助詞の相対頻度と漱石10作品の分類

留学生教育 第10号，2013
（琉球大学留学生センター紀要）

助詞の相対頻度と漱石10作品の分類

金 城 克 哉

1. はじめに

計量言語学の分野では単語の長さ、文の長さ、品詞の分布など様々な情報を用いて文章の著者の特徴を分析する研究が進められ、成果を上げている（金・村上2003）。村上（2001）は「書き手の文体の特徴は、意識して書くところにではなく、無意識に書くところに現れやすいと考えられる」とし、読点がどの文字の後にどの程度つけられているかの情報に基づき、井上靖・中島敦・三島由紀夫・谷崎潤一郎の4人の作家の9作品を分析し、「井上は『と』の後に読点が多く、中島は『て』『で』の後が少なく、『し』の後が多い」というように作家ごとの特徴を引き出すことに成功している。同じく村上（2010）は源氏物語54巻の成立順序を助動詞の出現率を用いて分析し、また金（1996, 2001, 2002a, 2002b, 2009）は読点と同じく助詞の分布も作家を特徴づける要因であることを論証している。

このような計量分析の成果を背景に、大久保・大久保(2011)は夏目漱石の10作品について、読点の前の文字とその相対頻度の情報を階層的クラスター分析と主成分分析で分析し、作品の分類結果が漱石の精神状態（病期と軽快期）と関連があることを明らかにしている。

この小論では読点の前の文字ではなく、作品に用いられる助詞の相対頻度情報を用いて大久保・大久保（2011）の分析を追試することを目的とする。

2. 大久保・大久保（2011）の分析

大久保・大久保（2011）は読点の施し方と漱石の精神状態との関連について調べるために「漱石作品における読点の付く文字の頻度を計量し、階層的クラスター分析と主成分分析による作品の分類」（2011：21）を試みている。分析対象となったのは以下の10作品である：

分析対象作品：「吾輩は猫である」、「坊っちゃん」、「草枕」、「虞美人草」、「三四郎」、「それから」、「門」、「こころ」、「硝子戸の中」、「道草」

30代後半の3作品（「吾輩は猫である」、「坊っちゃん」、「草枕」）、40代前半の作品（「虞美人草」、「三四郎」、「それから」、「門」）、40代後半の作品（「こころ」、「硝子戸の中」、「道草」）という3群分類（病期と軽快期にそれぞれ対応する、表1参照）が読点の付く文字の頻度と関連しているかどうかが大久保・大久保（2011）の研究の焦点である。この背景には高橋(2009)の作品の執筆時期と漱石の病期についての次のような指摘がある：

> このように漱石の人生には、20代後半、30代後半、40代後半にそれぞれ数年にわたって精神的な変調をきたす時期があったのだが、注目すべきは、漱石が30代後半、すなわち第Ⅱの病期の最中に創作活動を始めていることである。明治38年1月に『吾輩は猫である』の第１回を発表したのを皮切りに翌明治39年には『坊っちゃん』や『草枕』『二百十日』を発表するなど、漱石の初期の代表作はいずれも第Ⅱの病期に書かれている。また、大正2年から顕著になった第Ⅲの病期では、『行人』や『こころ』『道草』など後期の傑作が書かれているため、漱石の作品のかなりの部分は、彼の精神状態が不安定な時期に書かれたことになる。しかも漱石の精神変調は、いわゆる神経衰弱にとどまらず、いずれの病期でも幻覚や妄想を中心とした精神病レベルのものであるため、こうした精神状態が彼の創作活動に何の影響も及ぼさなかったと考える方が、むしろ不自然だろう。
>
> （2009：ii-iii）

漱石の創作活動に彼自身の精神状態（病期）が影響を及ぼしたであろうことは想像に難くない。上述したように、村上（2001）は「書き手の文体の特徴は、意識して書くところにではなく、無意識に書くところに現れやすいと考えられる」としている。大久保・大久保（2011）は同じ漱石の作品においても、病期に書かれたものと軽快期に書かれたものとでは文体に何らかの差があるのではないかと推測したのであった。

