Vol.2015-CH-107 No.7
2015/8/9

文学作品の計量分析：その方法と歴史

土山玄†

近年ディジタルヒューマニティーズが注目されるにともなって、計量文献学の手法を用いた文学作品の計量分析が盛んになっている。文章を計量的に分析するという点で、計量文献学は最近の学問のように思われるが、その研究の萌芽は 19 世紀の西欧にまで遡る。そこで、本論文では内外の計量文献学の達成の歴史とその研究方法を概観する。文章の計量分析では主に著者の識別、及び著作の成立年代あるいは成立順序の推定が研究目的となり、語の長さや語の頻度などの項目が分析に用いられてきた。本論文ではこれらの分析項目の特徴についても検討を加える。

Quantitative Analysis of Literary Text: The History and Methodology

GEN TSUCHIYAMA†

As the use of digitized information for humanities has become increasingly popular, the number of quantitative analyses of literature, called stylometry, has been increasing. Although many people consider stylometry as a new academic field, in fact, the pioneering figures in this field date back to the 19th century. Therefore, this paper reviews the history and methodology of stylometry. In stylometric research, the main purposes of investigation are authorship attribution and the estimation of the chronology of literary works, and researchers have proposed various textual measurements, such as word length, word frequency, and word richness. This paper discusses some characteristics of the measurements.

1. はじめに

文学作品などの文献を対象とし、計量的な分析を行う学問の 1 つに計量文献学がある。計量文献学は、著者の文体に関わる習慣的特徴を統計的に解析することを通じて、著者の識別、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[1]である。古典文学などの歴史的な文献には、著者や成立時期について議論の対象となっている場合が少なからず存在する。このような問題を扱うとき、従来は記述内容の検討や成立に関する歴史的事実の考証という観点から研究をおこなうのが主たる方法であったが、計量文献学はこのような方法とは一線を画し、文章から得られる計量的なデータを収集し、これを分析することによって当該文献の文体的特徴を把握して、結論を導き出すのである。

計量文献学において文体を取り扱うことの背景には、文体に著者の個性が現れるという了解があり、これは経験的に首肯しうるものと考えられる。「文は人なり (Le style c'est l'homme.) 」というフランスの箴言もこのことを指摘している。これは「ビュフォンの針」で有名な 18 世紀フランスの自然学者 Georges-Louis Leclerc, Comte de Buffon (1707-1788) のアカデミー入会演説のなかで述べた”Le style c'est l'homme même.” (文は人間そのものである)という一節に由来すると言われている[2]。日本の代表的な文体論の研究者である小林英夫の『小林英夫著作集第七巻 文体論の建設』(1975) [2]によれば、Buffon 自身の意図は別にあったとされるが、「文章というものはその作者の個性の反映である」と一般に解釈され、今日まで人口に膾炙している。また、P. Guiraud の『文体論』[3]においても、Buffon の箴言は当時から「文体が人間の本性そのものを表現している」と解釈され、読者は文体から著者に関する何かしらの情報、あるいは何かしらの印象を感受すると理解される。

ただし、文体という概念は多様であり、学問分野によって理解する内容は異なる。計量文献学においては、文体とは計数可能な記述形式のことであり、その内容は文字や語の頻度、語や文の長さの平均値などの文章を構成する量的な要素である。そしてそれら形式的な要素に、著者または文献の個性が反映されると考えるのである。そこでは語の意味や記述内容は考慮されず、したがって研究の対象とはならない。この点において人文学的な文献研究とは立場を異にしている。計量文献学は、統計的な解析を経た定量的なデータをもって、著者間に存在する文体の相違、あるいは特定の著者における文体的特徴の経時的な変化などを明らかにしようとするのである。

2. 計量文献学の嚆矢

2.1 著者の識別

文献を対象に統計学の手法を用いてアプローチをするという点で、計量文献学は最近の学問のようにも思われるが、計量文献学に関連した研究の歴史は長く、19 世紀にまでさかのぼる。

文章を計量的に分析するという本格的な意識の萌芽は、著名な 19 世紀イギリスの数学者 Augustus de Morgan (1806-1871) の発言に認められる。De Morgan は「2 人の

