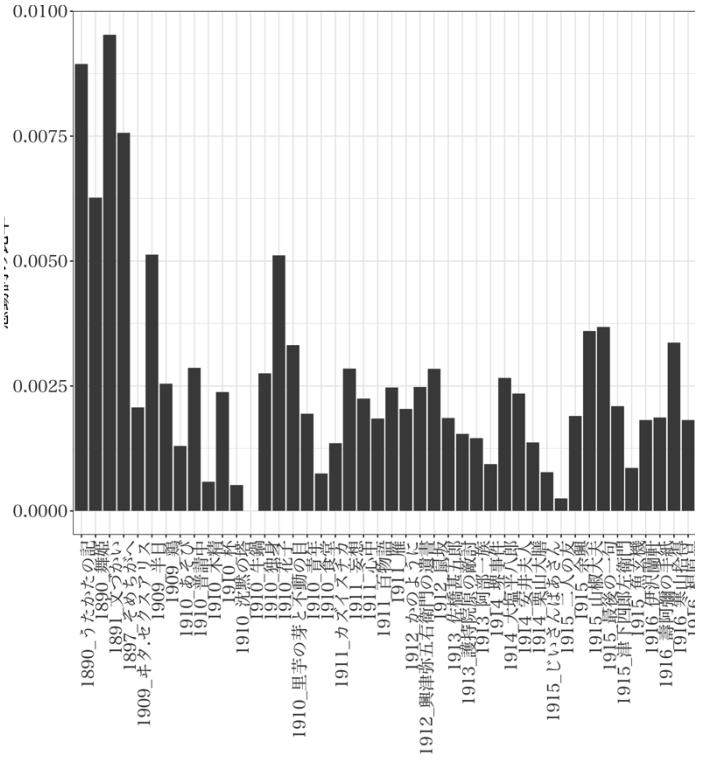

図 1　各作品における感動詞の比率

助記号は句読点やかぎ括弧などが含まれ、記号は文法的な機能を担わない記号が該当する。

これらの 15 のタグを説明変数とし、47 作品の出版年を目的変数としランダムフォレストを行った。その結果、表 2 に示すように推定された変数重要度は感動詞、連体詞、助動詞が高い。よって、森鷗外の小説 47 作品において、これら 3 品詞の出現傾向が継時的に変化している可能性が考えられる。そこで、変数重要度が最も大きかった感動詞の各作品における比率を可視化すると、図 1 に示すように初期の作品において感動詞の比率が高く、それ以降の作品に継時的な変化は認められないと考えられる。

次いで、15 の品詞タグを用いて、相関行列を用いた主成分分析を行った。図 2 は主成分分析によって求められた主成分得点の散布図である。横軸が第 1 主成分を、縦軸が第 2 主成分を意味している。なお、第 1 主成分の寄与率は 41.8%、第 2 主成分の寄与率は 13.0%であり、第 2 主成分までの累積寄与率は 54.9%である。図 2 において、初期 3 作品である『うたかたの記』『舞姫』『文づかひ』の第 2 主成分の主成分得点が小さく、これら 3 作品は類似した傾向を有していると考えられる。第 2 主成分の主成分負荷量は表 3 に示す通りであり、初期 3 作品は感動詞及び連体詞の比率が高く、助動詞の比率が小さい作品群であると解釈される。これは表 2 に示したランダムフォレストの結果と合致する。

