文学研究におけるデジタル・ヒューマニティーズの可能性
―文章心理学・計量文献学・マクロ分析―

杉浦 清人

1 デジタル・ヒューマニティーズと日本の統計的文体分析の歴史

近年の情報技術の目覚ましい発達に伴い、人文学にもデジタル技術を導入し、これまでの人文学を更新しようとする機運が高まっている。英語では Digital Humanities と呼ばれるそのような試みの日本語での表現としては、「デジタル人文学」「デジタル・ヒューマニティーズ」「人文情報学」などがあり、未だ訳が定まっているとはいえない状態であるが、ここではひとまずデジタル・ヒューマニティーズを使用しておく。

デジタル・ヒューマニティーズはその名のごとく人文学全体に関連するものだが、本稿ではその中で、特に文学研究の分野におけるデジタル技術の有力な使用法の一つである、文学テキストのコンピューターを用いた統計分析にまつわる問題を扱う。

昨今コンピューターの発達に伴い急激に進歩しているとはいえ、文学テキストの統計分析はけして近年始まったわけではない。むしろ世界のみならず日本においてもそれなりに長い歴史を持っている。金明哲は次のように書いている。

> 日本語の文章における早期の計量的研究としては , 波多野完治氏の『文章心理学』がある . 波多野完治氏の『文章心理学』の第 1 版は昭和 10 年に大日本図書で出版された . この本の新稿版（昭和 40 年）のなかの , 谷崎潤一郎氏と志賀直哉氏の作品における文の長さの比較に関する一節を紹介する .
>
> 「谷崎氏の文章は志賀氏の文章に比していちじるしく字数が多い . 平均すると前者は後者の倍ほどはないとにしても , それに近い長さをもっている .」
>
> このように , 当時の文章の計量分析は , 文の長さや品詞の使用率といった , その当時の技術で計量可能なデータの記述的な分析が主であった .
>
> 1960 年代にはいると , 安本美典氏が文章心理学のアプローチによる文章の計量分析に推測統計や因子分析などの手法を用いた研究を発表した . ほぼ同時期に樺島忠夫氏らは言語学のアプローチによる文体の計量分析を行っている . ただし , その時期はコンピュータが自然言語を自由に扱うことができなかったため , 文をサンプリングし , 目で確認しながら要素をカウントする方法に頼らざるを得なかった .
>
> 日本で , 本格的にコンピュータを用いて文章の計量分析を行ったのは , 村上征勝氏の日蓮遺文に関する研究である . 研究を始めた 1976 年は , コンピュータが日本語の漢字・カナを扱うことができなかったため , 分析に用いる日本語全文をローマ字表記に変換して用いたそうである .[1]（句読点ではなくカンマとドットを使用しているのはすべて原文ママ）

日本における文学テキストの統計的文体分析の歴史において、1935 年の波多野完治『文章心理学』が嚆矢であるが、波多野の用いた統計手法は未熟なものであり、技術の進歩に伴い、後の安本美典、樺島忠夫・寿岳章子、村上征勝、金明哲らの研究において統計手法が精緻化し発展していった、という認識が一般的であるといってよい。[2]

だから、ある意味では波多野の研究はすでに乗り越えられた過去のものである。しかし、本稿においてまず注目したいのは、日本の文学テキストの統計分析のはじまりに位置し、その統計手法において未熟であったために乗り越えられていった波多野の研究である。

この論文は、日本の文体分析の歴史と、最新のデジタル人文学の研究についての検討から、あるべき文学の統計分析の条件と可能性について考察を行うものである。

2 文章心理学とは何か

「文章心理学」とは心理学者である波多野完治が自らの研究の名前として使用したものである。[3] 波多野の文章心理学関係の著書は何冊もあるが、最も大きな影響を与えたのはその最初のものである『文章心理学』である。

国語学者の中村明は、文章心理学の画期性を次のように語っている。

> 文体論といえば、ひところ、数表とそこから大胆に結論を導く水際立った手並みがすぐ浮かんできたほど、波多野完治の文章心理学の影響力はすさまじいものであった。日本で文体論という学問が世に知られたのは、実に文章心理学の誕生によってであると言っても嘘ではない。[4]

一方、波多野に強く影響されて自身も文章心理学に取り組んだ安本美典は、文章心理学の登場の背景を次のように分析している。

> 新らしい時代思潮が、うちよせてくる。
>
> 心理学の分野ではヴントの心理学の批判として、ゲシタルト心理学があらわれる。言語や文章の研究の分野では、いっぽうでは、前代の時代思潮の批判者として、フォスラーやシュピッツァーのドイツ観念論学派があらわれる。また、いっぽうでは、バイイらがあらわれ、ソシュールの思想を発展させ、実証主義を刷新する。
>
> それに呼応して、わが国では、国語学者の時枝誠記氏、言語学者の小林英夫氏、心理学者の波多野完治氏などがあらわれ、言語あるいは文章の認識をふかめる。昭和の初期、おもに、昭和十年代のことである。[5]

