芥川龍之介における文体の経年変化
－ 主成分分析による計量的検証 －*

筒井昭博**

〈 目 次 〉

Key Words : 文体(style)、計量文体学(stylometry)、経年変化(chronological shift)、主成分分析(PCA: Principal Component Analysis)

Ⅰ. 序論

筆者推定(authorship attribution /authorship identification)は計量文体学(stylometry)[1]の中心領域の1つである。それは、あるテキストの文体[2]を計量的に分析することによりそのテキストの書き手を推定することである。文体の計量分析による筆者推定は欧米では長い歴史を持つ。その嚆矢は、使徒パウロが書いたとされる新約聖書中の書簡が本当にパウロ自身の手になるものかどうかに関

* 本論文は2011年度韓国外国語大学校校内学術研究支援費によるものである。
** 韓国外国語大学　日本語通翻訳学科　助教授,　日本語学・日本語教育学

1) 計量文体学とは文体の計量可能な特徴を分析する学問である(Kenny 1986, 1)。

2) 文体の定義は極めて難しいが(Wales 1990, 370)、ここでは「特定の内容を表すために可能なあらゆる表現の選択肢の中から書き手が特定の表現を選ぶ選択のしかた」(Argamon et al. 2010, 80)としておく。これはもともと文体にある範列的な関係(佐々木 1978, 39-40)を反映したものと言える。ただし、語のイデオム的な共起等により範列的な性質にも制約があること(Stubbs 2002, 225)に留意する必要がある。

してAugustes De Morganが1851年に発案した手法であるとされる(金・村上 2003, 5)。

筆者推定の根底にある基本的な論法は次のようなものである。すなわち、文体に関する特定の基準に基づき、2つのテキストが統計的に均一な母集団から得られたものと検証されるとき、この2つのテキストが同一の書き手になるとする推測の有利な根拠となる(Kenny 1986, 27)というものである。

この論法が成り立つためには、標本としての分析対象テキストの背後にある母集団として想定している文章の集合が文体に関して均一であるという前提が満たされなくてはならない。しかし特定の個人の文章の集合を母集団とみなしたとき、この母集団が文体に関して均一でないことはあり得る。Kenny(上掲書, 27)も指摘するように、主題や執筆時期の違いなどは特定個人の文体に変化をもたらす有力な要因である。そのため、特定の書き手の文体が執筆時期によって異なるか否か、つまり特定個人の文体の経年変化の有無を検証することは計量文体学における筆者推定において重要な意義を持つ。

筒井(2011)は、金(2009)の知見を受け、芥川龍之介の作品における文体の経年変化を検証し、助詞「が」「は」を文体指標とする限り芥川の文体の経年変化は否定できないことを明らかにした。しかし、文体とは多面的なものである。文体を捉える指標や手法が異なれば、結果が異なることもあり得る。逆に言えば、文体の変化が本当に存在するならば、異なる指標、手法で計測してもそれを検証できるはずである。また、筒井(上掲書)では、芥川作品の文体に経年変化があることは検証されたが、さらに踏み込んで「いかに」変化したかについては分析されていない。本稿は、芥川作品の文体がいかに変化したかを分析することをもって、その文体の経年変化を筒井(前掲書)とは異なる文体指標と手法を用いて再検証することを目的とする。

Ⅱ. 背景と目的

1. 文体の経年変化検証の重要性

前述のとおり、特定の書き手の文体の経年変化、つまり時間の推移とともに文体が変化するのかという視点を持つことは計量文体学における筆者推定において重要である。分析対象とするテキストの背後にある母集団としてのテキスト集合の文体的な均一性を前提とするか否かによって、筆者推定に関わる推論が全く異なり得るからである。それを端的に示すのが、村上(1996)による『源氏物語』の作者複数説の検証である。

村上(上掲書)は、『源氏物語』の作者複数説を検討すべく、テキストとして一定以上の規模を持つ44巻を対象とし、各巻の8品詞の使用率を用いて主成分分析を行った。その結果、作品の3群(第1部：1巻「桐壷」～33巻「藤裏葉」、第2部：34巻「若菜上」～44巻「竹河」、第3部(いわゆる「宇治十帖」)：45巻「橋姫」～54巻「夢浮橋」)に対応したグループが形成され、しかも第1部から2部、3部に移行するにしたがって使用率が安定していくことが示された。これに基づき、村上は従来の「宇治十帖」他作者説ばかりでなく、同一の作者の文体が時間推移とともに変化していったと考えることも可能であることを示唆した。

