

『計量国語学』アーカイブ

ID	KK290609
種別	ノート
タイトル	助詞・助動詞の使い方と漱石作品の分類
Title	Statistical distribution of postpositional particles and auxiliary verbs and the classification of the literary works of <i>Natsume Soseki</i>
著者	大久保 博美, 大久保 起延
Author	OHKUBO Hiromi and OHKUBO Tatsunobu
掲載号	29巻6号
発行日	2014年9月20日
開始ページ	227
終了ページ	234
著作権者	計量国語学会

研究ノート

助詞・助動詞の使い方と漱石作品の分類

大久保 博美 (駿河台日本大学病院精神神経科)

大久保 起延 (日本大学医学部附属板橋病院精神神経科)

要旨

夏目漱石の小説 9 篇と随筆 1 篇を助詞・助動詞の出現の仕方により分類した。対象とした作品 (略号) は、書かれた順に、吾輩は猫である (Wa), 坊ちゃん (Bo), 草枕 (Ku), 虞美人草 (Gu), 三四郎 (Sa), それから (So), 門 (Mo), こころ (Ko), 硝子戸の中 (Ga), 道草 (Mi) であった。助詞・助動詞の出現頻度を算出し、階層的クラスタ分析と主成分分析を行った。その結果、Wa, Bo, Ku, Gu の群と Sa, So, Mo, Ga, Ko, Mi の群に分類され、後者は Sa, So, Mo, Ga の群と Ko, Mi の群に分類され、随筆 1 篇 (Ga) を除けば小説 9 篇は書かれた時期ごとに群を形成した。この結果は作品の傾向・特徴の変遷と関係するとみられるが、作家の精神変調の推移と関係するとしても矛盾しないと考えられた。文学作品の助詞・助動詞を通して、作家の病的なものを含む情緒や思想へのアプローチができることが示唆された。

キーワード: 助詞・助動詞, 階層的クラスタ分析, 主成分分析, 夏目漱石, 精神状態

1. 背景と目的

これまで、日本文学を対象にして読点や助詞の出現の仕方に関する計量分析がなされてきている。金・樺島・村上 (1993) は、読点がどの文字の後に付くかで作品を分類すると、作家ごとの分類とよく一致することを見出した。大久保・大久保 (2011) は、このような読点の出現の仕方により分類すると、夏目漱石の作品が書かれた時期ごとに群を形成することを示した。このことから、読点の出現の仕方には作家間だけでなく作家内でも差異があることがわかった。また、金 (2009) は読点と助詞の出現の仕方により、芥川龍之介の作品の書かれた時期が推定できることを示した。助詞は助動詞などとともに文章の内容にあまり依存しないため、助詞・助動詞の出現の仕方を観察することにより、作品を通した作家の情緒や思想にかかわるプロフィールを描くことができるのではないかと考えられる。

夏目漱石の作品のいくつかは、作家自ら述べているところの神経衰弱と狂気のなかで書かれている (夏目漱石, 1995 再版)。漱石作品の読点の出現の仕方による分類 (大久保・大久保, 2011) は、病跡学の知見からの病期分類 (高橋, 2009) とよく一致する結果であった。しかし、読点の出現の仕方と作家の情緒や思想を結びつけるのはそれほど容易ではない。今回、著者らは読点から焦点を移し、夏目漱石の作品における助詞・助動詞の出現

頻度を測定し、階層的クラスター分析と主成分分析により作品の分類を行った。

2. 方法

2.1 対象とした作品

夏目漱石の中・長編小説 9 篇と随筆 1 篇の合計 10 篇を選択し、青空文庫 (<http://www.aozora.gr.jp/>) に公開されている作品で、新字体、新仮名遣の表記のものをを用いた。前述した読点の出現の仕方による分類で対象とした作品と同じ青空文庫の 10 篇とした。作品(略号)は書かれた順に、吾輩は猫である (Wa), 坊ちゃん (Bo), 草枕 (Ku), 虞美人草 (Gu), 三四郎 (Sa), それから (So), 門 (Mo), こころ (Ko), 硝子戸の中 (Ga), 道草 (Mi) であった。夏目漱石には中・長編小説 14 篇と随筆 2 篇の著作がある。取り上げた 10 篇は 38 歳 (Wa) から 48 歳 (Mi) までに及び、漱石が小説家として活動した時期を通じた傾向を見ることができる。