すなわち、波多野の文章心理学は、確かに日本の文体論において画期的なものであったのだが、それはけして何もないところから出てきたわけではなく、当時の心理学、言語学の世界的な流れを受けて登場したのである。彼は孤立していたわけではなく、日本におい

ても時枝国語学で知られる時枝誠記や、フェルディナン・ド・ソシュール『一般言語学講義』の翻訳で有名な小林英夫など、同時期に波多野と同じ潮流に属すると見なすことができるようなすぐれた言葉の学者が登場していたのだった。

さらに、学問の分野だけではなく、分析の対象とした文学創作の世界にもまた、彼の文章心理学が歓迎される環境が用意されていたことを波多野は述懐している。

> 昭和のはじめから十年くらいにかけて、日本では文体に関するイノヴェーション（革新）が行われた。これは一部はプロレタリア文学が弾圧され、作家がもっと抽象的な方向へむかわざるをえなかったことが原因であったろうが、もっと大きな原因は、新興芸術派、新感覚派、および心理小説などの、文学における新しい領域の開拓が行われたことに求めなければならないだろう。（中略）『文章心理学』自体は、このような文壇における文体革命の要求に応じて書かれたものではなかった。それは、一人の先覚者のあたまに宿ったアイデアにもとづき、そのアイデアにしたがって無我夢中でやみの中をほり進んだ、わかい学者の無鉄砲の結果として生まれたものであった。しかし、それが世の中にむかえられ、わかい学者がその後、ほそぼそながら三十年にわたって研究をし、ほとんど独力でその研究をしつづける根気をもたせてくれたのは、日本の現代文学に内在する文体革命の要求にささえられたからにほかならない。[6]

『文章心理学』が出た時期は日本の小説家たちが新しい文体を模索していた時期であり、1934 年には谷崎潤一郎『文章読本』が出版され話題になるなど、世間的に文章への関心が高まっていたという事情もあった。

つまり当時、学問においても創作においても言語についての様々な新たな考え、試みが現れており、波多野の文章心理学は当時の心理学や言語学の世界的な潮流、そして日本の文学の潮流にも合致したものだったため影響力を持ったのである。

しかし、その中で波多野の文章心理学が際立ったのは、やはり統計分析によってであるということもまた、忘れてはならない。

安本は次のように書いている。

> 波多野氏の『文章心理学』が、それまでの修辞学のゆきづまりを破った時代的意義は大きい。また、心理学による理論的なうらづけは、波多野氏のうでのさえを思わせるし、こまかくゆきとどいている。
>
> しかし、私は、心理学による理論的な説明自体は、『文章心理学』を、『文章心理学』たらしめた本質的なものではなかったであろうと思うものである。すくなくとも、小林英夫氏の文体論や、国語国文学の分野の研究などに共通してとりいれられたのは、『文章心理学』の心理学的説明の部分ではなかった。
>
> たんてきにいえば、『文章心理学』の斬新さは、理論よりも、むしろ、実践にあったと思

われる。さらにいえば、現代作家の文章をとりあげて分析したその分析の方法にあったと思われる。[7]

波多野が現代作家の文章を分析した方法とは、具体的には以下のようなものである。

まず、波多野は谷崎と志賀という二人の著名な作家の文章を読んだときの印象が非常に異なっており、「日本文の典型としても、二つの理念類型（イデアルティプス）を形成していること、どちらもすぐれた風格をもちながら、しかも一つはけっして他に還元することができない、いわば素数のようなものであること」[8]から出発し、二人の書いた文章のごく一部を抜き出して、手作業で計量分析を行っている。分析に用いる要素は、文の長さや句読点や品詞の割合や構文の複雑さである。そして、志賀と谷崎の文章におけるそれらの要素の多寡が何を意味するかを分析した後に、総合的に次のような見解を導いている。

> 谷崎氏と志賀氏との文章の相違は、両氏が文章に課する役割の相違からきていると考えられる。谷崎氏においては事件または事物の叙述が、あくまでも言語を主体として、言語が主役となって語られている。これに反して、志賀氏においては事件あるいは事物はことばによって語られていても、このばあいことばはただ媒介をするだけで、主になるものは事物そのもの、事件そのものである。ことばは事件を暗示するにとどまり、すべて事件じたいが直接の形で表現される。[9]