† 同志社大学 研究開発推進機構
Organization for Research Initiatives and Development, Doshisha University

人間が同じ主題について記述した２つの文章よりも、1 人の人間が異なる主題について記述した２つの文章の方がより類似度が高い、ということが明らかになることが期待される」といった見解や「ヘロドトスの著作のひとつに用いられている膨大な単語とすべての文字数を数え上げ、総文字数を総語数で割るとその本における単語の平均文字数が求められる。これと同じことをヘロドトスの別の本に対して行い、これによって得られた２つの値は非常に近いものになることが期待される。」という指摘をしていたことが、De Morgan の妻である Sophia Elizabeth De Morgan の著書に記されている (p.216) [4]。

この De Morgan の見解に着想を得た地球物理学者の T.C. Mendenhall が 1887 年に 19 世紀イギリスを代表する 3 人の文章家である Charles Dickens (1812-1870)、William Makepeace Thackeray (1811-1863)、そして John Stuart Mill (1806-1873) の著作を対象とし、計量分析によって著者の識別が可能であることを明らかにしている[5]。De Morgan が 1 語の平均文字数を想定していたのに対し、T.C. Mendenhall は文字数別の単語の出現頻度を集計した度数分布を分析に用いた。T.C. Mendenhall が用いた度数分布は、金属工学において無機物質の組成を明らかにするスペクトル写真になぞらえワードスペクトル (word spectrum) と称されている。なお、近年では計量文献学分野においては、ワードスペクトルは語の長さの分布 (word-length distribution) と呼ばれることが多い。

次に、T.C. Mendenhall はこのワードスペクトルを用いて、イギリスの劇作家である William Shakespeare (1564-1616) の詩劇に分析を加えている[6]。Shakespeare の作品には Shakespeare 自身によって書かれていないとされる説が古くからあり、これは Shakespeare 別人説などと称されている。この別人説の中に、哲学者である Francis Bacon (1561-1626) が当時の圧政抗議の社会風刺のために、Shakespeare という架空の人物の名を借りて数々の詩劇を書いたとする説がある。この説の真偽を検証するために、T.C. Mendenhall は Shakespeare と Bacon の著作を対象に、それぞれのワードスペクトルを示した。その結果、Shakespeare は 4 文字の単語を最も多く使用するのに対して、Bacon は 3 文字の単語を最も多く使用するという文体の習慣的特徴の違いを明らかにし、Shakespeare と Bacon の同一人物説を否定した[a]。この T.C. Mendenhall の計量的な研究は、日本国内において刊行された文献にも引用され、その影響の大きさがうかがえる。

2.2 作品の成立順序の推定

T.C. Mendenhall がワードスペクトルを用いて Shakespeare の詩劇を分析するよりも早く、L. Campbell は 1867 年にプラトンの 30 余りある『対話篇』の執筆年代を推定するために、語の出現頻度を用いて計量的な分析を行っている[7]。

プラトンの執筆時期は 50 年から 60 年に及ぶと考えられているが、『法律』がプラトンの最後の著作であることをアリストテレスが言及していることを除き、『対話篇』の執筆順序は不明であった。プラトンの思想には発展がみられることから、プラトンの思想を体系的に理解するためには、著作の執筆順序を明らかにすることがきわめて重要であった。Campbell (1867) では、"Lexicon Platonicum"[8]を用いて、プラトンの後期の『対話篇』とされる『ティマイオス』『クリティアス』『法律』に共通に用いられていながら、他の『対話篇』においてはほとんど用いられない語彙を調査し、これらの語彙の生起状況の調査することで、『ティマイオス』『クリティアス』『法律』に加え、『ソピステス』『政治家』『ピレボス』が後期の『対話篇』もまた後期の執筆であると指摘した。この業績により、L. Campbell は計量文献学的研究の先駆者と位置づけられる。

また、L. Campbell とは別に W. Dittenberge もまた 1881 年に『対話編』の執筆順序を研究しており、このときはペアとなる同義語の出現率に注目して L. Campbell と概ね同様の結果が得ている[10]。