2.2 計量分析

文章を電子化する際に付加された説明, ルビ, 記号を削除した後に, 茶筌(松本・高岡・浅原, 2007)により形態素解析を行い, 品詞タグにより助詞・助動詞を抽出した。助詞は, 格助詞, 副助詞等 6 種類の品詞タグ 2 階層目まで区別した。10 篇の総語彙量(句読点などの符号を除く)は延べ語数 911131, 異なり語数は 29723 で, そのうち助詞の延べ語数 306361, 異なり語数 143, 助動詞の延べ語数 108077, 異なり語数 198 であった。このうち, 漢字とかなという点でのみの違いの表記ゆれを統一した。また, 助詞・助動詞の両方の用法を持つ語が, 「なり」, 「や」, 「な」, 「ね」, 「で」, 「し」, 「た」の 7 種類観察されたが, すべて区別して扱った。その結果, 127 種類の助詞, 182 種類の助動詞を分析の対象とした。助詞・助動詞はどれも文を構成するうえで重要な役割を果たすことから, 今回はこれらすべてを対象とし, 今後に個別の形式の用法を吟味したい。分析の方法は金・樺島・村上(1993)に従った。語彙ごとの助詞・助動詞の出現頻度(助詞・助動詞の総数に対する比率)を算出し, 助詞・助動詞の出現頻度の上位 25 の語彙を変数とし, それより下位の語彙を全て加算した 1 変数(「その他」)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, ソフトウェア R (<ftp://ftp.ecc.u-tokyo.ac.jp/CRAN/>) を用いた。

3. 結果

3.1 助詞・助動詞の出現頻度

表 1 に作品ごとの総語彙量と助詞・助動詞の出現頻度と出現率を示した。総語彙量は 32444 (Ga) から 187946 (Wa) までであった。10 作品の助詞・助動詞の出現率を平均すると助詞 33.7%, 助動詞 11.9% であった。

表 2 に語彙ごとの助詞・助動詞のうち出現頻度 25 位までを示した。10 篇の出現頻度は高い順に, 「の」(格助詞), 「て」(接続助詞), 「た」(助動詞), 「は」(係助詞), 「を」(格助詞) などであった。「の」(格助詞)は, Wa, Ku, Gu, Ga の 4 篇で, 「て」(接続助詞)は Bo, Sa の 2 篇で, 「た」(格助詞)は So, Mo, Ko, Mi の 4 篇で頻度が最も高かった。

表 1： 漱石 10 作品の総語彙量と助詞・助動詞の出現頻度および出現率

	Wa	Bo	Ku	Gu	Sa	So	Mo	Ko	Ga	Mi
総語彙量*	187946	51847	48816	114927	94390	107286	85588	98463	32444	89424
助詞数	63446	17287	16884	38410	31794	35787	29224	33084	11030	29415
出現率 (%)	33.8	33.3	34.6	33.4	33.7	33.4	34.1	33.6	34	32.9
助動詞数	20729	6460	4718	12272	10826	12587	9929	14992	3831	11733
出現率 (%)	11.0	12.5	9.7	10.7	11.5	11.7	11.6	15.2	11.8	13.1

*:句読点などの符号を除く

表 2： 漱石 10 作品における助詞・助動詞の出現率の上位 25 位までとその他 (%)