波多野の分析のこの結論は実のところ、谷崎と志賀という特徴的かつ対照的な作家の文章の比較として、それほど独自性のあるものではない。当時は数字を使う分析が衝撃的に映ったとしても、意地悪く言えば読んだ印象を再確認し権威付けるために統計を使っているにすぎない。それでは統計分析が真に新しい知見を文学研究にもたらしていると言い切ることはできない。

ただ、先駆者としての功績は大きいし、単に文章を分析するだけではなく心理学やレトリック論を使って理論的に文体分析を検討していることや、『文章心理学』を世に出した後何十年も文章心理学の研究を続け、小説以外の新聞や広告などの文章についても研究を行うなど[10]、幅広い視野を持っていたこと、何より時代の要請に応え文学・文芸批評の領域と統計分析の領域を接続することに成功したのは高く評価できる。

中村明は『日本語文体論』において、波多野や波多野に影響された統計的な文体分析、すなわち客観的文体論の功罪を述べている。

客観的文体論の功績は、統計を用いることによって、評者の主観的な読解に依存する印象批評よりも客観的・科学的な分析を可能にし、文体研究を脱神秘化・脱名人芸化したことである。それまでの文体研究は、例えばレオ・シュピッツァーのような優れた批評家の素質を持つ人間にしかできないと思われていたのだが、統計分析ならば文学的才能がなくても可能であるし、別人によって検証可能でもある。

一方で以下のような客観的文体論の弊害をも挙げている。

まず、統計分析は文章を一定の単位に区切って集計する要素主義をとることになるが、それでは文章の有機性を正確に捉えることができず、具体的な作品の固有性を十分に掬い取れないという問題がある。

そして、計量的に把握できる部分が文体のなかでどの程度の割合を占めているのかもわからないにもかかわらず、あたかも数量化できる言語的性格だけが文体的特徴だという思いこみが生じてしまうといった、客体的な分析方法そのものの限界をどう認識するかという問題がある。

もう一つ、先行研究の受け継ぎに関する問題を、中村は次のように表現している。

> 言語調査の数量的な処理によって得たデータの解釈にあたって、波多野完治は心理学の知見を注ぎ、小林英夫は関連文献を精力的に参照し、それぞれ大胆な作品論・作家論を展開した。ところが、後継者たちの研究では、透徹した洞察力や豊かな想像力の要求される後半の過程がさほど重視されなくなり、極端な場合は、ひたすら文の長さを測り、品詞の出現率を調べる作業自体があたかも文体論であるかのような趣を呈するに至った。（中略）もしも言語調査を示して事足れりとする安易な文体論や、その結果を勝手に解釈して性急に結論を導く短絡的な文体論が、文体論の隆盛をもたらすようなことになるなら、文体論という学問にとってむしろ不幸なことと言わねばならないだろう。[11]

すなわち、優れた客観的な文体論には、文献の調査や作品・作家論といった人文的な研究と数量を用いた統計分析の両方が必要であるということである。もちろん両立させることは困難であり、この困難は文体分析に常につきまとっているといえよう。

3 波多野以降の統計的文体研究

ここまで波多野の研究について見てきたが、次に波多野以降の統計的文体分析の代表的な研究者である安本美典と村上征勝の研究から、日本の統計的文体分析がどのような発展の過程を辿ったかを見ることにする。

前述のように波多野に大きな影響を受けて文章心理学に取り組んだ安本は、『文章心理学入門』において波多野を踏まえつつも、さらに文章心理学を発展させている。

まず理論的には、安本は文章心理学の主な課題を、「(1) 文章の類型論を確立すること、(2) 文章の類型によって、その筆者の性格をあきらかにすること、(3) ある時代の文章によって、その時代の精神を確立すること[12]」としているように、波多野の文章心理学の心理学的・人文学的な志向を受け継いでいる。しかし、同時に安本は次のように、波多野とは明らかに違う立場をとっている。

> 波多野氏の立場は、その文章分析手法の合理主義的・実証主義的な色彩にもかかわらず、

根本には、意識といった精神的なものをふまえている。その意味では、時枝氏や小林氏の立場と共通するところがある。ゲシタルト心理学の影響をうけ、そこから、観念論がそそぎこんでいる。

ここで、私の立場から、意見をのべておこう。文章の研究において、意識は、排除されるべきではない。なぜなら、文章は、ことばの形をとった人間の意識を、紙の上に定着させたものともいえるから、しかし、文章が、人間の意識とか主観とかから生み出されたからといって、文章研究の出発点に主観をおかなければならないという理由はない。[13]

つまり、波多野は例えば谷崎と志賀の文章が対照的であるというような人間の印象による分類からスタートしてそれを統計によって裏付けているが、そうではなく、統計による分類を先行させるべきだということだ。