大久保・大久保（2011）は青空文庫を用いて、読点の前の1文字とその頻度（読点総数に対する文字の比率）を求め、「上位25文字の頻度とそのほかすべての文字をひとまとめにした頻度の合計26変数を算出し、親縁性の指標であるユークリッド距離による最小分散法（ウォード法）を用いた階層的クラスター分析と主成分分析」を適用して作品を分類している。この結果、階層的クラスター分析「吾輩は猫である」と「坊ちゃん」からなる群、「草枕」、「虞美人草」、「三四郎」、「それから」、「門」からな

表1　漱石略年表（高橋(2009)をもとに作成）

1903年	(明治36年)	1月	英国留学を終え帰国。	病期Ⅱ
1903年	(明治36年)	4月	一高・東大の講師として英文学を講ずる。	病期Ⅱ
1905年	(明治38年)	1月～8月	「吾輩は猫である」を『ホトトギス』にて連載。	病期Ⅱ
1906年	(明治39年)	4月	「坊っちゃん」を『ホトトギス』に発表。	病期Ⅱ
1906年	(明治39年)	9月	「草枕」を『新小説』に発表。	病期Ⅱ
1907年	(明治40年)	3月・4月	東大退職、朝日新聞社入社	軽快期
1907年	(明治40年)	6月～10月	「虞美人草」を『朝日新聞』にて連載。	軽快期
1908年	(明治41年)	9月～12月	「三四郎」を『朝日新聞』にて連載。	軽快期
1909年	(明治42年)	6月～10月	「それから」を『朝日新聞』にて連載。	軽快期
1910年	(明治43年)	3月～6月	「門」を『朝日新聞』にて連載。	軽快期
1914年	(大正3年)	4月～8月	「こころ」を『朝日新聞』にて連載。	病期Ⅲ
1915年	(大正4年)	1月～2月	「硝子戸の中」を『朝日新聞』にて連載。	病期Ⅲ
1915年	(大正4年)	6月～9月	「道草」を『朝日新聞』にて連載。	病期Ⅲ

る群、そして「こころ」、「硝子戸の中」、「道草」からなる群という3群分類が得られたとする（図1参照）。

この結果に対し、大久保・大久保(2011)は次のように評価している（アルファベットの略字はそれぞれ、Wa:「吾輩は猫である」，Bo:「坊ちゃん」，Ku:「草枕」，Gu:「虞美人草」，Sa:「三四郎」，So:「それから」，Mo:「門」，Ko:「こころ」，Ga:「硝子戸の中」，Mi:「道草」を表す）：

> 著者らの読点の施し方による作品の分類は、例外的なKu、Guも見られるものの、作者の神経衰弱と関連すると推測される。ただし、Wa、Bo、Kuが書かれた時期とKo、Ga、Miが書かれた時期はともに病期であるにもかかわらず作品がひとまとまりとはなっていない。この点は、40代後半は30代後半より症状の烈しさが減じるなどの病像の違いも知られており、そのことが関連しているかもしれない。
>
> (2011：25)

本稿ではこの大久保・大久保(2011)の読点の施し方とは別の、助詞の使用頻度をもとに同様の分類が可能かどうかを検証する。

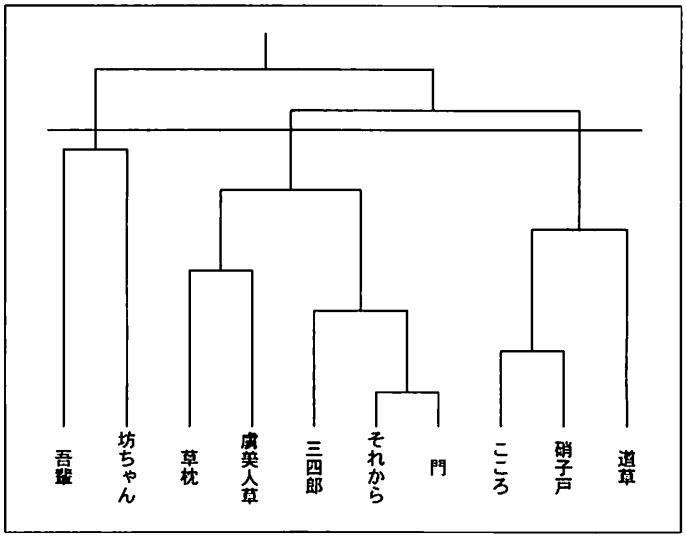

図1　大久保・大久保(2011)の分類結果（註1）

3. 本稿の分析方法

助詞の頻度の観点から上述した大久保・大久保(2011)の追試を行うという目的のために大久保・大久保で分析の対象となった同じ10作品を分析することとした。作品は青空文庫（http://www.aozora.gr.jp/）から当該作品のテキストファイルをダウンロードし、説明、ルビ、記号などを削除するクリーニングを行った後、立命館大学の樋口耕一氏によって開発されたフリーのテキストマイニングソフトウェア「KH Coder」を用いて機械的に助詞を抽出した。