これは、母集団と想定するテキスト集合の文体の均一性に対する考え方の違いが異なる推論を導くことを示している。すなわち、母集団の文体が常に均一であることを前提とすれば、標本としての『源氏物語』テキストの文体が異なるとき、その母集団が単一ではないと推論することになる。すなわち、『源氏物語』作者複数説である。反対に、母集団の文体が異なることもあり得るという前提に立てば、標本としての『源氏物語』テキストの文体が異なっても、母集団自体が異なると直ちに推論することはない。つまり作者の文体も異なると考えることになる。これが村上(前掲書)の基底にある考え方である。このように、個人の文体の経年変化に対する洞察は、計量文体学における筆者推定においてきわめて重要なのである。

2. 文体指標

計量文体学における筆者推定は、特定の書き手の文体特徴を捉える計量可能な指標が存在するという仮定の上に成り立っている(Kenny 前掲書, 1)。そのような指標をここでは「文体指標」と呼ぶことにする。文体指標はあたかもテキストの指紋(fingerprint)である(Baayen et al. 2002, 1；Mikros & Argiri 2007, 29)。しかし、指紋とは異なり、あらゆるテキストに用いることができる絶対的な指標は存在しない(Biber et al. 1998, 234)。文体指標は言語によっても異なり得るし、対象とする書き手によっても異なると考えられている(金 1994, 317; 金 1997, 358; 金・村上 2003, 9)。そのため、特定の分析にどのような指標を用いるかが重要となる(Argamon et al. 2007, 35-36; Biber et al. 1998, 234; 金 1997, 357; 金 2002, 15)。

しかし、文体指標の要件はある。先行研究が挙げるのは主に次の3点である。使用頻度が高く、かつ書き手がその使用を意識しないこと(Argamon et al. 2005, 1)、テキストの内容の影響を受けないこと(Argamon et al. 2007, 15-16; Golcher 2007, 1 ; Santiani 2004, 1; 金 2000, 278-9)、書き手において安定性を持っていること(金 1997, 357; 金 2000, 278-9; 金 2002, 15)である。安定性を持つとは、テキストの中で一定の出現率を維持し、テキストの規模に左右されないことを意味するものと考えられる。

日本語のテキストを対象とし、計量文体学の手法を適用した研究を概観しても、文体指標の上記要件の重要性が確認できる。主な研究例のうち分析に使用された文体指標のみを列挙すれば以下のとおりである。助詞・助動詞の相対頻度(鈴木・景浦 2006)、「接続語句」・「助詞相当句」の相対頻度(村田 2000)、読点を打つ位置(金 1994、筒井 2010a, 筒井 2010b)、助詞の使用率(金 1997；金 2000)などである。特に助詞は他の品詞に比べて使用率が高く、かつ文章の内容への依存の度合いが低いため(金・村上 前掲書, 13-14)、助詞を文体指標として用いた書き手の判別は比較的短いテキストにおいても高い精度が得られることが報告されている(金 1997; 金 2000)。

3. 芥川龍之介の文体の経年変化

特定の書き手の文体の経年変化を計量的に検証した研究として金(2009)がある[3]。金(上掲書)は特定の指標を用いて特定個人の作品の執筆時期を推定できるかに関心を持ち、芥川作品の場合、助詞を用いれば一定の精度でその執筆時期を推定できることを示した。その前段階として、1000文字以上の規模を持つ芥川作品250編を対象とし、助詞38種に読点を加えた計39項目の相対頻度を用いて主成分分析を行った。その結果、芥川の作品の場合、執筆時期の推移とともに格助詞「が」の相対頻度は低下し、係助詞「は」の相対頻度は上昇する傾向が確認された。

筒井(2011)はこれを受け、芥川作品における助詞「が」・「は」の執筆時期による変化が本当に経年変化であるのかを検証した。特に、作品の類別という要因が執筆時期と当該助詞使用率の変化という2つの要因の間に介在しているのではないかという仮説を立て、3要因間の関係をそれぞれ検証した。その結果、類別という要素の影響は極めて限定的なもので、芥川作品においては助詞「が」・「は」の経年変化は否定できないという結論に至った。つまり、芥川龍之介の作品の場合、助詞「が」・「は」を文体指標とする限り、時間の推移とともにその文体が変化しているということである。

4. 本研究の目的

Biber(1988)は、テキスト間の多様性は多次元的なものであるためテキスト間の比較も多次元的に行うべきであると主張した(Biber 上掲書, 9-20)。相対的な比較の上に成り立つ計量文体学における「文体」についても同様のことが当てはまるであろう。つまり、1つの指標のみによる文体の把握が充分ではないこともあり得るということである[4]。その意味において、助詞使用率のみ