日本国内では上述の T.C. Mendenhall のワードスペクトルが計量文献学の発展に繋がる初期の研究として知られているが、それ以前に Campbell (1867) [7] 及び Dittenberger (1881) [10] が公表されており、計量的な観点による文体研究として先駆的意義を有するものと言えよう。なお、プラトンの『対話篇』の執筆年代の推定に関する研究は W. Lutoslawski によって著された”The origin and growth of Plato’s logic; with an account of Plato’s style and of the chronology of his writings”[11]において網羅的にまとめられており、これは統計学者の G.U. Yule にも多大な影響を与えた業績として知られる。

3. 計量文献学の方法

文章の計量分析では、すでに述べたように、語の長さや語の頻度が用いられる。さらに語の長さに類似した分析項目として、文の長さという項目もある。ここではこれらの分析項目を用いた研究事例を概観する。また、上に紹介した統計学者の G.U. Yule は語彙の豊富さ指標を提案していることから、その代表的な指標についても示す。

3.1 語の長さ

語の長さを用いて、著者の識別を行った代表的な研究として、Brinegar (1963) [12]がある。内容は以下の通りであ

a Mendenhall (1901) [6]については、その分析方法に問題があることが Williams (1975) [9]において報告されている。これはすなわち、Shakespeare の文章は詩劇、つまり韻文 (verse) であるのに対し、Bacon の文章は散文 (prose) であるということである。そこで、Williams (1975)では、16 世紀のイギリスの詩人・政治家・軍人であり、散文および韻文のどちらも書き残している Philip Sidney (1554-1586) の著作を分析している。分析の結果、Sidney の散文は Bacon の散文と類似した分布を示し、反対に Sidney の韻文は Shakespeare の韻文と類似した分布を示したことが報告されている。したがって、Mendenhall (1901)において示された単語の長さの分布の相違は、執筆者の相違ではなく、ジャンルの相違である可能性が指摘されている。

る。

1861 年に、Quintus Curtius Snodgrass と名乗る人物の投書によるアメリカの南北戦争を批判する 10 通の手紙がニューオリンズ・デイリー・クレセント誌に掲載された。この 10 通の手紙は QCS Letters と呼ばれ、投稿者すなわち書き手は『トム・ソーヤの冒険』の作者として知られる Mark Twain (1835-1910) であるという見解があった。そこで、Brinegar (1963)は QCS Letters が掲載される以前の Twain の文章、同時期の文章、以後の文章を収集し、語の長さを集計し、それぞれカイ二乗検定および t 検定を行い、この見解を否定した。

O'Donnell (1966) [13]は、"The O'Rudy" (1903) を対象に分析を加えている。"The O'Rudy" は 33 の章 (chapter) から構成される文学作品である。この作品は作者である Stephen Crane (1871-1900) が執筆中に死亡しており Robert Barr (1849-1912) によって書き継がれたとされるが、どの章で作者が交代したかが不詳であった。そこで、語の長さの分布などの 18 項目について、2 群判別分析を行っている分析の結果、24 章までが Stephen Crane による執筆であり、25 章で作者が交代し、26 章から 33 章までが Robert Barr による執筆であると結論づけている。この問題によく似た問題に日本の古典文学作品である『源氏物語』最終 10 巻の宇治十帖において論じられる複数作者説がある。これについては後述する。

Radday (1970) では、旧約聖書の一書である『イザヤ書』について検討を加えている。『イザヤ書』は 66 章によって構成されるが、イザヤの真筆が確実とされるのは最初の 12 章までであり、その他は 2 人から 4 人の複数の人物によって執筆されたと考えられている。そこで、音節および音素を単位とした語の長さの平均値、エントロピーなどについてカイ二乗検定を行うことで、イザヤが執筆したとみられる 12 章と類似した特徴を有する章と、別人の執筆である可能性が高い章を指摘している[14]。なお、Radday (1970) では写本研究ではなく、あくまで計量的な研究であることから、校定された現行版の旧約聖書を分析に使用しており、写本間の相違については意図的に考慮していないという研究上の特徴を持つ。

Forsyth et al. (1999) では、長期間にわたって散逸していて 1583 年にヴェネツィアにおいて発見された、古代ローマの哲学者であるマルクス・トゥッリウス・キケロ (106 B.C.-43 B.C.) の長期間にわたって散逸していた著書である『慰め (Consolatio) 』が、キケロのオリジナルであるのか、あるいは偽書であるのか検討するために語の長さについて判別分析を行った。その結果、この書物の発見者とされる Carlo Sigonio(1520-1584)の偽作であることが明らかになった[15]。