そのために安本はこの本の中で作品を分類する方法として因子分析を採用している。これは、分析対象の多くの特徴から、重要な少数の因子を探索し、データの分類や理解を可能にするものである。彼は 1900 年から 1954 年までに発表された現代作家 100 人の作品からそれぞれ一部を抜き出し、直喩の数・色彩語の数・センテンスの長さ・名詞の数・名詞の長さなどの 15 項目に対して調査を行い、そこから分析によって A・B・C の 3 つの因子を見出した。

A 因子は「用言型（和文型）―体言型（漢文型）」と呼ぶべき因子であり、A 因子得点が高い作品の文章は体言的（漢文的）で、逆に低い作品の文章は用言的（和文的）となる。同様に B 因子は修飾型―非修飾型の因子であり、C 因子は文章型―会話型の因子である。そして、これら 3 つの因子の計 8 通りの組み合わせによって作品と作家は分類される。

そして安本は分類を行った後に、それぞれの型に属する作家の共通項を見出し、分析している。その際用いる概念は、「グループ (abc)[用言―非修飾―文章型] に属する作品には、いずれも、文章そのもののリズムのもつ美しさや、味わいが、表面に出ているように思える[14]」といった具合に、一般的な文学史や印象批評と変わらないものである。

以上のように、安本の文章心理学は、波多野よりも統計的に発展した技術を用い、主観的なものを排除しようとしながらも、なお心理学的・文芸批評的な部分を色濃く残しているといえる。

対して、近年の日本の統計的文体分析の代表者である村上は自らの研究を計量文献学と呼んでおり、著者推定 (Authorship Attiribution) を主な目標とする。計量的な著者推定自体は安本が先駆けてはいるが、コンピューターを使えない時代のことであり、本格的にコンピューターを利用した著者推定は村上の登場を待たねばならなかった。

村上は日蓮遺文及び関連文献 50 編のデータベースを構築し、日蓮宗の開祖である鎌倉時代の仏教者日蓮の作とはされているが、著者に関して古くから疑問が出されていた 5 つの文献の著者推定を行った。これがコンピューターを積極的に利用した最初の日本語文献の研究とされている。

村上らの方法は以下のようなものであった。まず日蓮の真作とわかっているテキストのグループと、日蓮の贋作や日蓮門下の書いたテキストのグループに分けてそれぞれの文章の計量的特徴（言葉の出現率や、文長・単語長・品詞の出現率など文の構造に関する情報）を測る。そしてそれらのデータを用いて、似た性質の文献を集めて分類するクラスター分析を行い、日蓮の作とそうでない人物の作をうまく区別して分類することが出来た特徴をピックアップし、それらの特徴を用いて今度は真贋が疑わしいテキストを分類し、日蓮の作のグループに分類されるものは真作の可能性が高いとした。

結果として 5 編のうち 1 つは日蓮の書いたものであるである可能性が非常に高く、もう 1 つは日蓮作である可能性がかなり高いが、残り 3 編は贋作である可能性が高いという結論を導いている。[15]

著者推定はしばしば指紋鑑定に例えられる。それは、文章から書き手特有の癖を探し出し、誰が著者であるかを推定するものである。そこでは作者の心理は問題にならないし、文章がどのような性質を持っているかということは副次的な問題である。つまり文章心理学とは大きく異なっている。

すなわち、波多野から村上への日本の文体分析の変化はさらなる統計化・非心理学化・非文芸批評化と要約することができるが、その過程の中で統計的かつ心理学的・文学的である安本の研究は波多野と村上の中間に位置するということができる。

ここで着目したいのが、波多野の文章心理学にしても村上の計量文献学にしても、それらの性格の相違にもかかわらず、どちらも作家個人の文体分析を主としているということである。だが、文体分析が個人を対象とすることはけして当然のことではない。

4「文は人なり」

文体分析を理論付けるために頻繁に引かれる言葉が、「文は人なり」というものである。『文章心理学』の第一章「文章心理学とは何か」を波多野は次のように書き始めている。

> 「文は人なり」ということば。文章を語る人が口ぐせにいうことばである。しばらく、このことばを手がかりとして、すすむことにしよう。[16]

ここで登場する「文は人なり」という言葉は、元々は十八世紀フランスの著名な博物学者ジョルジュ＝ルイ・ルクレール・ド・ビュフォンがアカデミー・フランセーズの会員に就任した際の記念講演に現れた、”Le style est l'homme même” という言葉の日本語訳であり、日本に広めたのは高山樗牛であるとされている。文体論に関して波多野と影響を与え合った小林英夫も、『文体論の建設』の冒頭でこう書いている。

> 文体とは何か、という問を起すときに、すぐさま連想されるのは、「文は人なり」というアフォリズムである。[17]