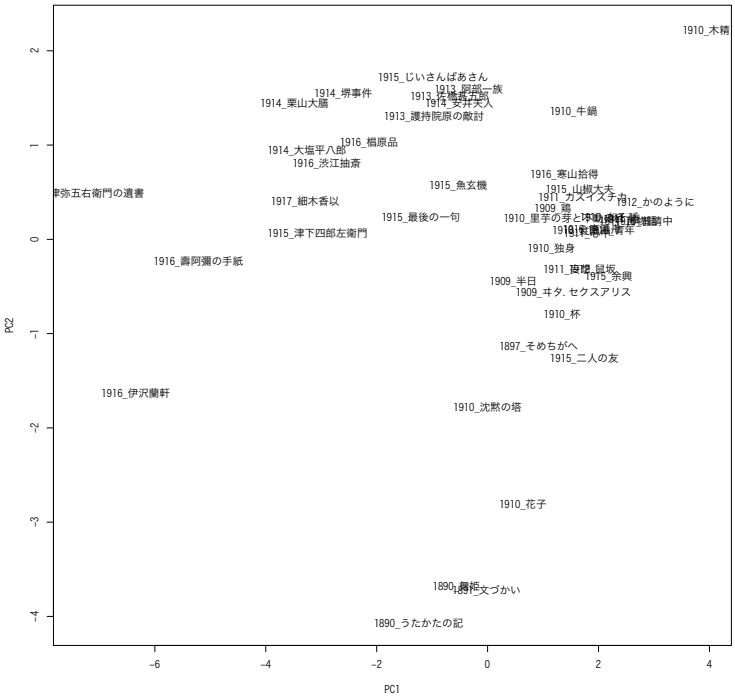

図 2　品詞の比率の主成分分析の結果

よって、初期 3 作品を除き 44 作品を対象として改めてランダムフォレストを行った。表4はランダムフォレストの結果として得られた変数重要度であり、形容詞の重要度が最大となった。これら 44 作品の形容詞の比率は図 3 に示す通りである。形容詞の比率は単調な変化を示していないが、1912 年以降の作品では顕著に形容詞の比率が認められる。従って、ここに森鷗外の小説における 1 つの文体的特徴の継時的な変化が明らかになったと言える。

次に、小説 47 作品を対象とした単語の出現率を特徴量としてランダムフォレストを行った。先に述べたように本研究では助詞と助動詞を採り上げ、分析を行った。まず助詞の出現率に対してランダムフォレストを行った。分析によって求められた変数重要度は表 5 に示す通りである。「のみ」の変数重要度が最大となり、「し」及び「ど」などの変数重要度も高い。そこで、「のみ」の各作品における出現率を可視化すると、図 4 に示すように初期 4 作品における出現率が顕著に高く、その他の作品ではおよそ出現しない。そこで、品詞に対する分析と同様に、初期 3 作品を分析対象から除外し、改めてランダムフォレストを行った。その結果、「に」「か」「も」と言った助詞の変数重要度が高く推定された。図 5 は各作品における「も」の出現率であり、1912 年より出現率が減少傾向にあると考えられる。

表 3　主成分負荷量

	PC1	PC2
名詞	-0.384	0.078
助詞	0.356	0.072
動詞	0.349	0.096
補助記号	0.067	-0.298
助動詞	0.239	0.228
接尾辞	-0.356	-0.033
副詞	0.310	-0.027
代名詞	0.164	-0.209
接頭辞	-0.235	-0.020
形容詞	0.327	-0.042
連体詞	0.072	-0.477
形状詞	0.346	0.015
接続詞	-0.062	0.106
記号	-0.032	-0.545
感動詞	0.012	-0.506

表 4　44 作品を対象とした品詞の変数重要度

品詞	重要度
形容詞	106.736
感動詞	93.422
補助記号	92.383
形状詞	31.325
助詞	27.502
接続詞	25.297
代名詞	20.821
連体詞	19.389
記号	14.778
接尾辞	12.550
名詞	8.121
助動詞	8.056
副詞	5.664
動詞	4.352
接頭辞	4.226

表 5　47 作品を対象とした助詞の変数重要度

	重要度
のみ.助詞	150.178
し.助詞	144.760
ど.助詞	140.729
ば.助詞	136.973
など.助詞	136.318
こそ.助詞	121.418
とて.助詞	121.092
しき.助詞	117.250
にて.助詞	72.315
より.助詞	70.872
に.助詞	66.114
きに.助詞	57.908
なり.助詞	55.214
から.助詞	48.221
で.助詞	38.027

表 6　47 作品を対象とした助動詞の変数重要度

	重要度
ず.助動詞	333.909
たる.助動詞	328.051
た.助動詞	255.677
なり.断定.助動詞	194.917
たり.断定.助動詞	165.736
ない.助動詞	105.531
だ.助動詞	71.238
べし.助動詞	52.153
ごとし.助動詞	50.207
や.助動詞	49.674
り.助動詞	32.982
てる.助動詞	20.603
しめる.助動詞	14.139
たい.助動詞	13.628
まじ.助動詞	11.138