このように、De Morgan によって考案され、T.C. Mendenhall によって使用された語の長さの分布は近年にいたるまで、著者の識別に有効な分析項目として、用いられている。

3.2 文の長さ

語の長さと類似した分析項目に文 (sentence) の長さがある。先にふれたように、W. Lutoslawski の著書[11]の影響を受けたイギリスの統計学者の G.U. Yule は 1 文に含まれる語数、すなわち文の長さ (sentence-length) が著者の特徴を示す 1 つの指標になり得ることを提案している[16]。これは、Yule (1939) において、Francis Bacon、Samuel Taylor Coleridge (1772-1834) 、Charles Lamb (1775-1834) 、Thomas Babington Macaulay (1800-1859) の 4 人の文の長さを調査し、その結果、文の長さの分布は著者間では相違し、同一著者では安定することを明らかにしたことによる。また、この結果を踏まえて、G.U. Yule は『キリストに倣いて (De imitatione Christi) 』の著者を推定した。『キリストに倣いて』はカトリック教徒にとって必読の書とされた文献であるが、その著者については古くから議論されてきた。主たる著者の候補は聖アウグスティノ修道院副院長の Thomas á Kempis (1380-1471) とパリ大学総長である Jean Charlier de Gerson (1363-1429) であったことから、Yule (1939) では、両候補者の文の長さの平均値・中央値・四分位を求め、その結果から Thomas á Kempis が『キリストに倣いて』の著者であると結論づけている。

Morton (1965)は、ヘロドトス、トゥキディデス、リシアス、デモステネス、イソクラテスらの古代ギリシャ散文を対象に分析を試みている[17]。それぞれの著作から 100 文以上を抜き出し、1 文あたりの語数の分布を求め、実証的な研究を行い、著者それぞれの文体に、長期にわたって変わらない習慣的特徴が認められることを示した。これによって、著者別の文体を分析する際に、文の長さの分布が有効であることを論じている。

また、上述の『イザヤ書』における著者の識別を行っている Raddy (1970) [14]においても文の長さは分析項目の 1 つとして用いられている。

3.3 語の頻度

Ellegård (1962) は文中において語彙的意味ではなく文法的機能をあらわす機能語 (function words) を分析に用いて、1769 年から 1772 年にかけてイギリスの新聞に掲載された Junius Letters と呼ばれる投稿記事を対象とし、自信が考案した、対象となる語の Junius Letters における相対頻度を Junius Letters 以外の文献からサンプリングした 100 万語における相対頻度で割った値である distinctiveness ratio という指標を用いて執筆者の推定を行っている[18]。これらの記事はイギリス政府の政策を批判しており、Junius という偽名が使用されていたことから Junius Letters と称されており、著者は不明であった。しかし、分析によって 100 名以上いる Junius の候補者から

投稿者を推定している。

Mosteller and Wallace (1964) は『ザ・フェデラリスト (The Federalist Papers) 』と称される 1787 年 10 月から 1788 年 8 月までの間にニューヨークの新聞紙に連載されたアメリカ合衆国憲法の批准を推進するために書かれた 85 編の連作論文の著者不明の論文について分析を行っている。これら 85 編の論文はすべて James Madison (1751-1836) 、 John Jay (1745-1829) 、 Alexander Hamilton (1755-1804) の 3 人によって著されたことが判明しているが、上記の 3 人のうち誰が執筆したのか不明な論文が 12 編あり、これらの著者不明の論文について著者が明確である文章から選出した論文の内容とは無関係に出現する語彙である"on"、"upon"、"while"、"whilst"などの 30 語について判別分析やベイズの定理を用いて検討を加えている。分析の結果、Hamilton に”while”、Madison に”whilst”が頻出していることを示した。これをもって、上述の語彙の多くは文法的機能を表す機能語であることから、機能語が著者の識別に有効であることを論じた[19]。