	Wa	Bo	Ku	Gu	Sa	So	Mo	Ko	Ga	Mi
の/格助詞	8.98	7.41	11.33	12.28	9.44	9.82	10.08	9.04	11.44	10.89
て/接続助詞	8.72	9.34	8.61	9.73	9.94	9.62	10.12	7.97	9.41	8.10
た/助動詞	4.83	7.17	4.75	7.25	9.67	11.11	11.87	11.34	10.26	12.24
は/係助詞	7.70	7.29	7.74	11.31	8.19	8.61	7.35	8.63	7.75	8.14
を/格助詞	7.27	6.97	7.81	10.61	7.57	8.34	8.26	6.70	7.23	7.45
に/格助詞	6.26	5.43	8.80	10.02	6.90	7.88	8.16	7.37	8.49	8.41
と/格助詞	6.19	5.93	5.43	7.14	5.61	5.31	4.95	4.92	4.84	3.76
が/格助詞	5.16	5.47	5.83	6.03	6.03	4.49	4.70	4.15	4.45	4.18
も/係助詞	3.87	3.47	4.00	4.20	3.18	3.27	3.72	3.83	3.50	4.12
だ/助動詞	3.24	4.90	1.75	3.76	2.58	1.89	1.48	1.44	1.58	1.95
で/助動詞	2.53	1.61	1.99	2.59	2.42	2.18	2.22	1.54	2.73	1.86
に/助動詞	2.19	2.02	1.66	2.84	2.05	1.86	1.84	2.14	1.76	2.09
で/格助詞	2.34	2.40	1.99	2.20	2.06	1.80	1.81	1.94	2.11	2.10
の/準体助詞	2.21	1.73	1.23	1.41	1.39	1.43	1.74	3.35	2.82	1.47
な/助動詞	1.67	1.76	1.12	1.63	1.48	1.59	1.50	1.44	1.67	1.79
ない/助動詞	1.30	2.05	0.75	1.62	1.80	1.19	1.13	1.40	1.58	1.48
へ/格助詞	1.15	1.89	1.29	1.33	1.69	1.45	1.69	1.06	1.03	0.96
から/格助詞	1.12	1.17	1.25	1.40	1.41	1.52	1.53	1.21	1.55	1.25
か/副助詞	1.24	1.39	1.31	1.12	1.05	0.98	0.99	0.89	1.33	1.05
から/接続助詞	1.44	1.82	0.77	1.15	1.08	0.95	0.69	0.84	0.83	0.84
です/助動詞	1.15	0.78	0.89	1.29	0.84	0.61	0.46	2.35	0.53	0.67
か/終助詞	1.18	1.02	1.16	1.87	1.09	0.95	0.41	0.85	0.76	0.78
と/接続助詞	1.03	1.26	0.87	0.73	1.23	0.82	1.03	0.70	0.84	0.53
が/接続助詞	1.26	1.72	0.87	0.92	0.84	0.73	0.80	0.52	0.71	0.37
ん/準体助詞	0.90	1.00	0.49	1.31	0.71	0.66	0.47	0.45	0.22	1.03
その他	15.08	13.00	16.32	18.84	9.73	10.93	10.99	13.92	10.59	12.48

