Vol.2019-CH-120 No.6
2019/5/11

文末表現の計量分析に基づく夏目漱石の小説の分類

土山 玄[†1]

概要：本研究では夏目漱石の小説を採り上げ、文末における文体的特徴の出現傾向の変化について検討を加えた。句点などの文末記号とその直前の 1 語及び文末記号とその直前の 2 語の出現率を特徴量として主成分分析を行った。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ると考えられる。特に 1909 年に発表された『それから』以降の 7 作品はどちらの分析においても同様の傾向を有している。主成分負荷量に対する考察から 1908 年あるいは 1909 年に認められる文末における文体的特徴の変化として、文末に助動詞を用いるようになったことがあげられる。特に、助動詞の「た」の使用が増加していると考えられる。

キーワード：計量文献学　夏目漱石　多変量解析

Quantitative Analysis of Sentence-final Expressions in Soseki Natsume's Novels

Gen TSUCHIYAMA[†1]

Keywords: Stylometry, Sosuke Natsume, multivariate analysis

1. はじめに

文章の計量分析では、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、数多くの研究成果が報告されている。これは、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顕著に相違するという考えに基づいている。その一方で、1 人の著者の文章において、文体的特徴の出現傾向が一定ではないことは想像するに難しくはない。このような研究に採り上げられる文体的特徴とは著者の文章にあらわれる習慣的な表現形式のことである。一般に、助詞や助動詞などの文体的機能を担う単語や句読点などの記号が分析に用いられる。そのような文体的特徴の中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。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。

本研究では夏目漱石の小説を採り上げ、文体的特徴の出現傾向の変化について検討を加える。分析に採り上げるのは文末表現である。文末表現を採り上げる背景に、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[1]。つまり、夏目漱石の同時代の作家にとって、文末は文語表現から口語表現へと移行する上で容易に操作できる要素の 1 つであったと考えられる。

そこで、本研究では文末表現の変化について指摘した夏目漱石の小説を対象とし、句の終わり、すなわち文末表現に変化が認められるのか検討を加える。

表 1　分析対象と発表年

タイトル	掲載時期
吾輩は猫である（一～六）	1905年1月
幻影の盾	1905年4月
琴のそら音	1905年5月
一夜	1905年9月
薤露行	1905年11月
吾輩は猫である（七～十一）	1906年1月
趣味の遺伝	1906年1月
坊っちゃん	1906年4月
草枕	1906年9月
二百十日	1906年10月
野分	1907年1月
虞美人草	1907年6月23日～1907年10月29日
坑夫	1908年1月1日～1908年4月6日
文鳥	1908年6月
夢十夜	1908年7月
三四郎	1908年9月1日～1908年12月29日
それから	1909年5月31日～1909年8月14日
門	1910年3月1日～1910年6月12日
彼岸過迄	1912年1月1日～1912年4月29日
行人	1912年12月6日～1913年11月15日
こころ	1914年4月20日～1914年8月11日
道草	1915年6月3日～1915年9月14日
明暗	1916年5月26日～1916年12月14日

また、文献の成立順序の推定を目的とした研究は、芥川龍之介の文章について分析を行った研究が有名である。これは芥川龍之介の文章 309 編について分析を行った結果、

†1 一橋大学森有礼高等教育国際流動化機構
Mori Arinori Institute for Higher Education and Global Mobility

ⓒ 2019 Information Processing Society of Japan

表 2　形態素解析の例

本文	発音形	語意素読み	語意素	品詞	活用型	活用形
吾輩	ワガハイ	ワガハイ	我が輩	代名詞		
は	ワ	ハ	は	助詞-係助詞		
猫	ネコ	ネコ	猫	名詞-普通名詞-一般		
で	デ	ダ	だ	助動詞	助動詞-ダ	連用形-一般
ある	アル	アル	有る	動詞-非自立可能	五段-ラ行	終止形-一般
。			。	補助記号-句点		
名前	ナマエ	ナマエ	名前	名詞-普通名詞-一般		
は	ワ	ハ	は	助詞-係助詞		
まだ	マダ	マダ	未だ	副詞		
無い	ナイ	ナイ	無い	形容詞-非自立可能	形容詞	終止形-一般
。			。	補助記号-句点		