先にふれた Morton (1965) では、語の頻度についても分析を行っている[17]。新約聖書の『パウロ書簡』を対象にギリシャ語で”and”、”but”、”in”や人称代名詞を意味する共通語彙 (common words) と称される 4 語の頻度を集計し、カイ二乗検定を行っている。『パウロ書簡』はパウロが執筆したとされるが、検定の結果、『パウロ書簡』の一緒である「ガラテヤ信徒への手紙」の著者をパウロであると仮定すると、最低でも他に 6 人の著者が存在すると結論づけている。

また、Holmes and Forsyth (1995) は Mosteller and Wallace (1964) と同様に、『ザ・フェデラリスト』を対象とし、高頻度機能語 49 語について主成分分析を行い著者について議論の余地がある論説の著者についてのより詳細な検討を加えている[20]。

近年の研究としては、Binongo (2003) において、アメリカのファンタジー小説の”Oz”の最終巻である第 15 巻の著者の識別がある。”Oz”の作者は Lyman Frank Baum (1856-1919) であるが、第 15 巻が発刊されたのは 1921 年であり、第 15 巻には児童書作家である Ruth Plumly Thompson (1891-1976) 執筆したとの疑いが論じられており、このことから、第 15 巻は Baum の遺稿であるか、あるいは Thompson の執筆によるものか不明であった。そこで Binongo (2003) では、機能語を採り上げ、主成分分析を行い第 15 巻の作者は Thompson であると結論づけている[21]。

語の頻度は語の長さと同様に、文章の計量分析においては一般的な分析項目である。欧米諸語を対象とした分析において、前置詞・接続詞・助動詞・冠詞などを指す機能語は著者の識別に有効であることが実証的に明らかにされている。

3.4 語彙の豊富さ

語彙の豊富さ、いわゆる語彙力は著者によって相違するものと予想されることから、これを用いて著者を識別しようとする立場もある。そのため、語彙の豊富さを表す指標は多数提案されている。最も基礎的な指標は TTR と称されるもので、対象となる文章の延べ語数を N、異なり語数を V としたとき、V/N で求まる値である。

次に、上述の G.U. Yule も K 特性値 (characteristic K) という指標を提案している[22]。延べ語数が N であるときに i 回出現した語数を V_i としたとき、K 特性値は

$$K = 10^4 \left(\sum i^2 V_i - N \right) \Big/ N^2$$

という数式で求められる。

K 特性値の他に、著名な指標として、Guiraud (1954) [23] によって提案された

$$R = V/\sqrt{N}\ 、$$

Herdan (1960) [24]の

$$C = \log V / \log N\ 、$$

Tuldava (1977) [25]の

$$LN = \left(1 - V^2\right) \Big/ \left(V^2 \log N\right)$$

などがある。

また、Grieve (2007) によれば、*LN* が著者の識別の精度が良いことが報告されている[26]。

3.5 他の分析項目

語や文の長さの分布、語の頻度、語彙の豊富さ指標の他に品詞の分布がたびたび分析に使用される。Antosch (1969) では形容詞に対する動詞の比率を求め、この比率は文章のジャンルに依存することを明らかにした[27]。なお、民話ではこの比率が高く、科学系の記事では低いことを報告している。

また、句読点 (punctuation) の頻度についても研究されている。上述の O’Donnell (1966) [13]において、語の長さの他に句読点の頻度についても分析が加えられている。また、Chaski (2001) でも 5 人の女性が記述した文章を体操に、句読点の頻度を集計し、カイ二乗検定を行っている。その結果、句読点の頻度は著者の識別に有効であることを明らかにしている[28]。

4. 日本語を対象とした研究事例

日本における文章の計量分析における初期の研究は 1935 年に波多野完治によって著された『文章心理学』である[29]。波多野完治が論ずるところの文章心理学は文章の特徴と著者の性格の結びつきを考慮することから、厳密に言うと計量文献学とは性質のことなる学問であるが、研究方法では類似点が多々認められる。波多野 (1935) では、

谷崎潤一郎の『金と銀』および志賀直哉の『雨蛙』を分析対象として採り上げ、前者は動詞の割合が高く、後者は名詞の割合が高いことを指摘している。これに加え、文の長さの平均値を計測しており、前者は 49.2 字、後者は 32.1 字であることから文体的特徴が著者によって相違することを明らかにした。なお、上述の小林英夫は波多野 (1935) の影響を受けたことを記している[2]。