日本語の助詞には格助詞・並立助詞・終助詞・間投助詞・副助詞・係助詞・接続助詞などがあるが、これらを一括して高頻度順に並べ上位の「の」から「ぞ」まで34の助詞を抽出し、金（2002）を参考に頻度順に上から25の助詞とそれ以外（「など」「くらい」「さえ」「わ」など）の9つを「その他」としてまとめた26の変数についてその相対頻度を求めた（表2）（註2）。これに対し、大久保・大久保(2011)と同様、ユークリッド距離による最小分散法（ウォード法）を用いた階層的クラスター分析と主成分分析による作品の分類を行った。階層的クラスター分析と主成分分析のための統計処理にはエスミの「Mac多変量解析Ver.2.0」を用いた。

4. 結果と考察

表2に助詞とその相対頻度を示した。頻度では「の」が最も多く使用され、それに

「に」、「は」、「て」、「を」が続いている。「の」の相対頻度が最も高かった作品は「硝子戸の中」で、それに「こころ」と「道草」が続いているところを見ると、後期作品では「の」の使用がおよそ全体の18%以上を占めている点が特徴としてあげられる。前期2作品の「吾輩は猫である」と「坊ちゃん」は第2位の「に」の相対頻度がおよそ11%以下となっているところが他の作品とは異なっている点としてあげられる。

表2：夏目漱石の10作品における助詞とその相対頻度（%）

	吾輩	坊ちゃん	草枕	虞美人草	三四郎	それから	門	こころ	硝子	道草
1 の	15.96	13.14	17.32	15.74	15.29	16.22	16.21	18.83	20.06	18.62
2 に	10.75	9.84	13.25	13.31	11.51	13.41	13.07	13.05	13.47	14.30
3 は	11.05	10.60	10.66	12.97	11.56	12.64	10.50	13.40	11.12	12.41
4 て	11.85	13.23	11.10	10.74	13.41	9.82	13.34	10.98	12.40	11.02
5 を	10.39	10.19	10.75	12.12	10.78	12.28	11.82	10.40	10.48	11.37
6 が	9.18	10.46	9.17	7.92	9.61	7.99	7.78	7.08	7.40	6.96
7 と	9.60	9.68	7.88	8.38	8.39	7.92	7.57	7.06	6.25	5.14
8 も	4.42	4.06	4.56	3.76	3.34	3.73	3.98	4.36	3.63	4.68
9 から	3.48	4.31	2.67	2.74	3.24	3.29	2.93	2.82	3.10	2.99
10 で	3.46	3.78	2.72	2.52	3.03	3.54	2.82	3.00	2.94	3.07
11 か	2.86	2.75	2.90	2.96	2.50	2.45	1.60	2.22	2.39	2.32
12 へ	1.69	2.81	1.76	1.55	2.44	2.35	2.42	1.67	1.50	1.50
13 ば	1.08	0.77	1.51	1.24	0.61	0.69	0.57	0.75	0.60	0.92
14 まで	0.53	0.63	0.62	0.57	0.65	0.01	0.72	0.56	0.56	0.64
15 や	0.35	0.49	0.26	0.27	0.15	1.31	0.36	0.28	0.29	0.28
16 だけ	0.41	0.32	0.38	0.37	0.46	0.00	0.57	0.56	0.44	0.56
17 でも	0.47	0.45	0.38	0.41	0.22	0.29	0.33	0.37	0.55	0.49
18 ながら	0.29	0.33	0.33	0.40	0.44	0.47	0.56	0.31	0.29	0.33
19 より	0.28	0.38	0.29	0.38	0.28	0.42	0.38	0.44	0.33	0.46
20 ばかり	0.39	0.57	0.25	0.40	0.46	0.08	0.27	0.24	0.39	0.34
21 ほど	0.26	0.25	0.47	0.40	0.39	0.01	0.44	0.26	0.22	0.28
22 けれども	0.05	0.08	0.01	0.11	0.37	0.58	0.42	0.43	0.32	0.38
23 ので	0.26	0.13	0.04	0.02	0.24	0.28	0.64	0.30	0.65	0.29
24 くらい	0.38	0.18	0.08	0.11	0.15	0.00	0.10	0.19	0.23	0.00
25 さえ	0.14	0.10	0.21	0.16	0.05	0.00	0.15	0.13	0.16	0.37
26 その他	0.42	0.48	0.42	0.45	0.42	0.22	0.43	0.34	0.23	0.27