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[2]。しかし、このような文体的特徴の継時的な変化についての研究は広くなされておらず、まだ十分に展開しているとは言えない。

2. データについて

本研究で採り上げた小説は表 1 に示した 22 作品である。また、これら 22 作品の発表年も表 1 に示した。『吾輩は猫である』は第一から第十一までの 11 章で構成されており、第一から第六までは 1905 年に、第七から第十一は 1906 年に発表されている。そこで本研究では『吾輩は猫である』を発表年に応じて 2 つに分割した。また、『行人』の初出は 1912 年 12 月 6 日の朝日新聞であるが、完結したのは 1913 年 11 月 15 日であり、連載期間の大部分が 1913 年ため、本研究では 1913 年の作品として扱った。

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。青空文庫において公開されているテキストデータにはルビなどの注釈が本文中に記入されているが、本研究ではこれらの文字列を削除してから形態素解析を行った。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。また、夏目漱石の小説では英語などの日本語以外の表現や数式が本文中に現れるが、それらは分析から除外した。これに加え、会話文と地の文では作者が文体に対する意識が異なることが考えられる。金 (2009) においても、分析では会話文が削除されている[2]。よって、会話文を除外し地の文のみを対象とした。

先にふれたように、本研究では文末表現に注目する。文末を示す記号として文末として句点、閉かぎ括弧、エクスクラメーションマーク、クエスチョンマークの 4 つを選出した。本研究ではこれら 4 つの記号を文末記号と称する。また、本研究では文末記号の直前にはどのような単語が出現するのかという観点から分析を行った。上掲の夏目漱石の「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」という指摘にある「である」は助動詞「だ」及び動詞「ある」の 2 語であり、「のだ」