また、日本語の品詞の構成比率を採り上げた研究として、1956 年に発表された大野晋による研究が著名である。『万葉集』『枕草子』『徒然草』『方丈記』『紫式部日記』『土佐日記』『讃岐典侍日記』『竹取物語』『源氏物語』を対象とし、名詞の出現率が減少するにつれて、動詞、形容詞、形容動詞の出現率が増加することを明らかにした[30]。なお、これは「大野の法則」と称される。また、「大野の法則」は水谷 (1965) において数理的に定式化されている[31]。なお、大野 (1956) は上述の Antosch (1969) [27]よりも早い。

次に、日本文における単語の頻度を対象とした文章の計量分析の初期の研究としては安本 (1957) [32]があげられる。欧米で古典文学である Shakespeare の詩劇が採り上げられたことと同様に、安本 (1957) においては平安時代に紫式部 (973-1014)によって著されたとされる『源氏物語』を分析対象としている。『源氏物語』を宇治十帖とその他の 44 巻に二分し、統計的な仮説検定を行っている。検定に用いた項目は名詞の使用度、用言の使用度、助詞の使用度、助動詞の使用度など 12 あり、これらの使用度は、各巻からランダムに 1000 字抽出し、その 1000 字における各項目の頻度によって求められている。従って、『源氏物語』の全文が分析の対象になっているのではない。

安本 (1957) もまた文章心理学に位置づけられる研究であり、検定の結果、宇治十帖の文体は作り物語的、用言的、緊密かつ連続的な構想による詳細な描写を特徴とし、一方、他の 44 巻の文体は歌物語的、体言的、飛躍的、断続的な構成による直感的描写を特徴とすると考察されている。それゆえ、宇治十帖の作者は他 44 巻の作者と同一人物であるとは言い難い、と結論づけている。

これに加え、安本 (1977) においては、安本 (1957) と同様の標本抽出によって得られた 12 の変数を用いて因子分析を行い、再び宇治十帖の複数作者説に検討を加えている。分析の結果、これまでと同様に宇治十帖の文体が他の 44 巻の文体と相違すると指摘するが、ここでは作者が相違するとは結論づけていない[33]。

他方、新井 (1997) は、各巻の中央部から各巻の長さに応じて標本を抽出し、五十音図の頭子音行別頻度や母音列別頻度に対して統計的検定を行っている[34]。検定の結果、宇治十帖の作者が他の諸巻の作者と別人であるとは考えられないと述べている。

上述の『源氏物語』を対象とした 3 つの計量的な研究はそれぞれに意義を有するが、物語の全文を用いて分析をおこなってはいない。これに対して、村上・今西 (1999)は『源氏物語』の全文を対象とし、多変量解析の手法を用いて本格的な計量的研究をおこなった。この研究では機能語である助動詞の出現率を用い、数量化Ⅲ類により『源氏物語』の成立巻序の推定を行っている[35]。なお、『源氏物語』を対象に主成分分析などの多変量解析を行った研究として土山・村上 (2014) [36]があり、宇治十帖とそれに先立つ匂宮三帖において論じられる複数作者説について、助動詞の出現率の他に助詞などの出現率や語の長さの分布を用いて検討を加えている。その結果、匂宮三帖および宇治十帖の作者が他の諸巻と異なる可能性の低いことを論じている。

また、現代文を対象とした研究として、金 (1994)があげられる。金 (1994) では井上靖、三島由紀夫、中島敦の著作を対象とし、これまで用いられていた分析項目とは異なり、読点の前の文字の出現状況を調査し、これを分析に用いるという新しい視点から研究を行っており、読点の使用方法に著者の特徴があらわれることを明らかにした[37]。

また、金 (2002) ではこれまで研究対象としてきた職業作家の文章ではなく、一般人の日記文と作文を対象に判別分析を用いて書き手の識別を行っている。分析の結果、日記文の場合は助詞の bigram に、作文の場合は trigram に書き手の特徴があらわれていることを明らかにした[38]。なお、助詞の bigram とは、文中の助詞以外の品詞を無視したときの、助詞の連続して出現するパターンのことである。同様に trigram は 3 連続のパターンである。