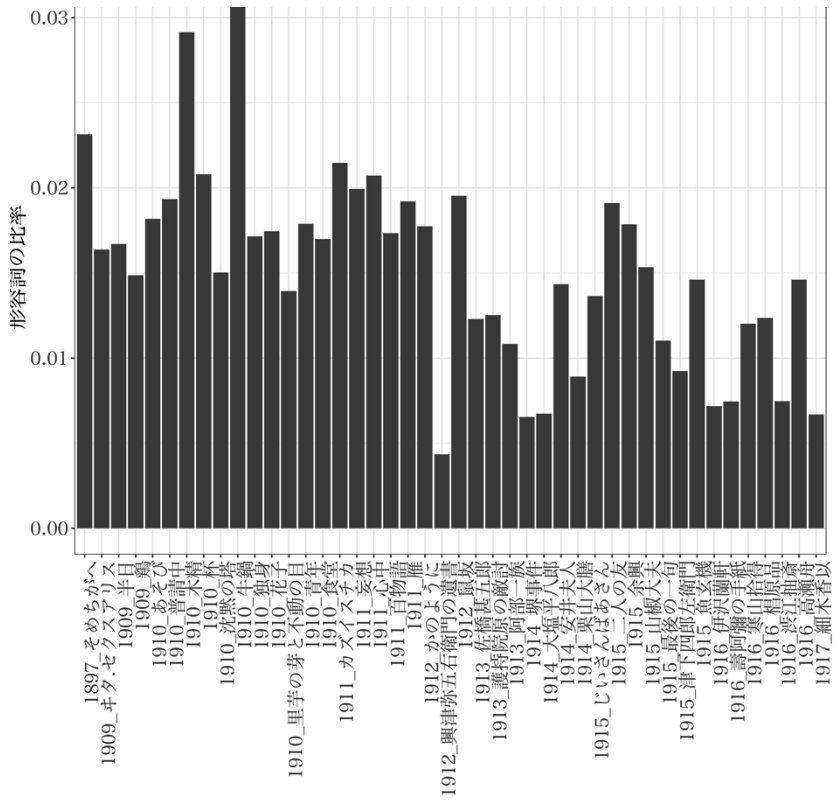

図 3　各作品における形容詞の比率

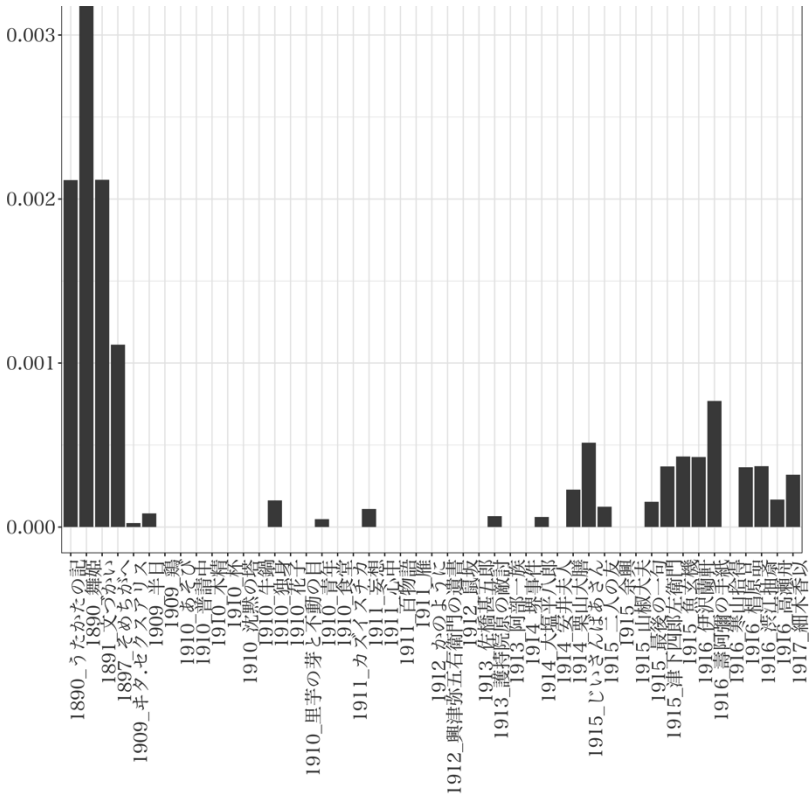

図 4　各作品における助詞「のみ」の出現率

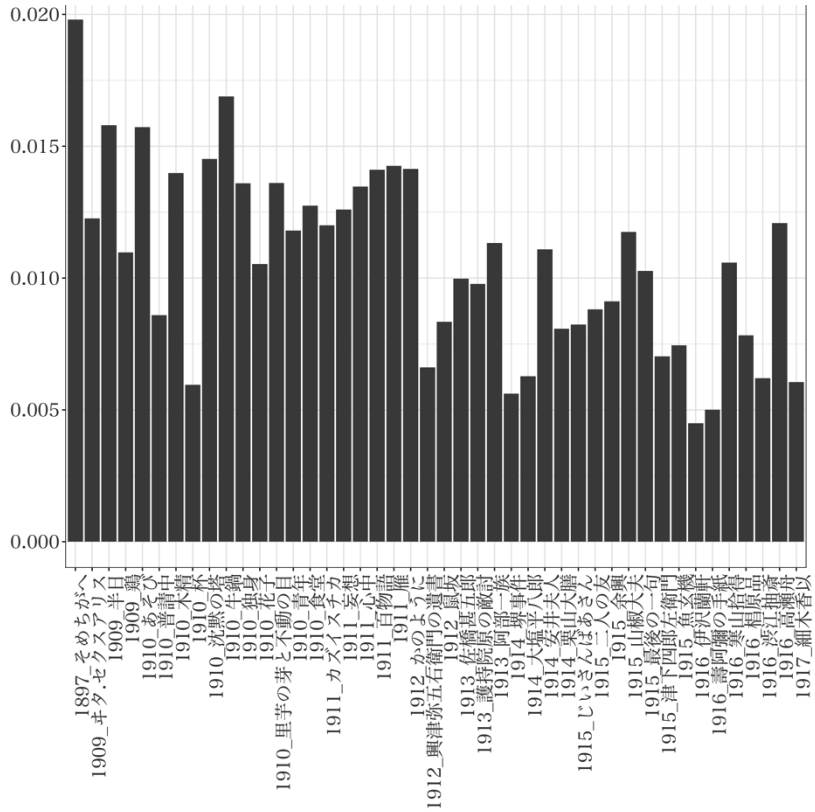

図 5　各作品における助詞「も」の出現率

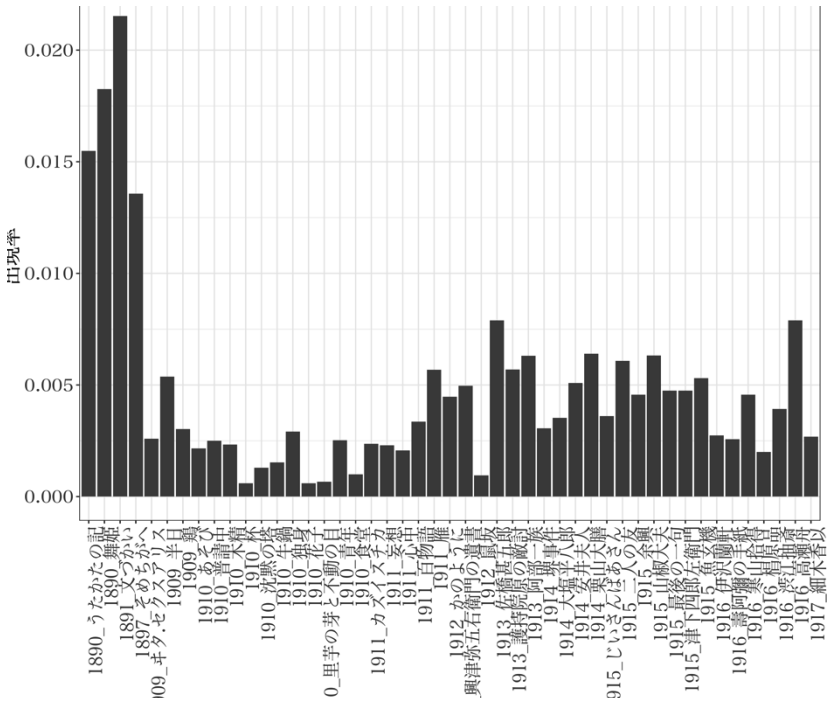

図 6　各作品における助動詞「ず」の出現率

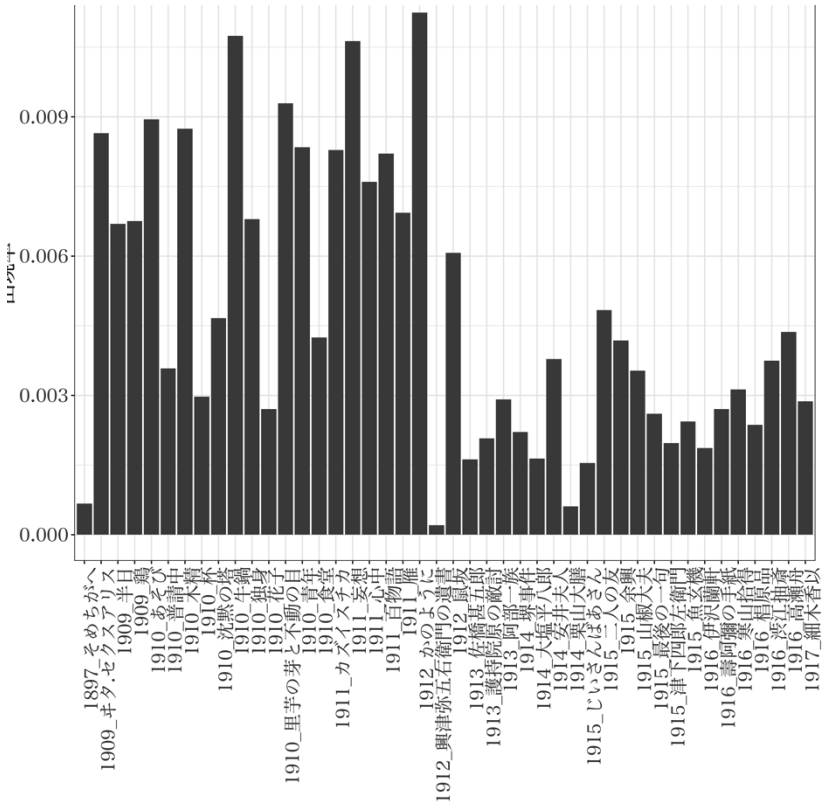

図 7　各作品における助動詞「ない」の出現率

最後に、助動詞の出現率についてランダムフォレストを行った。分析の結果として求められた変数重要度は表 6 に示す通りである。表 6 より「ず」「たる」と言った助動詞の重要度が高く推定された。しかし、図 6 において示した「ず」の出現率のグラフのように、これまでの分析と同様に初期の作品における出現率が偏って高いため、このような結果となったと考えられる。