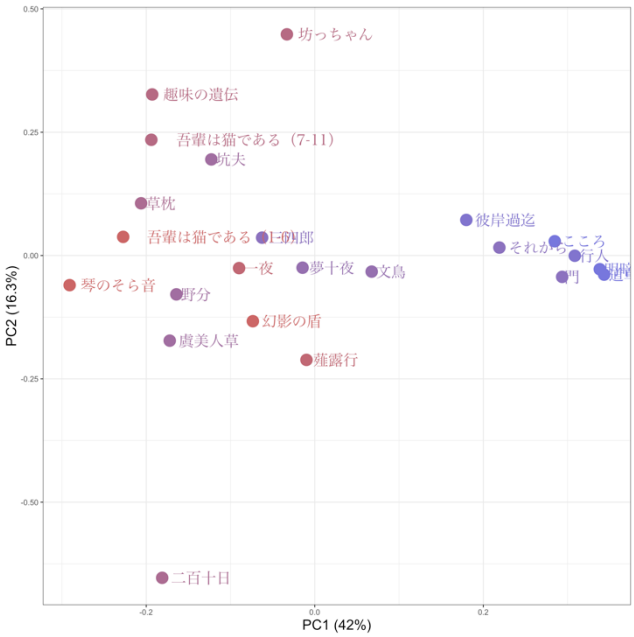

図 1　文末の bigram についての主成分分析の結果

は準体助詞「の」及び助動詞「だ」の 2 語であることから、

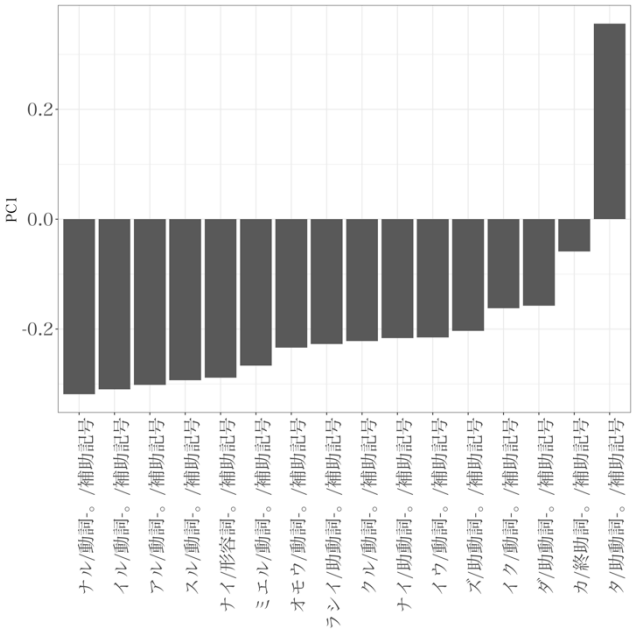

図 2　文末の bigram の主成分負荷量

本研究では文末記号の直前の 1 語に加えて、直前の 2 語について調査を行った。つまり、本研究の特徴量は文末の bigram 及び trigram であると言える。表 2 は『吾輩は猫である』の冒頭の形態素解析の結果である。『吾輩は猫である』の冒頭を例とすると、文末の bigram を分析するときに集計される特徴量は「ある/動詞 - 。/補助記号」及び「無い/形容詞 - 。/補助記号」である。次に、文末の trigram を分析するときに集計される特徴量は「で/助動詞 - ある/動詞 - 。/補助記号」及び「まだ/副詞 - 無い/形容詞 - 。/補助記号」である。

ⓒ 2019 Information Processing Society of Japan

3. 分析

3.1 特徴量について

本研究での特徴量の出現率について、相関係数行列を用いた主成分分析を行った。特徴量の出現率は bigram 及び trigram の総度数に対する各々の特徴量の出現頻度の割合である。

3.2 文末の bigram についての分析

まず、文末記号と直前の 1 単語、すなわち文末の bigram について分析を行った。分析には出現頻度上位 16 変数であり、これは全体の 90%を超える最小の累積度数である。分析の結果、求められた第 1 主成分の主成分得点は図 1 に示す通りであり、『文鳥』及び『それから』以降の 7 作品の主成分得点が正の値となる。なお、散布図中においてドットの色が赤い作品は出版年が早く、青い作品は出版年が遅いことを意味している。また、図 2 は第 1 主成分の主成分負荷量を棒グラフにした図であり、分析に用いた 16 変数のうち「タ/助動詞 - 。/補助動詞」のみが正の値となり他 15 変数は負の値となる。すなわち、『文鳥』と『それから』以降の 7 作品は他の小説に比べて助動詞の「た」で文が終わることが多いと考えられる。『文鳥』は 1908 年に発表された小説であるが、『それから』は 1909 年に発表された小説であるから、遅くとも 1909 年には夏目漱石の文末に対する意識は変化したと考えることができる。

3.3 文末の trigram についての分析

文末記号とその直前の 2 単語、すなわち文末の trigram について検討を加えた。分析では出現頻度上位 582 変数を分析に使用した。582 変数までの累積度数は全体の 90.2%となり、これは 90%を超える最小の度数である。これら 582 変数について主成分分析を行った結果、図 3 に示したように第 1 主成分の主成分得点は『文鳥』、『三四郎』以降の 8 作品が正の値となる。1908 年に発表された小説は『坑夫』『文鳥』『夢十夜』『三四郎』の 4 作品であるが、『坑夫』及び『夢十夜』の主成分得点は負、『文鳥』及び『三四郎』の主成分得点は正となる。このことから、文末に現れる 2 単語に注目し、分析を行った場合も 1908 年頃に夏目漱石の小説が文末の表現が変化していると考えられる。次いで、図 4 は第 1 主成分の主成分負荷量であり、主成分負荷量の大きい 10 変数と小さい 10 変数を図示した。主成分負荷量が正となる 10 変数はすべて助動詞の「た」が含まれるのに対し、負となる 10 変数は助動詞の「た」が含まれず、用言で文末となることが多い。よって、図 3 において主成分得点が正になった 9 作品では「スル/動詞 - タ/助動詞(した)」「ナイ/助動詞 - タ/助動詞（なかった)」「アル/動詞 - タ/助動詞（あった)」「イル/動詞 - タ/助動詞（いた)」などの表現が文末に多用されていると言える。一方で、図 3 において主成分得点が負となった小説では、「テ/接続助詞 - イル/動詞」「テ/接続助詞 - ミル/動詞」「ハ/係助詞 - ナイ/助動詞」「テ/接続助詞 - アル/動詞」などが文末に多用されていると考えられる。