---

3) 特定個人の文体の経年変化を検証したものとして廖(2008)を挙げることができるが、計量的な手法によるものではない。また、対象作品も極めて限定的であるため、文体の違いが果たして経年変化なのか作品による差異なのか判別できない。

4) 明示的ではないが、安本・本田(1981)にもこのような思考があると思われる。安本・本田(上掲

をもって芥川作品の文体の経年変化を捉えようとした筒井(前掲書)は限界を持つ。さらに、変化の様相を明らかにすることに主眼が置かれていなかったという点も筒井(前掲書)の限界である。

本研究は、筒井(前掲書)にあるこの2つの限界を乗り越えようとするものである。すなわち、筒井(前掲書)とは異なる複数の文体指標を用いて筒井(前掲書)の知見を検証するとともに、芥川の文体の変化の様相を明らかにしようとするのである。逆に言えば、文体というものが一般的に多面的なものであり、芥川作品の文体に本当に経年変化があるならば、異なる文体指標、異なる手法を用いてもそれを検証できるはずである。また、複数の文体指標を用いて文体を把握することにより、その変化をより具体的に捉えることが可能になると考えられる。以上より、本稿は筒井(前掲書)とは異なる複数の文体指標、および異なる手法を用いて芥川作品の文体の経年変化を検証するとともに、文体の経年変化が認められる場合、その変化はいかなるものかを探ることを目的とする。

Ⅲ. 方法

1. 対象データ

「青空文庫」からダウンロードできる芥川龍之介の作品のうち、次の対象除外項目に該当するものを除いた結果、分析対象データとして最終的に147編を得た。

除外基準
- 翻訳作品
- 戯曲

書)は作家100名の文体を分類するのに12の項目を用いた。具体的には、直喩、声喩、色彩語、会話文の量、句点、読点、漢字、名詞、人格語、会話文、名詞の長さ、動詞の長さである。川崎(1967)の手法についても、文体を多面的にとらえるという点で同様のことを指摘できる。

・ 擬古文で書かれているもの
・ 形態素の総数(述べ語数)が1000以下のもの
・ 執筆時期(初出年月)を年月の単位で確定できないもの、および執筆期間が 1年以上に渡るもの

さらに、上記対象データを次の要領で精製・形態素解析した。基本は、各作品を地の文のみにするということである。

精製作業
・作品情報、ルビ等、本文以外の情報を削除
・会話、詩歌等の部分を削除

形態素解析
対象データの形態素解析(MeCab0.98を使用)

また、関口(1999, 317)に従い、芥川作品を以下の4期に分けた[5]。

芥川の執筆時期の4区分

第1期： 1913年　—1919年3月
第2期： 1919年4月—1920年12月
第3期： 1921年1月—1924年12月
第4期： 1925年1月—1927年7月

その結果、分析対象とする芥川作品147編の執筆時期別分布は次のようになった。

分析対象作品の執筆時期別分布

第1期： 36編
第2期： 24編

5) 関口は時期区分の明確な基準を示していない。しかし、ここでの目的は芥川における「執筆時期の推移」という要素を便宜的にカテゴリカル変数(名義尺度)にすることだけが目的であるので、方法論上の問題とはならない。

第3期： 48編
第4期： 39編

2. 変数

変数として次の5種の文体指標を選定した。すなわち、文長平均、文長標準偏差、Guiraud値、名詞比率(以下、N比率)、MVRである。これらの計測単位はすべて形態素である。変数選定の第1の基準は、「II.2. 文体指標」で述べた文体指標の要件、すなわち書き手がその使用を意識しないこと、内容の影響を受けないこと、安定性を持っていることの3つを満たすことである。第2の基準は、文体の特徴を捉えるのに効果的なことが先行研究によって示されていることである。例えば、筒井(前掲書)で用いられた助詞使用率は文体の相違を明らかにするには有効であるが(金 1997； 金 2000)、文体の様相を明らかにするにはさほど効果的であるとは考えられない。そのため、このような指標は本稿では変数として用いない。

上記5種の文体指標は大きく3つに分類できる。すなわち、文長に関するもの、語彙の多様性に関するもの、品詞比率に関するものである。文長、語彙、品詞比率はすべてテキスト全体に渡るもの、換言すれば、あるテキストにおける特定の文の長さ、特定の語彙、特定の品詞の使用を意識するものではない。そのため、これら3種の範疇は上記文体指標の第1の要件を満たしていると考えられる。

文長平均と文長標準偏差は文長、すなわち文の長さに関わるものである。文長平均はテキスト、ここでは芥川の個々の作品におけるすべての文の長さの平均(mean)である。同じく、文長標準偏差は芥川の個々の作品におけるすべての文の長さの標準偏差(standard deviation)であり、文の長さのばらつきの度合いを表す。文体指標としての文長は古くから用いられ、金・村上(前掲書,11)によれば、文長を文体指標として用いた最初の研究はSherman(1888)[6]であると言われている。日本語における文体分析に関しても、文長を文体指標

6) Sherman, L.A., Some observations upon the sentence-length in written prose : University of Nebraska Studies 1, 1888, pp.119-130.