そこで、初期 3 作品を分析対象から除外し、改めてランダムフォレストを行った。その結果、表 7 に示すように「ず」「たい」「ない」と言った単語の重要度が高く推定された。図 7 は「ない」の各作品における出現率を可視化したグラフである。図 7 においても 1912 年以降の作品における出現率の低下が認められると考えられる。

5. 考察

本研究では、森鷗外の小説 47 作品を対象に、機械学習の手法の 1 つであるランダムフォレストを用い、出現傾向が継時的に変化する文体的特徴の抽出を試みた。その結果、品詞の比率の分析、助詞及び助動詞の出現率に対する分析において、出現傾向が変化する文体的特徴が明らかになった。従って、継時的に出現傾向が変化する文体的特徴の抽出を目的とするとき、ランダムフォレストは有効な分析手法の 1 つであると考えられる。

また、本研究における品詞に対する分析及び単語の出現率に対する分析によって、1890 年に出版された『うたかたの記』及び『舞姫』、1891 年に出版された『文づかひ』の 3 作品は他の小説と異なる傾向を有していることが明らかになった。これに加えて、形容詞の比率や助動詞の「ない」の出現率などにおいて 1912 年以降の作品ではそれ以前の作品と異なる出現傾向を有する可能性が認められた。

参考文献

[1] 漱石全集第 34 巻. 岩波書店, 1957.

[2] 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.

[3] 土山玄. 夏目漱石の小説における文語表現について. じんもんこん 2018 論文集, 2018, Vol. 2018, pp. 269-276.

[4] 土山玄. 文末表現の計量分析に基づく夏目漱石の小説の分類. 研究報告人文科学とコンピュータ, 2019a, 2019-CH-120, Vol. 6, pp. 1-4.

[5] 土山玄. 森鷗外の文体的特徴の変化に関する計量的な考察. 人文・自然研究, 2019b, Vol. 13, pp. 107-115.

[6] 工藤彰; 村井源; 徃住彰文. 計量分析による村上春樹長篇の関係性と歴史的変遷. 情報知識学会誌, 2011, Vol. 21, No. 1, pp. 18-36.