この言葉は、「文章は書いた人の性格を表す」という意味に解釈するのが自然なように思える。高山もそのように使用し、一般にはそのような意味で流通した。しかし、これは誤解であり元々の意味は違うのだということを小林は指摘している。[18]

ただ、ビュッフォンの元々の意図はともかく、「文は人なり」という言葉は、文章は書き手の性格を表すという意味だと考えられて流通したのであり、波多野もその意味で使用しているし、本稿の以下においてもそちらの意味で使用することとする。

この「文は人なり」というテーゼは、簡潔に文章心理学が必要とする理論的前提を表している。文章が書き手の性格を表すからこそ文章の心理学が成立するのであり、そうでないのなら文章から作家の心理を分析することは不可能になる。村上もまた「文は人なり」というテーゼを引用している。[19]

しかし、この「文は人なり」というテーゼは、必ずしも常に正しいというわけではなく、限界を持っている。波多野は『文章心理学入門』において、近代日本文学の歴史を、自然主義以前・自然主義・自然主義以後の 3 つに分けている。[20] その区分の持つ意味は次のようなものである。

> 近代日本の文章の歴史は近代意識の表現意欲の歴史であるとも言い得られる。明治維新によって解放された精神は「普通文」で自己の表現を充分に行うことに困難を感じた。いわんや漢文や擬古文では自己を現わすことなど出来ないと感じたのである。
>
> 自己を語るとは自己の経験した事件を語ることおよび自己の経験の心理を語ることである。
>
> そのためには、自己を他から区別してはっきりと示すことのできるスタイルが確立されなければならぬ。またスタイルが確立されるほどに国語の表現性に幅が増大していなければならぬ。このような要求のうち、前者すなわち自己の経験した事件を語る方が自然主義によっていちおうの完成に達し、自己の経験した心理を語るほうが―それにおうじてスタイルの完成が―やはり自然主義を契機として白樺及び新思潮の時代にできあがった、と私は考えるのである。
>
> だからほんとうの意味で自由な柔軟な日本文のスタイルは自然主義の時代に礎石をすえられ、大正初期に確立された事件なのであった。武者小路、有島、里見、志賀、谷崎、芥川、久米等々とならべただけで私の意味がくみとっていただけるであろう。
>
> 文章は、それに並んで国語は、ここにいたってともかく近代化をなしとげたと考えるのである。[21]

作家が文章で自らを表現することは使用している言語、すなわちこの場合では近代日本語そのものに充分な表現力がなければ不可能なのだ。[22] そのため「文は人なり」が成立するのは近代的な文体が確立して以後に限られるのであって、それ以前の時代には通用しないということになる。だから明治日本の作家たちは江戸時代とは異なる、西洋のような近

代文学を書くことが可能な新しい日本語を作ろうとした。日本文学史においてこのことは言文一致運動と呼ばれる。そして、自然主義の作家たちによってひとまず言文一致体の文章が確立されたというのが、波多野のみならず日本文学史の一般的な見解である。逆に言えば、自然主義以前の作家たちは十分に自己を表現する能力のない日本語で作品を書いていたのである。それでは「文は人なり」が成り立たない。

ここで、近代以前の作家については文章心理学の前提が成り立たない、すなわち文章心理学による分析が不可能だという問題があることがわかった。もちろん近代以前の作家の文体分析が実際に不可能だというわけでも、無意味だというわけでもない。現に村上や安本らが近代以前の作品である紫式部『源氏物語』などに対して著者推定を行っているし、日本以外においても、近代以前の作家であるシェークスピアの著者推定は計量文献学の最も有名なテーマの一つである。[23] ここで指摘しているのは、「文は人なり」を前提とする文体分析の抱える理論的な問題である。

また、近代以後においても「文は人なり」というテーゼは破綻してしまうのだ。

20 世紀初頭からのモダニズムはそれまでの近代文学の伝統を覆した。20 世紀を代表する作品であるジェイムズ・ジョイス『ユリシーズ』は 18 の挿話で構成された作品であるが、ジョイスは、特に作品の後半において、それぞれの挿話ごとに異なる文体を採用している。

また、前衛的文学グループ「ウリポ」の中心人物であるレーモン・クノーの『文体練習』は 99 の断章から成っている作品だが、それらの断章はすべて同じ些細な出来事を扱っている。それを、「1・メモ　2・複式記述 3・控えめに 4・隠喩を用いて 5・遡行」[24] などというように、99 通りの異なった文体で書いたのがこの作品である。

このような作品全体を一つのまとまった文体として分析し、そこに作家と直結する文体を見いだすことは明らかにナンセンスである。作者は意図的に複数の文体を使い分け、単一の文体から構成される作品を拒否しているのだから。