として用いることはつとに波多野(1953)によって提唱されている。文長を文体指標として用いた実際の研究事例としては、樺島・寿岳(1965)、川崎(前掲書)、新井(2001)、鈴木・影浦(2008)などがある[7]。また、文体指標としての文長標準偏差は同じ波多野(前掲書)によって提唱されている。実際の使用例としては、樺島・寿岳(前掲書)、鈴木・影浦(前掲書)などがある[8]。

Guiraud値(Guiraud's Index)は、異なり語数を述べ語数の正の平方根で除したものとして定義され、語彙の多様性を示す指標の1つである。語彙の多様性については数多くの指標が提案されているが(金・村上前掲書, 20-23)、そのうちTTR(Type/Token Ration)はテキストのサイズに影響されやすいのに対し、Guiraud値はテキストのサイズの影響を受けにくく安定しているとされる(石川 2008, 77)。

N比率、MVRはともに品詞比率に関わるものである。N比率は非自立語を除く全品詞に対する名詞の比率、MVRはMVR=100M÷V(M：形容詞・形容動詞・副詞・連体詞、V：動詞)と定義される(樺島・寿岳前掲書, 25-31)[9]。樺島・寿岳(前掲書)によれば、N比率は文章がどの程度要約的であるかを示す。つまり、要約的な文章はN比率が高く、描写的な文章はN比率が低い。また、MVRにかんしては、M(形容詞・形容動詞・副詞・連体詞)が「ありさま」を表すことが多いのに対し、V(動詞)は「動き」を表す語と考えられることから、Vに対するMの比率、つまりMVRは文章が「ありさま」的か「動き」的かを表す。すなわち、M＞Vであれば「ありさま」＞「動き」、反対にM＜Vであれば「ありさま」＜「動き」と解釈できるわけである(樺島・寿岳前掲書, 32)。実際の適用例としては、樺島・寿岳(前掲書)がN比率、MVRとその他の指標[10]を用いて文学作品の文体の相違を数値化して説明している。

---

7) ただし文長を測る単位は同一ではない。樺島・寿岳(1965)は文節数、川崎(1967)、新井(2001)は文字数、鈴木・影浦(2008)は形態素数である。

8) 樺島・寿岳(1965)においては標準偏差の2乗である分散を用いている。

9) 樺島・寿岳(1965)では名詞比率の算出式の分母は自立語のみとしている、つまり助詞、助動詞は除いている。

10) その他の指標とは、「指示詞比率」、「字音語比率」、「文長」、「引用文の比率」、「接続詞を持つ文の比率」、「現在止めの文の比率」、「表情語比率」、「色彩語比率」である。

3. 主成分分析

上記5種の指標によって得られたデータに対して主成分分析(PCA: Principal Component Analysis)を行う。主成分分析とは、「相互に関係を持つ多変量のデータ集合の分散をできるだけ保持しつつ、変数がなす次元を縮約すること」(Jolliffe 2002, 1)により、データ集合に潜む情報の理解を容易にする多変量解析の1種である。主成分分析は計量文体学において多変量データに基づく分析の標準的な手法として定着している(Holmes 1998, 114)。本稿では、上記5種の変数間の相関係数行列(correlation matrix)に対して主成分分析を行い、芥川作品147編の分布の状況を見ることとする。相関係数行列に対して主成分分析を行うのは、すべての変数を標準化することにより、変数の計測単位の違いが分析結果に影響を与えるのを排除するためである(Jolliffe 前掲書, 22)。

Ⅳ. 結果および考察

分析対象作品147編における5種の変数の数値からなるデータの相関係数行列に対して主成分分析を行った。

まず、変数間の相関係数行列は以下の通りであった。

表1 変数間の相関係数行列

	文長平均	文長標準偏差	Guirald値	MVR	N比率
文長平均	1.000				
文長標準偏差	0.851	1.000			
Guirald値	0.357	0.300	1.000		
MVR	0.103	0.048	0.125	1.000	
N比率	-0.183	-0.105	0.045	-0.006	1.000