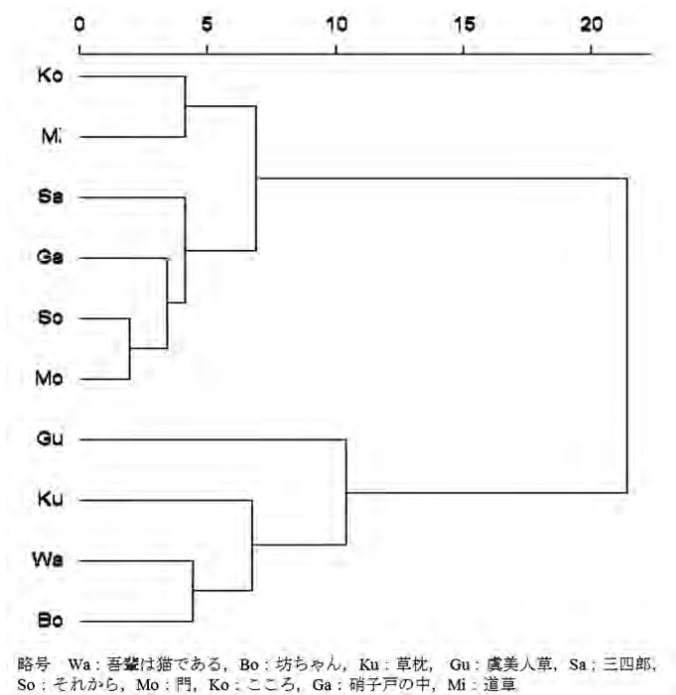


図1：漱石10作品における助詞・助動詞の出現率による階層的クラスター分析の樹形図

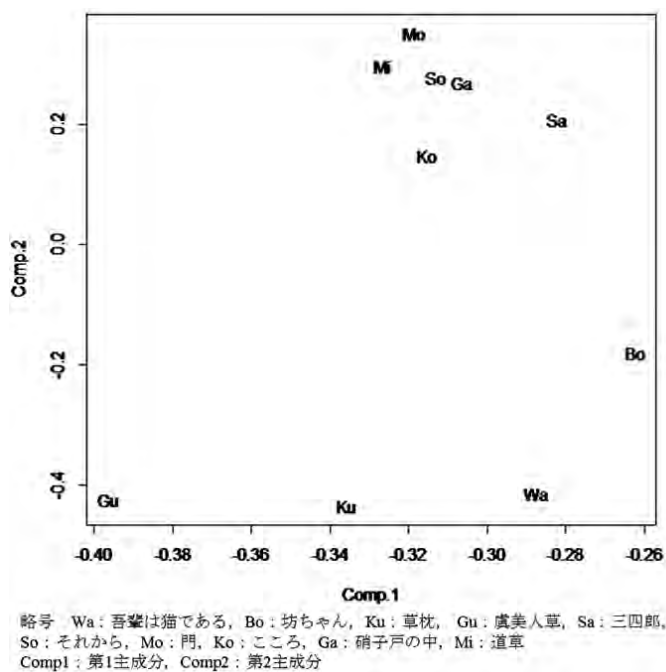


図2：漱石10作品における助詞・助動詞の出現率による主成分分析の散布図

3.2 作品の分類

図1に階層的クラスタ分析の樹形図を示した。コーフェン相関係数は0.815であった。樹形図を高さ15付近で区切ると Wa, Bo, Ku, Gu (書かれた順, 以下同様) からなる群と Sa, So, Mo, Ko, Ga, Mi からなる群に分かれる。Gu を除けば, それぞれ30代の作品と40代の作品である。前者は高さ10付近で区切ると Gu が分かれ, 高さ5付近で区切るとさらに Ku が分かれる。後者は高さ5付近で区切ると Sa, So, Mo, Ga からなる群と Ko, Mi からなる群に分かれる。Ga を除けば, それぞれ40代前半と40代後半の作品である。

図2に第1主成分得点と第2主成分得点を用いた主成分分析の散布図を示した。第1主成分と第2主成分の累積寄与率は0.982であった。散布図では Wa, Bo, Ku, Gu の群と Sa, So, Mo, Ko, Ga, Mi の群とが, それぞれ下部と上部とに分かれて分布した。そして, 前者はひとつひとつの作品が離れて分布し, 後者は作品が近くにまとまって分布した。これらの主成分分析の結果は, Gu は Ku, Wa, Bo から分離でき, Ko, Mi は Sa, Ga, So, Mo から分離できることも含め, クラスタ分析の結果と矛盾しない。

4. 考察

4.1 対象としたデータについて

対象とした作品は, 夏目漱石が作家として活動したさまざまな時期に書かれたものを選んだ(朝日新聞2007)。この研究では, 文章を新字体, 新仮名遣の表記のものに統一しており, 原文のものは含まれないことを記しておく。10篇には, 小説9篇とともに随筆1篇(Ga)が含まれていた。随筆を除外しなかった理由は, 前述した読点の出現の仕方の結果と比較するためであった。

表1に示したように10作品を通して助詞, 助動詞とも安定して出現しており, 助詞と助動詞を合わせた出現率は約45%に及ぶため, 統計的に作品の傾向や特徴をよく捉えられると考えられる。助詞については茶釜の6種類の品詞タグ2階層目まで区別して扱ったが, 接続助詞, 終助詞など個々を抜き出での検討は行わなかった。今回は, 自動解析により客観的な作品の傾向や特徴を観察したいと考えた。個々の助詞を抜き出し, その際選別が過度に恣意的なものとなることを避けつつ検討することはこれからの課題としたい。