図2にクラスター分析結果を、図3に主成分分析の結果を示す。第1クラスター（病期１）に「吾輩」と「坊ちゃん」が分類されたものの、期待された「草枕」は第2クラスターに分類されている。第2クラスターは、第１クラスターに分類が期待された「三四郎」が入っているが、その他の3作品が同一クラスターに分類された。第3クラスターは病期2に対応するもので、3作品が分類された。

大久保・大久保(2011)でも「草枕」が第１クラスターに分類されないという結果となっていたが、「草枕」のころは既に回復しつつあったと考えられる。再び高橋の指摘を引用する：

「このような見地からすれば、漱石が『草枕』でとなえた『非人情』や『明暗』執筆時の『則天去私』といった心境も、従来の東洋哲学的な解釈の他に、病的体験からの回復による心理的安定という、より漱石の内面に即した捉え方ができるのではあるまいか」（2009：92）。

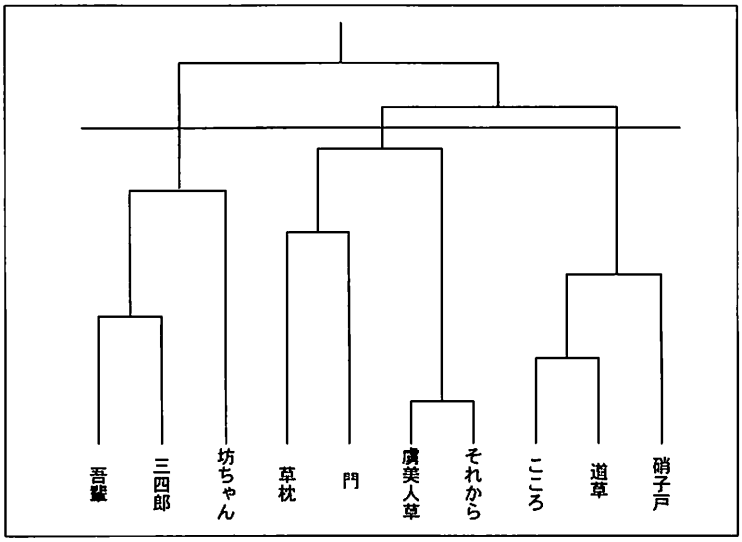

図2　夏目漱石10作品における助詞の相対頻度による分類樹形図

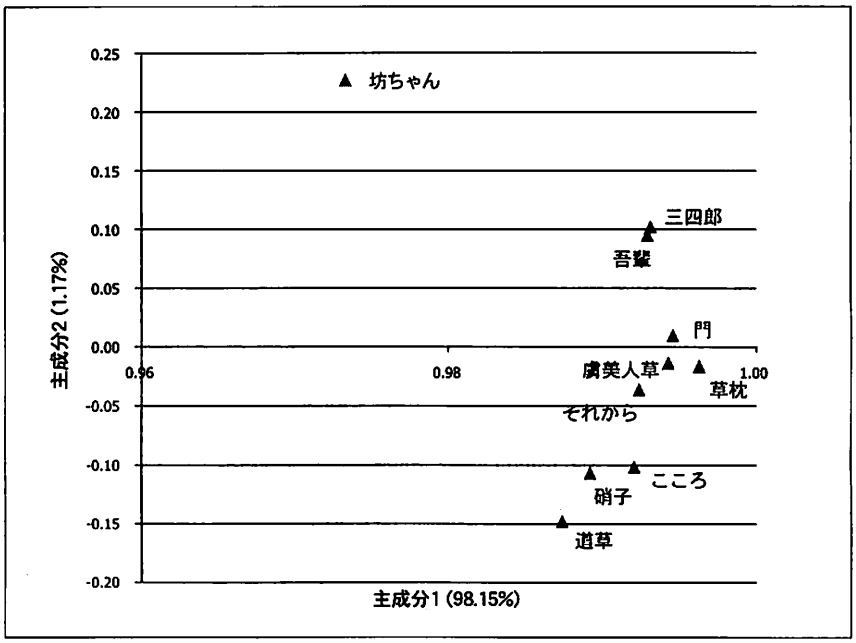

図3　夏目漱石10作品における助詞の相対頻度による主成分分析の散布図

仮にこの高橋の分析が正しいとすれば、「草枕」が大久保・大久保(2011)においても今回の追試においても軽快期である第2クラスターに分類されたことも納得がいく。問題は「三四郎」が第１クラスターに分類されている点である。これについては、軽快期にありながらも実は病いに苛まれていたという推測も可能であろう。しかし現段階ではそれを裏付ける証拠はない。これについては今後の検討課題としたい。

5. まとめ

この小論では大久保・大久保(2011)において読点の施し方の観点から夏目漱石の10作品の分類がなされた研究について、助詞の相対頻度という別の角度から同様に作品分類が可能であるかの追試を行った。結果、概ね漱石の病期と軽快期で作品が分類される、すなわち助詞の相対頻度という文体上の特徴が漱石の精神状態と関わりがあることがわかった。本稿では扱うことができなかったが、多変量解析という分析手法には階層的クラスター分析や主成分分析以外にも様々な方法があり、特に金（2002b）で取り上げられている自己組織化マップは有効な解析方法になる可能性があると思われる。将来的にはそのような分析方法を視野に入れた研究を行いたいと考える。

註