思想でも同様の傾向が現れた。フランスの批評家・哲学者であるロラン・バルトは「作者の死」を説いた。[25] 彼によれば、前近代では物語は個人に帰属するものではなく、物語の語り部はいても、彼らはあくまで仲介者であり、作家としての天才性を評価されるようなことはなかった。「作者」というのは、西洋の近代化にあたり、個人の人格が尊重されるようになるとともに生み出された概念なのである。そして、この「作者」概念が現代にあってはすでに死を迎えているのだ。

バルトはこの作者の死とともに、「作品からテクストへ」という文学作品の捉え方の変化が起こったのであり、「読者の誕生は、「作者の死」によってあがなわなければならないのだ[26]」と書いている。文学作品の意味は作者によって決定されるわけではなく、文章を読者がどのように読むかということこそが重要だというわけである。

つまり、20 世紀において文学や哲学などに共通する変化が起こり、「文は人なり」という考えを否定するような考えが出てきたのである。とはいっても、文体は作家の性格を表現しているという考えがなくなったわけではもちろんない。「文は人なり」という考えが

完全に正しいわけでも完全に間違っているわけでもなく、真実はその間にある。つまり、文章は作者を含めた複数の要素によって重層的に決定されているのであり、重要なのは、対象のテクストを説明する要因として作者がどの程度の比重を占めているかを考慮し、場合によっては別の要因によって説明することである。また、文体のどのような側面がどのような要因によって影響されるのかを調査することである。

5 人でないなら何か

「文は人なり」という考えを否定したなら、当然文体は作家以外の要素によって決定されているということになる。作品に関わる要素は作家以外にいくつもあり、そのどれもが程度の差はあれ文体に影響するだろう。

例えば非常にスケールの大きい要素として国家がある。アメリカ文学、ドイツ文学などの国や言語ごとに、「ドイツ文学は重苦しい」「フランス文学はお洒落である」といったような大まかな傾向の違いがあると考えられており、それは文体と無関係ではないだろう。国ごとに文学を分けることは空間的な区分だが、時間的な区分、すなわち時代による文学の移り変わりも当然文体に影響する。

よりスケールの小さい要素として、ストーリーやキャラクター、テーマなどもある。例えばラブロマンスとゴシックホラーでは必然的に文体が異なるだろうし、方言を話すキャラクターが登場すれば標準語を話すキャラクターと違った文体を持つことになる。死をテーマとする作品は死に関係した言葉を多く含む。

このように、文体に関わる要素は多様に存在し、そのうちのどれに着目するかによって分析方法や分析結果が大きく変わると考えられる。

比較文学者のフランコ・モレッティは、「世界文学への試論」において、世界文学に取り組むために、伝統的な精読 (close reading) ではない、テクストから距離を取る読み方である遠読 (distant reading) を提唱した[27]。そして、「技巧、テーマ、文彩―あるいはジャンルやシステム[28]」といった「テクストよりずっと小さく、ずっと大きい単位に焦点を合わせる[29]」ことを推奨している。これは、特に世界文学を意識しない場合でも示唆的な提言である。

作家や作品を主体として考える文学観に対してモレッティは、ジャンルや文章技法やテーマといった、非個人的・非作品的なものに着目することで、見慣れたはずの文学も違った様相を見せることを説いている。

彼は論文「文学の進化について」で、ダーウィンの進化論を援用して文学史を考えることを主張した。そこでは進化論において個々の個体よりも種が重要なのと同じように、個々の文学作品ではなくそれが属するジャンルが重要な単位となる。

> 文学のダーウィニズム的歴史において、様々な文学形式は互いに戦い、その環境（コンテクスト）によって選択され、進化し、消滅してゆく。自然界の種と同じように。ここに、

> 文学批評が現代の形而上学的無価値を捨て去り、ある種の物質主義に立ち返った時に見られる素晴らしい情況を予想することができる。[30]

つまり複数のジャンルが誕生し、そのうちのあるものは人気を博して多くの作品が書かれ、あるものは人気が出ずに消滅し、あるものはその内容を変化させてさらに発展していく。このようなジャンルの変遷の歴史として文学史を理解するということである。

具体的には彼は、ヨーロッパ近代の文学史を次のように説明している。イギリスで近代小説が誕生した 18 世紀に様々な試行錯誤が行われて多くのジャンルが誕生したが、19 世紀には、18 世紀末に起こった産業革命と市民革命という社会の変化の影響によって多くのジャンルが自然淘汰され、最もうまく適応することが出来たビルドゥングスロマン（ここでのモレッティのビルドゥングスロマンという語の用法はかなり拡大解釈されており、近代リアリズム小説を指すと考えて良い）というジャンルが生き残り、長い間繁栄した。そして 20 世紀にはこの 19 世紀小説に対する批判が起こり、再び多くのジャンルが登場した。