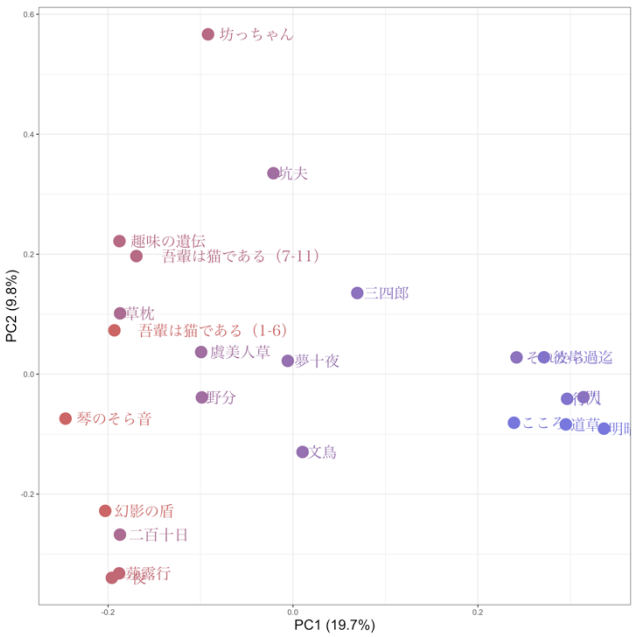

図 3　文末の trigram についての主成分分析の結果

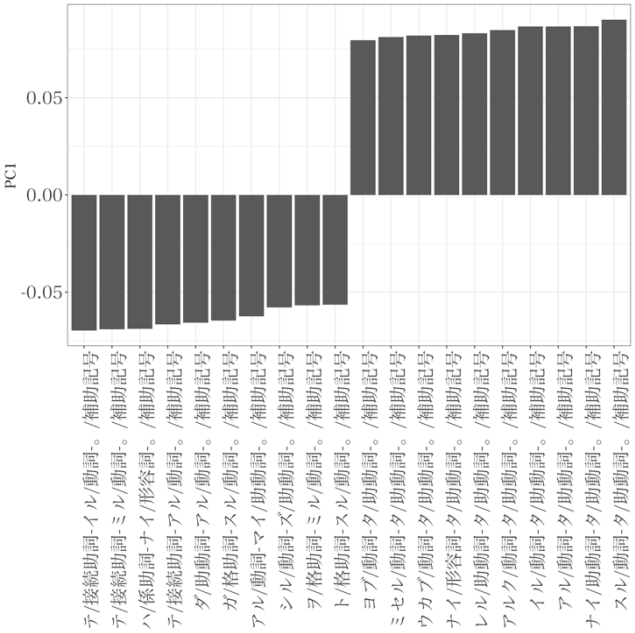

図 4　文末の trigram についての主成分負荷量

このように、会話文以外の文末に現れる単語について分析を行った結果、1908 年頃に文末表現が変化していると考えられる。特に、文末に助動詞の「た」が 1909 年以降の小説に多用されるようになったと言える。

4. おわりに

本研究における分析を通じて、主成分得点の考察から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

れると考えられる。特に 1909 年に発表された『それから』以降の 7 作品はどちらの分析においても同様の傾向を有している。主成分負荷量に対する考察から 1908 年あるいは 1909 年に認められる文末における文体的特徴の変化として、文末に助動詞を用いるようになったことがあげられる。特に、助動詞の「た」の使用が増加していると考えられ、会話文を除外した地の文に対する分析ではこの傾向は顕著である。

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. 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, 2019, Vol. 13, pp. 107-115.
[5] 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.

ⓒ 2019 Information Processing Society of Japan