松浦・金田 (2000) では、文字レベルの n-gram 分布を用いている。これは国木田独歩や菊池寛なでおの近代の 8 人の作家の文章を対象に、各 n-gram の確率分布を求め、それらの確率分布間の疎遠度を計算し、文体の距離を測ることで著者推定に成功した[39]。

これらの分析は著者の識別を目的としたものだが、芥川龍之介の著作を対象に執筆年代の推定を行った実証的な研究としては金 (2009) がある。助詞の出現率の経時的な変化を明らかにすることにより、高い精度での執筆年代の推定が可能であることを指摘している[40]。

5. 今後の展望

冒頭に記したように、本論文では、文学作品を含む文献を対象とした計量的な研究事例を概観した。計量文献学の主たる研究目的は著者の識別および執筆順序の推定であるが、著者の識別に比べ著作の執筆順序や執筆時期の推定に関する研究は十分に展開しているとは言えない。この原因として、執筆順序の推定が問題になるのは古典文学作品に多いこと、しかしながら古典文学作品には年代設定の基準となる外部資料が乏しいことが主な原因であると考えられる。

日本の文献を見渡すと、平安時代に成立した物語である『宇津保物語』や『源氏物語』を始め、成立順序が議論の

対象となる古典文学作品は数多く存在する。

例えば、『源氏物語』に関しては、『源氏物語』の成立過程に関する見解の 1 つとして、登場人物の出現状況の調査に基づく客観的なデータから、第 1 巻から第 33 巻までの 33 巻には「紫上系」と称される 17 巻と「玉鬘系」と称される 16 巻の 2 系統が内在しており、それぞれが別個に成立したとする説が論じられている[41]。これは、初出が「紫上系」となる登場人物は「玉鬘系」においても登場するが、初出が「玉鬘系」である人物は「紫上系」に登場しないという事実に基づいている。

『源氏物語』のような古典文学作品の場合、成立過程について記述されている外部資料が乏しい。それ故、計量文献学の方法を用いて、文献内部から古典文学作品の執筆順序の推定を行うことは、成立過程を解明するための資料を提供できるだろう。