表1より、5種の変数のうち文長平均と文長標準偏差の相関係数が大きいことが分かる。これは、作品全体の文の長さが平均的に長くなれば、その分、

文の長さのばらつきの度合いも大きくなるということを意味する。その他の変数の組の相関係数に関して注目すべきものはない。

次に、主成分分析による各主成分の固有値、寄与率、累積寄与率は次の通りであった[11]。

表2　各主成分の固有値と寄与率

主成分(PC)	固有値	寄与率	累積寄与率
1	2.102	0.420	0.420
2	1.066	0.213	0.633
3	0.967	0.193	0.827
4	0.724	0.145	0.972
5	0.141	0.028	1.000

主成分(PC：principal component)とは、実際の変数の持つ情報を数学的手続きによって算出した合成値である。合成された主成分は、データ群のばらつきの大きい方向の軸に第1、第2…の順で相当している。つまり、データ群に関する説明力がこの順序で大きい。主成分は、理論的には変数と同数、合成されることになるが、もとの変数が持つ情報[12]を縮約することが主成分分析の主目的であるため、説明力が一定の基準以下の主成分は無視することになる。主成分同士は数学的に互いに直交であるため、各主成分は対象データ群の特徴のそれぞれ異なる側面を捉えていると考えなければならない。固有値(eigenvalue)は、データ群が持つ全情報量のうち各主成分が持つ情報の量を示す。相関係数行列から主成分分析を行った場合は元の変数が持つ情報量は1であるため、固有値が1未満の主成分は充分な価値を持たないものとして無視する。よって、ここでは第1、第2主成分のみに注目することとなる。寄与率(contribution ratio)とは、各主成分が表わす情報量を比率換算したものである。累積寄与率(cumulative contribution ratio)とは、第1主成分から当該主成分

11) 主成分分析に関するここでの説明は、主に慶応SFCデータ分析教育グループ編(2008)、中村(2009)に依っている。

12) データの分散(散らばり)をここでは情報と考えている(中村 2009, 101)。

までの寄与率の累積値である。よって、ここでは第1、第2主成分が全情報量のそれぞれ42.0%、21.3%を集約しており、第2主成分までで全情報量の63.3%までを表わしていることになる。

次に、第1、第2主成分と各変数との間の主成分負荷量を以下に示す。

表3　第1、第2主成分と各変数との間の主成分負荷量

	PC1	PC2
文長平均	-0.937	0.109
文長標準偏差	-0.906	0.107
Guirald値	-0.563	-0.455
MVR	-0.192	-0.544
N比率	0.220	-0.734

主成分負荷量(principal component loading)とは、主成分と元の変数との相関係数であり、当該主成分が何を表しているかを解釈するための手掛かりとなる。表3を見ると、第1主成分では文長平均と文長標準偏差の主成分負荷量の絶対値が大きいことから、第1主成分の負の方向は文の冗長さを表すと考えられる。換言すれば、第1主成分の正の方向は文章の端正さを表すと考えられる[13]。また、第2主成分ではN比率とMVRの絶対値が大きいことから第2主成分の負の方向は文章が描写的で「ありさま」よりも「動き」に比重が置かれていると解釈できる。正負の方向を反対にして述べれば、第2主成分の正の方向は文章がより直截的であると解釈できる[14]。以上より、第1、第2主成分は文章の端正さと直截性をそれぞれ表すものと解釈できる。

次に、個々の芥川作品における第1主成分と第2主成分の主成分得点を求め、第1主成分と第2主成分よりなる座標上に配置したものが下図である。主成分得点(principal component score)とは主成分を軸とする座標上の座標値であるため、これによって主成分からなる座標上の位置が定まる。