では、このようなジャンル中心の文学史の理解は、作家や作品中心の理解と比較してどのようなメリットをもたらすだろうか。

まず､従来の少数の傑作や偉大な作者中心の文学史から脱することができることである。モレッティは書いている。

> これ（引用者注:ジャンル概念の重視）は見方の変更となるが､その結果は予測しがたいし､また予測などいくぶん怠慢である。しかし、確かなことがひとつある。このことが、文芸批評の基盤からその歴史学的位置づけまで再検証させることになるということだ。文芸批評はこの点では動揺し時代遅れなものとなっているが、偉大な作品あるいは偉大な個人が「事件」たり得るような、そんな事件史でなかったためしはない。歴史的大論争でさえ、結局、きわめて少数の作品や作者の解釈にばかり汲々とする。この手続きはジャンル概念を、周縁的な下位の機能にまで貶めるものである。[31]

今までに多くの文学作品が書かれているが、現在まで生き残って読まれている作品は極めて少数である。歴史的傑作や優れた作家中心の文学史では大多数の忘れられた作品はほとんど無視されたり、大衆に媚びる価値の低い作品とみなされ、傑作の凄さを引き立たせるために引き合いに出されたりするだけである。しかしそれら多数の平凡であったり下らない作品はもはや読まれていないのであり、本当に読む価値がないのか、つまらないのかは充分に検証されていない。本当はそこにまだ発見されていない文学の可能性があるかもしれないし、たとえ本当にそれらの作品自体は取るに足らないとしても、どんな傑作も多くの平凡な作品と同じ時代に、同じ環境の中で生まれたのであり、傑作を充分に理解するためにも多くの読まれざる作品、特に同じジャンルの作品の理解が必要なのである。

多くのもはや読まれない平凡な作品の集合であるジャンルの理解が必要である。しかし、実際に平凡な作品群に取り組もうとすれば、あまりにも読むのに時間がかかるという問題が持ち上がる。一つの作品を精読するには時間がかかるからこそ、通例文学研究者は扱う作品を少数に絞るのである。つまり精読では不可能なのだから、モレッティが提唱するところの遠読が必要になる。遠読という概念の意味するところは一つではなく、他人の精読の結果の分析を読むことで自らが精読しないで済ますことなどもあてはまるが、精読の代わりにコンピューターによる分析を行うこともまた遠読の一つのかたちである。

つまり、少数の優れた作家・作品の精読中心の、いわばミクロな文学研究とは異なる、多くの平凡な作品を含むジャンルの統計分析中心の、いわばマクロな文学研究にこれまでの文学研究とは違った知見を導き出す可能性があるのである。

6 マクロ分析の実践

では、ジャンルを主体としたマクロ分析の具体的な例を見てみよう。

モレッティは文学研究の中でグラフや地図や樹形図を積極的に利用することを目指した *Graphs,Maps,Trees*(Verso,2005) において、18 － 19 世紀のイギリス小説のジャンルごとの刊行点数をグラフにすることで、1760 年から 1790 年までは書簡体小説が、1790 年から 1815 年まではゴシック小説が、1815 年から 1840 年代までは歴史小説が流行していることを示し、そこから、ジャンルの流行は 25 年から 30 年の周期を持つのではないかという仮説を立てている。また、より多くのジャンルの変遷を見ると、10 年ほどで消えるジャンルと 25 年ほど続くジャンルの二つに分かれており、前者はジャコバン小説や反ジャコバン小説など政治的な状況に呼応したジャンルであり、それらはより短い流行の周期を持つとした。

そしてモレッティはこの 30 年周期のジャンルの変遷が、文学史を裏から規定する隠されたサイクルであると主張した。

このような視点を採用するとき、文学史から優れた作家や作品は見えなくなり、周期的に生まれ、消えていくジャンルこそが文学史の主人公となる。一般的な文学史に必ず登場する傑作も、今では忘れられ顧みられることのない駄作も、グラフにすれば同じ一つの作品にすぎない。もちろんこのような視点が単純に精読よりも優れているなどというわけではない。しかし違う見方をさせてくれるのは確かだ。

Stanford Literary Lab[32] でモレッティと共同研究を行うマシュー・L・ジョッカーズは、*Macroanalysis :Digital Methods & Literary History* において *Graphs,Maps,Trees* におけるモレッティのマクロ分析をさらに発展させ、統計を駆使しつつも、統計が何を可能にするかを注意深く検討している。彼はマクロ分析の目的を大規模な文脈化と位置づけ、それがもたらすことができる知見の例として以下のような事柄を挙げている。