今後は著者の識別を目的とした研究がより展開されることとともに、著作の執筆順序の推定を目的とした研究が新たな発想に基づき発展されることを期待したい。

参考文献

1) 村上征勝: 文化を計る−文化計量学序説. 朝倉書店 (2002).
2) 小林英夫: 小林英夫著作集第 7 巻 文体論の建設. みすず書房 (1975).
3) Guiraud, P. (佐藤信夫 訳): 文体論−ことばのスタイル−. 白水社 (1954).
4) De Morgan, S. E.: Memoir of Augustus de Morgan, Longmans, Green, and Company (1882).
5) Mendenhall, T. C. : The characteristic curves of composition, Science, (214S), pp. 237-246 (1887).
6) Mendenhall, T. C.: A mechanical solution of a literary problem. Popular Science Monthly, Vol. 60, No. 2, pp. 97-105 (1901).
7) Campbell, L.: The Sophistes and Politicus of プラトン. Clarendon Press (1867).
8) Ast, F. :Lexicon プラトン nicum, 3 vols. Lipsiae (1835).
9) Williams, C. B.: Mendenhall's studies of word-length distribution in the works of Shakespeare and Bacon. Biometrika, Vol. 62, pp. 207-211 (1975).
10) Dittenberger, W.: Sprachliche Kriterien für die Chronologie der プラトン nischen Dialoge. Hermes, Vol. 16, No. 3, pp. 321-345 (1881).
11) Lutoslawski, W.: The Origin and Growth of プラトン's Logic, With an Account of プラトン's Style and of the Chronology of His Writings. Longmans, Green, and Company (1897).
12) Brinegar, C. S.: Mark Twain and the Quintus Curtius Snodgrass letters: a statistical test of authorship. Journal of the American Statistical Association, Vol. 58, pp. 85–96 (1963).
13) O'Donnell, B.: Stephen Crane's The O'Ruddy: A Problem In Authorship Discrimination. In Leed (ed.), The Computer and Literary Style. Kent, OH: Kent State University Press, pp. 107-115 (1966).
14) Radday, Y. T. Isaiah and the computer: A preliminary report. Computers and the Humanities, Vol. 5, No. 2, pp. 65-73 (1970).
15) Forsyth, R. S., Holmes, D. I., and Tse, E. K.: Cicero, Sigonio, and Burrows: investigating the authenticity of the "Consolatio". Literary and Linguistic Computing, Vol. 14, No. 3, pp. 375-400 (1999).
16) Yule, G. U.: On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship, Biometrika Vo. 30, No. 3-4, pp. 363-390 (1939).
17) Morton, A. Q.: The authorship of Greek prose, Journal of the Royal Statistical Society. A-128, pp. 169-233 (1965).
18) Ellegård, A.: A Statistics Method for Determining Authorship : The Junius Letter, 1769-1772, Gothenburg Studies in English, 13 Acta Universitaits (1962).
19) Mosteller, F. and Wallace, D.: Inference in an authorship problem. Journal of The American Statistical Association, Vol. 58, pp. 275–309 (1963).
20) Holmes, D. I. and Forsyth, R. S.: The federalist revisited : new direction in authorship attribution. Literary and Linguistic Computing, Vol. 10, No. 2, pp. 111-117 (1995).
21) Binongo, J. N. G.: Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. Chance, Vol. 16, No. 4, pp. 375-388 (2003).
22) Yule, G. U.:A statistical study of vocabulary, Univ. Press, Cambridge, England (1944).
23) Guiraud, P.: Les caractères statistiques du vocabulaire. Presses universitaires de France (1954).
24) Herdan, G.: Type-token mathematics, Mouton (1960).
25) Tuldava, J. : Quantitative Relations between the Size of the Text and the Size of Vocabulary, Journal of Linguistic Calculus, Vol.4, pp28-35 (1977).
26) Grieve, J. : Quantitative authorship attribution: An evaluation of techniques. Literary and linguistic computing, Vol.22, No.3, pp.251-270 (2007).
27) Antosch, F.: The diagnosis of literary style with the verb-adjective ratio, pp. 57-68, New York (1969).
28) Chaski, C. E.: Empirical evaluations of language-based author identification techniques, Forensic Linguistics, Vol.8, pp. 1-65 (2001).
29) 波多野完治.: 文章心理学, 三省堂 (1935).
30) 大野晋.: 基本語彙に関する二三の研究, 国語学, Vol. 23, pp. 34-46 (1956).
31) 水谷静夫.: 大野の語彙法則について, 計量国語学,Vol. 35, pp. 1-13 (1965).
32) 安本美典.: 宇治十帖の作者−文章心理学による作者推定−, 心理学評論, Vol. 2,No. 1, pp. 147-156 (1957).
33) 安本美典.: 現代の文体研究, 岩波講座『日本語』, Vol. 10, pp. 395-423, 岩波書店 (1977).
34) 新井皓士.: 源氏物語・宇治十帖の作者問題：一つの計量言語学的アプローチ, 一橋論叢, Vol. 117, No. 3, pp. 397-413 (1997).
35) 村上征勝, 今西祐一郎.: 源氏物語の助動詞の計量分析, 情報処理学会論文誌, Vol. 40, No. 3, pp. 774-782 (1999).
36) 土山玄, 村上征勝.: 『源氏物語』第三部の成立に関する計量的な考察, じんもんこん 2014 論文集, Vol. 2014, No. 3, pp. 213-220 (2014).
37) 金明哲.: 読点の打ち方と文章の分類, 計量国語学, Vol. 19, No. 7, pp. 317-330 (1994).
38) 金明哲.: 助詞の n-gram モデルに基づいた書き手の識別, 計量国語学, Vol. 23, No. 5, pp. 225-140 (2002).
39) 松浦司 , 金田康正。: n-gram の分布を利用した近代日本語文の著者推定, 計量国語学 Vol. 22, No. 6, pp. 225-238 (2000).
40) 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, Vol. 36,No. 2, pp. 89-103 (2009).
41) 武田宗俊.:源氏物語の研究, 岩波書店 (1954).