13) ここでは、文が短く、かつ文の長さが均一である場合、文章の印象がきびきびしたものになる、つまり端正になると考えている。

14) 文章が要約的、かつ「ありさま」>「動き」であることを「直截的」であると解釈した。

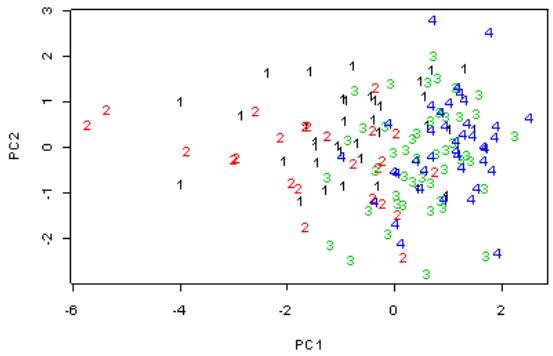

図1. 主成分得点による芥川作品147編の散布図

図の1～4の数字はそれぞれ芥川の執筆時期を表す。図1より2つのことが読み取れる。1つは、執筆時期別作品群の右方向への移動である。すなわち、1期の作品群は図のほぼ中央に分布し、2期の作品群は一旦やや左側に移動している。そして、3期、4期の順で作品群は図右側に次第に移動している。もう1つの点は執筆時期別作品群の密集度の違いである。すなわち、1期と2期の作品群は分布のばらつきが大きいのに対し、3期、4期の順で作品群の分布は特に横の軸(第1主成分)に関して次第に密集している。もちろん、3期、4期ともに各群の周縁部に各群から離れて位置している作品もあるが、それらは少数にすぎず例外とも見なし得る。また、第1主成分は時期別芥川作品の弁別に有効であるが、第2主成分は弁別にあまり有効ではないことも分かる。

以上のことを芥川作品の文体の経年変化という点に照らしてより具体的に述べれば次のようになる。第1に、執筆時期別作品群の図1における右(正)方向への移動という点は、芥川作品の文体が執筆時期の推移とともに次第に端正さを増してきたことを示す。なぜなら、第1主成分の正方向は端正さを表すからである。第2に、執筆時期別作品群の第1主成分に関する密集度の違いは芥川作品が執筆時期の推移とともに文体的に安定してきたことを示す。換言すれば、芥川が初期には端正さの点でばらつきのある文体の作品を書いてい

たが、時間の推移とともにその文体を確立し、晩年には端正さの点でほぼ均一な文体で作品を執筆していたと解釈できる15)。

この点を実際の作品にあたって確認してみよう。2つ目の点は相当数の作品を群として比較しなければならないためこの場で確認することは難しいが、1つ目の点は典型的な作品2点の比較をもってある程度確認することができる。第1主成分の主成分得点に関して対極に位置するのが「妖婆」と「浅草公園」である。「妖婆」は第2期の作品で第1主成分の主成分得点が－5.7033であり、「浅草公園」は第4期の作品で第1主成分の主成分得点が2.5373である。図1の散布図では前者が図のもっとも左にある「2」、後者が図のもっとも右にある「4」でそれぞれ示され、文章の端正さの点では分析対象147編のなかで両極にある。つまり、前者は文章が冗長、後者は端正であると予測される。以下に示すのは両作品の各冒頭である16)。

あなたは私の申し上げる事を御信じにならないかも知れません。いや、きっと嘘だと御思いなさるでしょう。昔なら知らず、これから私の申し上げる事は、大正の昭代にあった事なのです。しかも御同様住み慣れている、この東京にあった事なのです。外へ出れば電車や自働車が走っている。内へはいればしっきりなく電話のベルが鳴っている。新聞を見れば同盟罷工や婦人運動の報道が出ている。――そう云う今日、この大都会の一隅でポオやホフマンの小説にでもありそうな、気味の悪い事件が起ったと云う事は、いくら私が事実だと申した所で、御信じになれないのは御尤もです。が、その東京の町々の燈火が、幾百万あるにしても、日没と共に蔽いかかる夜をことごとく焼き払って、昼に返す訣には行きますまい。ちょうどそれと同じように、無線電信や飛行機がいかに自然を征服したと云っても、その自然の奥に潜んでいる神秘な世界の地図までも、引く事が出来たと云う次第ではありません。それならどうして、この文明の日光に照らされた東京にも、平常は夢の中にのみ跳梁する精霊たちの秘密な力が、時と場合とでアウエルバッハの窖のような不思議を現じないと云えましょう。時と場合どころではありません。私に云わせれば、あなたの御注意次第で、驚くべき超自然的な現象は、まるで夜咲く花のように、始終我々の周囲にも出没去来しているのです。 「妖婆」

15) 図1において執筆時期が遅くなるほど横方向(PC1, すなわち文章の端正さ)のばらつきが小さくなっていることからこのように解釈できる。

16) いずれも「青空文庫」に拠った。ただし、漢字の振り仮名に関する付加情報は除去した。

1

浅草の仁王門の中に吊った、火のともらない大提灯。提灯は次第に上へあがり、雑沓した仲店を見渡すようになる。ただし大提灯の下部だけは消え失せない。門の前に飛びかう無数の鳩。

2

雷門から縦に見た仲店。正面にはるかに仁王門が見える。樹木は皆枯れ木ばかり。

3

仲店の片側。外套を着た男が一人、十二三歳の少年と一しょにぶらぶら仲店を歩いている。少年は父親の手を離れ、時々玩具屋の前に立ち止まったりする。父親は勿論こう云う少年を時々叱ったりしないことはない。が、稀には彼自身も少年のいることを忘れたように帽子屋の飾り窓などを眺めている。