4.2 読点の出現の仕方による分類との比較

助詞・助動詞の出現の仕方によるクラスタ分析では, Wa, Bo, Ku, Gu の群と Sa, So, Mo, Ko, Ga, Mi の群に分かれ, 後者はさらに, Sa, So, Mo, Ga, Mi の群と Ko, Mi の群に分かれた。著者らの読点の出現の仕方によるクラスタ分析では, Wa, Bo の群と Ku, Gu の群, Sa, So, Mo の群および Ko, Ga, Mi の群に分かれた(大久保・大久保, 2011)。すなわち, 読点でも助詞・助動詞でも, 出現の仕方から分類すると Wa, Bo が同じ群, Sa, So, Mo が同じ群, Ko, Mi が同じ群となることで一致していた。これら7篇はそれぞれ30代後半, 40代前半, 40代後半の作品である。残る Ku, Gu, Ga の3篇は読点と助詞・助動詞との間で, 出現の仕方による分類は完全には一致しなかった。Ku には文語が用いられ, Gu は擬古文体で書かれていることが, 読点と助詞・助動詞の出現の仕方に差異をもたらしたのかもしれない。Ga はここでは唯一随筆であり, 読点の出現の仕方では執筆時期の近い小説 Ko, Mi と同じ群に分類されたが, 助詞・助動詞の出現の仕方では同じ群には分類されなかったことは, 小説と随筆というジャンルの違いのため

めと考えられた。Ga は 48 歳のときの作品であるのに、40 代前半の作品と同じ群に分類された理由を、さらに後で検討したい。

読点と助詞・助動詞の間で出現の仕方が近似した理由の 1 つとして、読点の分析の際には読点の前の 1 文字の出現頻度により測定していたため、助詞や助動詞も含まれていたことが考えられる。10 篇のうち 7 篇が同じように分類されたことから読点と助詞・助動詞の使い方はともに、書き手の精神の特性と状態を反映すると考えられる。

4.3 助詞や助動詞の出現の仕方を用いた他の研究との比較

最近、金城（2013）はわれわれと同じ 10 作品を対象に、形態素解析器や統計処理はわれわれとは異なるソフトウェアを用いて助詞だけの出現頻度を調べている。また助詞の抽出の際に階層分類による抽出をしないで一括して分析しているが、その結果は Wa, Bo, Sa の群, Ku, Gu, So, Mo の群, Ko, Ga, Mi の群に分かれ、概ね 30 代後半, 40 代前半, 40 代後半の作品群に分かれたと報告している。われわれの結果で示されているように、助動詞は助詞とともに出現頻度が高く、作品の傾向・特徴を示すものとして注目すべきと考えられる。

鈴木・影浦（2008）は第二次世界大戦後の 1945 年から 2006 年までの歴代の総理大臣 27 人の国会演説のオンラインデータベースを用い、助詞・助動詞の出現頻度を分析した。その結果、1945 年～1954 年の群、1954 年～1987 年の群、1987 年～2006 年の群と演説時期の違いにより 3 つの群に分かれると報告した。われわれの助詞・助動詞の結果はそれほど長くない期間の観察であり、作家を取り巻く時代の変化による影響は比較的小さいと想定される。

小林・小木曾（2013）は紫式部の『源氏物語』（うち「桐壺」「若紫」「橋姫」「夢浮橋」）と『紫式部日記』、および対照として菅原孝標女の『更級日記』を用いて、助詞・助動詞の出現頻度における書き手とジャンルによる影響を検討した。その結果、日記文学では格助詞の出現率が高く、物語文学では助動詞の出現率が高いことを観察した。そして対象とした中古和文では、ジャンルによる助詞・助動詞の出現頻度への影響が大きいと結論した。このことは、作品が書かれた時代は異なるものの、随筆（Ga）が同時期の小説群から分離されたわれわれの結果を支持している。また、この報告には、膠着語である日本語の助詞・助動詞が表現の論理や情緒を表すにあたって重要な働きを持つ点が述べられているが、われわれの漱石作品の助詞・助動詞の結果はこのことに矛盾しないと考えられる。

4.4 作家の精神変調との関係

文学作品は作中人物を通して作家自身を語るといわれる。作品の独創性は、しばしば精神医学の視点から、作家の精神の逸脱との関係が問われてきた。ある作家に精神の逸脱があれば、その文章は統計学的にも標準からの逸脱として観察できると考えられ、それは助詞・助動詞などの使用にも反映すると考えられる。

夏目漱石は 30 代後半に幻聴・妄想などの病的体験が出現し、40 代前半には病的体験は顕著ではなくなり、40 代後半にはそれに代わって抑うつに苦しんでいたことが指摘されている（高橋 2009）。作品 Wa, Bo, Ku は 30 代後半, Gu, Sa, So, Mo は 40 代前半, Ko, Ga, Mi は 40 代後半に書かれている。小説に限れば、40 歳のとき書かれた Gu が 30 代後半の作品と同じ群に分類されるものの、助詞・助動詞の出現の仕方と病期の分類がほぼ一致している。ただ、どちらも病期であるにもかかわらず 30 代後半と 40 代後半では同