(1) この図は大久保・大久保(2011：23)の表１「夏目漱石の10作品における読点の前の文字とその相対頻度」をもとに筆者が作成したもので、大久保・大久保(2011：24)の図1と同等のものである。しかし、この階層的クラスター分析について、大久保・大久保(2011：22)は「ユークリッド距離による最小分散法（ウォード法）を用いた」としているが、今回筆者が大久保・大久保(2011：23)の表１のデータを用いて改めて分析を行ったところ、ウォード法では図1の結果は得られなかった。本稿の図1は最長距離法によるものである。なお、以下で述べる本稿の階層的クラスター分析では、最長距離法・ウォード法いずれでも同じ結果が得られた。

(2) 今回は相対頻度をもとに分析した結果を示したが、100語当たりの調整頻度で分析した結果も同様のものとなった。

謝辞

KH Coderの操作に関して立命館大学の樋口耕一先生にご指導いただきました。この場を借りて感謝申し上げます。

参考文献

大久保起延・大久保博美（2011）「読点の施し方と漱石作品の分類」『計量国語学』28（1），21-26.

金明哲（1996）「S8－5助詞の分布に基づいた文章の原著者の認識」『日本行動計量学会大会発表論文抄録集』24，144-147.

金明哲（2001）「助詞のN-gram分布に基づいた書き手の識別」『日本行動計量学会大会発表論文抄録集』29，298-299.

金明哲（2002a）「助詞の分布における書き手の特徴に関する計量分析」『社会情報』11(2)，15-23.

金明哲（2002b）「自己組織化マップと助詞文を用いた書き手の識別と特徴分析」『日本行動計量学会大会発表論文抄録集』30，194-197.

金明哲（2009）「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』36（2），89-103.

金明哲・村上征勝（2003）「文章の統計分析とは」甘利俊一ほか（編）『言語と心理の統計：ことばと行動の確率モデルによる分析』岩波書店，3-57.

高橋正雄（2009）『漱石文学が物語るもの：神経衰弱者への畏敬と癒し』みすず書房

村上征勝（2001）「"ことば"新研究―統計分析への誘い―」『ことば工学研究会』 8，23-27.

村上征勝（2010）「文献の計量分析―源氏物語を中心に」『日本語学』29(1)，50-61.

ソフトウェア

株式会社エスミ　『Mac多変量解析Ver.2.0 アカデミック版』

樋口耕一　『KH Coder Ver.2b29』

（琉球大学法文学部）