「浅草公園」

文章が与える印象の解釈は主観的なものであるため、上記2編の作品の文体に関して本稿執筆者が断定的に論じることは控えたい。しかし、主成分分析の結果から予測できた当該2編の文体の違いをある程度確認できたのではないだろうか。

以上、主成分分析の結果より次のことが明らかになった。芥川の文体は時間の推移とともに文章が端正になっていった。また、時間の推移とともに端正さの点で安定していった。ただし、文章の直截性という点では時間推移による変化は確認できない。

本稿における課題は、筒井(前掲書)とは異なる文体指標と手法で芥川作品の文体の経年変化を検証すること、および変化の様相を明らかにすることにあった。実際には、筒井(前掲書)とは異なる5種の文体指標、および主成分分析を用いて、芥川作品の文体の経年変化のありさまを部分的ながら明らかにすることによって、その文体の変化を確認することができた。その点において、筒井(前掲書)における知見、すなわち芥川作品の文体に経年変化が見られるという結論は本研究によっても支持されると結論付けられる。

V. 結論

本稿は、芥川龍之介作品の文体の経年変化を筒井(前掲書)とは異なる変数と手法で検証すること、および変化の様相を探ることを目的とした。変数としては文長平均、文長標準偏差、Guiraud値、名詞比率、MVRの5種、手法としては主成分分析(PCA)を用いた。その結果、第1主成分と第2主成分がなす座標上において執筆時期別作品群が執筆時期の推移により第1主成分の正方向へ移動していること、さらに執筆時期別作品群の分布が後期になるほど第1主成分に関して密集化していることが明らかになった。よって、芥川作品の文体は執筆時期が遅くなるほどより端正になっていくとともに、文章の端正さの点でより安定していったと解釈できる。このことより、芥川龍之介作品の本文の経年変化は筒井(前掲書)とは異なる変数と手法をもちいても確認できると結論付けることができる。

【参考文献】

新井皓士「文長分布の対数正規分布性に関する一考察 : 芥川と太宰を事例として」『一橋論叢』125(3), pp.205-223, 2001

石川慎一郎『英語コーパスと言語教育－データとしてのテクスト』, 大修館書店, 2008

樺島忠夫・寿岳章子『文体の科学』, 綜芸舎, 1965

川崎宏「文学作品の因子分析的研究1」『長崎大学教養部紀要 人文科学』7, 1967, pp.1-38

慶応SFCデータ分析教育グループ編『データ分析入門』第13章「主成分分析」, 2008, http://www.keio-up.co.jp/kup/pdf/15240.pdf 検索日2011.06.01

金明哲「読点のうち方と文章の分類」『計量国語学』第19巻第7号, 1994, pp.317-330

金明哲「助詞の分布に基づいて日記の書き手の識別」『計量国語学』第20巻第8号, 1997, pp.357-367

金明哲「自然言語における統計手法を用いた情報処理」『統計数理』第48号第2号, 2000, pp.271-287

金明哲「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』Vol.11 No.2, 2002, pp.15-23

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』第36

号第2号, 2009, pp.89-103
金明哲・村上征勝「文章の統計分析とは」金明哲・村上征勝・永田昌明・大津起夫・山西健司『言語と心理の統計　ことばと行動の確立モデルによる分析』岩波書店, 2003, pp.1-57
鈴木崇史・影浦峡「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), 2008, pp.113-122
関口安義『芥川龍之介とその時代』, 筑摩書房, 1999
筒井昭博「読点の現れを捉えるモデル-主観性・客観性ベクトルの競合という観点から-」『日本研究』第43号. 韓国外国語大学日本研究所, 2010a, pp.387-411
筒井昭博「読点は書き手の違いを現すか-文構造に基づく読点の分布による分析-」『日語日文学研究』第73集. 韓国日語日文学会, 2010b, pp.477-499
筒井昭博「文体の経年変化の検証-芥川龍之介作品における助詞「が」「は」を例として -」『日語日文学研究』第77集. 韓国日語日文学会, 2011, pp.335-356
中村永友『Rで学ぶデータサイエンス 2 多次元データ解析法』, 共立出版, 2009
波多野完治『文章心理学』, 新潮社, 1953
村上征勝「『源氏物語』の言葉を分析する」『統計数理』第44巻第1号, 1996, pp.127-131
安本美典・本田正久(1981)『因子分析法』, 培風館, 1981
廖育卿「森鴎外訳『即興詩人』における文体表現 : ドイツ三部作との比較及び再検討」『熊本大学社会文化研究』6, 2008, pp.365-379
Argamon, Shlomo, Casey Whitelaw, Paul Chase, Sushant Dhawle, Sobhan R. Hota, Navendu Garg and Shlomo Levitan, Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology* 58(6), 2007, pp.802–822.
Argamon, Shlomo and Moshe Koppel, The Rest of the Story: Finding Meaning in Stylistic Variation. Argamon, Shlomo, Kevin Burns and Shlomo Dubbov. eds. *The Structure of Style: Algorithmic Approaches to Understanding Manner and Meaning*, New York: Springer-Verlag, 2010, pp. 79-126..
Baayen, Harald, Hans van Halteren and Fiona Tweedie, Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution, 1996
Baayen, Harald, Hans van Halteren, Anneke Neijt and Fiona Tweedie, An experiment in authorship attribution. *Proceedings of the 6th International Conference on the Statistical Analysis of Textual Data (JADT 2002), 2002*
Biber, Douglas, *Variation across speech and writing*, Cambridge: Cambridge University Press, 1998
Holmes, David I., The Evolution of Stylometry in Humanities. *Literary and Linguistic Computing* Vol.13, No.3, 1998, pp.111-117