じ群とはならなかった。また、樹形図と散布図の結果は、30歳代後半と比べると40代後半の作品群はまとまっており、助詞・助動詞の出現の仕方が収束していることが観察される。このことは病像の違いと関係するのかもしれない。夏目漱石が作品を重ねるごとに自己への洞察を深め、精神の安定を得ていたことも指摘されている（三浦 2008）。作品を書くことのこのような効果は、随筆である Ga が症状が軽快していた40代前半の小説と同じ群に分類されたことを説明するのかもしれない。小説では作家は作中人物ではないが、作中人物にならずにはその作品を書くことはできず、随筆では作家は主人公となるため自己洞察の効果も大きいと考えられる。しかし、ここでは1篇だけの結果であり、作品数を増やして検討されるべきことである。

英語圏では機能語の計量により、話し手（書き手）の心理が推測できることを示す知見が集積されてきており、機能語の中でも特に代名詞に焦点が当てられている（Chung & Pennebaker, 2007）。文学作品の機能語も分析され、作家の情緒や思想を推測することがなされつつある（Pennebaker & Ireland, 2011）。日本語では機能語として助詞・助動詞は主要なものであり、これらの出現の仕方を観察することは話し手（書き手）の情緒や思想を推測するのに客観性の高い指標となる可能性があると考えられる。

文献

- Chung, Cindy. & Pennebaker, James W. (2007) The psychological functions of function words, Fiedler, K (ed.) Social Communication. 343-359. New York: Psychology Press.
- Pennebaker, James W. & Ireland, Molly E. (2011) Using literature to understand authors. The case for computerized text analysis. *Scientific Study of Literature*, 1(1):34-48.
- 朝日新聞（2007）「漱石作品売り上げ十傑（新潮文庫）」10月17日夕刊3版13面・朝日新聞社.
- 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（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, 樺島忠夫, 村上征勝（1993）「読点と書き手の個性」『計量国語学』18(8):382-391.
- 金城克哉（2013）「助詞の相対頻度と漱石10作品の分類」『留学生教育：琉球大学留学生センター紀要』10:35-42.
- 小林雄一郎, 小木曾智信（2013）「中古和文における個人文体とジャンル文体」『国立国語研究所論集』6:29-43.
- 鈴木崇史, 影浦峯（2008）「総理大臣国会演説における基本的文体特徴量の探索的分析」『計量国語学』26(4):113-122.
- 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.
- 夏目漱石（1995再版）『漱石全集第十四巻. 文学論』14-15. 岩波書店.
- 松本裕司・高岡一馬・浅原正幸（2007）形態素解析システム『茶釜』version 2.4.4 (<http://chasen-legacy.sourceforge.jp/>).
- 三浦雅士（2008）『漱石. 母に愛されなかった子』岩波書店.

（2014年3月23日受付, 2014年5月31日再受付）

Note

Statistical Distribution of Postpositional Particles and Auxiliary Verbs and the Classification of the Literary Works of *Natsume Soseki*

OHKUBO Hiromi (Department of Psychiatry, Nihon University Surugadai Hospital)

OHKUBO Tatsunobu (Department of Psychiatry, Nihon University Itabashi Hospital)

Abstract:

The present study maps the use of postpositional particles and auxiliary verbs to explore how these words can reflect writer's pathological mental state. Ten literary works written by *Natsume Soseki* (1867-1916), who is a foremost novelist in modern Japan and is known to suffer from recurrent mental symptoms by biographical and pathographical studies, were analyzed to extract linguistic information using a Japanese morphological analysis system (*Chasen*). Data on the frequencies of the postpositional particles and auxiliary verbs were tested using hierarchical cluster analysis and principal component analysis. The results showed that the literary works were clustered roughly into three groups of his late thirties, early forties and late forties, which in accord with the course of the illness. Behaviormetric analysis of such function words can be a useful approach to the evaluation of writer's mental state.

Keywords: postpositional particles and auxiliary verbs, hierarchical cluster analysis, principal component analysis, *Natsume Soseki*, mental state