Jolliffe, I.T., *Principal Component Analysis 2nd Edition*, NewYork: Springer-Verlag, 2002

Kenny, Anthony, *A Stylometric Study of the New Testament*, NewYork: Oxford University Press USA, 1986

Mikros, George and Eleni Argiri, Investigating topic influence in authorship attribution. *Proceedings of the International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection*, 2007, pp. 29–35

Santini, Marina, A Shallow Approach to Syntactic Feature Extraction for Genre Classification, *Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics*, 2004, http://www.cs.bham.ac.uk/~mgl/cluk/papers/santini.pdf (検索日: 2010.12.03)

Stubbs, Michel, *Words and Phrases: Corpus Studies of Lexical Semantics*. 南出康世・石川慎一郎監訳『コーパス語彙意味論―語から句へ―』研究社, 2006

Wales, Katie, *A Dictionary of Stylistics 2nd Edition*, New York: Longman Publishing Group, 2001

【分析資料】

青空文庫 芥川龍之介の作品　計147編

http://www.aozora.gr.jp/index_pages/person879.html#sakuhin_list_1(検索日: 2010.10.1~2010.10.26)

▣ 접 수 일 : 2011년 06월 30일
심사개시 : 2011년 07월 18일
심사완료 : 2011년 08월 12일
게재결정 : 2011년 12월 06일

〈要旨〉

아쿠타가와류노스케 문체의 경년(經年)변화
PCA를 사용한 계량적 검증

본 연구 목적은 아쿠타가와류노스케 작품들간에 시간 추이에 인한 문체 변화가 있는지를 쯔쯔이(2011)와 상이한 변수를 바탕으로 검증하는 데에 있다. 사용한 변수는 문장(文長) 평균값, 문장(文長) 표준편차, Guiraud 치, 명사 비율, MVR(동사 빈도에 대한 형용사, 부사, 연체사[連体詞] 빈도의 비율)의 5 가지이며, 이들 간의 상관행열(correlation matrix)에 대해 PCA(Principal Component Analysis)를 실시했다. 그 결과 아쿠타가와의 문체는 시간 흐름에 따라 간결해 졌음과 동시에 문장(文章)의 간결성에 관해 안정성이 커진 것으로 확인되었다. 그러므로 아쿠타가와류노스케 작품의 문체에 경년변화가 있음이 쯔쯔이(2011)와 다른 변수를 대상으로 해도 검증됨으로 결론지을 수 있다.

Chronological Style Shift in *Akutagawa Ryunosuke*
− Stylometric Analysis with the Application of PCA −

This study explores whether it is possible to demonstrate *Akutagawa Ryunosuke*'s stylistic shift over time, with the use of variables different from those that are used in Tsutsui's previous study. The variables used in this investigation are five style-indicating features: average sentence length, standard deviation of sentence length, Guiraud' Index, the frequency ratio of the occurrence of noun, and MVR, which is a ratio of the frequency of the occurrence of adjectives, adverbs, and *rentaishi* to that of verbs. PCA (Principal Component Analysis) is conducted on the correlation matrix among the data indicated by those five features. The result of PCA illustrates *Akutagawa*'s stylistic shift over time, suggesting the establishment of his writing style towards the last stage of his career. It is concluded that *Akutagawa*'s stylistic change over time is confirmed by variables other than those that are used in Tsutsui's previous study.