・大きな文学的文脈の中での個々のテキスト、作家、ジャンルの歴史的な位置

・時代や地域や人口構成による文学生産の増大と減少
・時代や地域や人口構成ごとに採用される文学的パターンや語彙
・文体と文体の進化に影響を与える文化と社会の力
・個々の作者やテキストやジャンルを文学文化全体に束ね、あるいは束ねなかったりする文化的・歴史的・社会的連関
・文学的エスタブリッシュメントの嗜好と、それが大衆の嗜好に一致するか否か[33]

また、マクロ分析は以下のような問題に実践的なアプローチの方法を提供できるとしている。

・特定のジャンルに特有の文体があるか
・文体は国家によって決定されるか
・ある国のトレンドが他国に影響するか、するとしたらどのようにか
・サブジャンルがそれの属するより大きなジャンルをどの程度反映しているか
・文学的トレンドが歴史的出来事と関連しているか
・ある国や地域の生み出す文学が、その人口構成や時代や相対的な自由度や教育水準などの関数であるか
・文学は進化論的か
・成功した文学作品が流派や伝統を生じさせるか
・正典に含まれる作者と伝統的に傍流とされてきた作者に差があるか
・性別・民族・国籍といった要素が直接的に文学作品の文体や内容に影響するか[34]

以上のリストからもわかるように、マクロ分析は精読に対して、文学の歴史的・社会的側面を研究するのに大きなアドバンテージを持っている。ただ、以下では歴史的・社会的問題ではなく、「何が文体を決定するか」という問題を扱う。

統計的文体分析において著者推定が最も方法的に発達しているため、ジョッカーズは著者推定で使われている技術をベースにしてマクロ分析に応用している。著者推定は文章の傾向から作家に特徴的な文体を検出するものであるとはいえ、実際には文体は作家以外の複数の要素をも含む複合的な作用によって形成されている。そこで著者推定は作家という要素を把握するため、作家以外の要素をできるだけ排除できるように条件を統制するわけだが、ここでのマクロ分析では逆に作家以外の要素の文体に対する寄与を積極的に測定しようとすることとなる。

著者推定の方法は作家の文体を検知するために発達してきたのであるから、著者推定と同じ方法を採用するだけではもちろん作家の文体を測るだけになってしまう。そこでジョッカーズが採った方法は次のようなものである。著者推定は、分析対象となる複数のテキストが同じ作者の書いたものという同じ属性を持つグループに属するかどうかを、語

の使用頻度などの指標が充分近いか否かによって判定するものだといえる。その結果として高い精度で正しい判定が可能ならば（そして多くの実験によって実際可能であることが示されている）、作者という要素は強く文体を規定しているといえる。そこで同様に、同じジャンルの作品という同じグループに属するかを判定し、高い確率で判定が成功するなら、ジャンルが文体に大きな影響を与えていると考えることができるし、どのカテゴリによる判定の精度がより高いかを比べることで、ジャンルと作者のどちらが文体に大きく影響しているかといったことを比較できるというわけだ。

より具体的には、ジョッカーズは頻出語と句読点を指標として、106 作の、歴史小説・ゴシック小説・ビルドゥングスロマンといったジャンルに属する 19 世紀に英語で書かれた小説をそれぞれ 10 個に分割した、あわせて 1060 のテキストを分類した。その結果、ジャンルについては、ランダムよりもはるかに高い確率で正しいジャンルへの分類に成功するという結果が出た。ここから、ジャンルは文体にかなり寄与していると考えることができる。また、ジャンル判定の間違いはすべてのジャンルに均等に起こるわけではなく、あるジャンルと間違えられやすいジャンルや、そうでないジャンルがある。これは、ジャンル同士の関係として近いジャンルと遠いジャンルがあるためだと考えられる。どのジャンルとどのジャンルが近いかというのは文学史を書くために重要な情報になるだろう。

さらにジョッカーズは、作品（ひとつの作品を 10 個に分割して使用しているため、ある作品の一部分と同じ作品の別の部分が同じグループに分類されるかもテストされることになる）・作者・作者の性別・ジャンル・年代の 5 つのカテゴリについて、どれが強く文体を規定するかを調べる実験を行っている。その結果として、作品と作者は非常に高い精度での判定を可能にし、ジャンル・年代・作者の性別はそれらには劣るがある程度の精度で判定が可能であるという。同一作品の中のテキスト同士が言語的特徴を共有し同じグループに分類される確率が高いのは当然だが、作者もそれに劣らないほど、非常に強く文体を規定している。ジャンルや年代はそれらに比べて文体を規定する力が弱い。また、ジャンル自体が年代によって移り変わるするものであるため、ジャンルによる変化と年代による変化を区別するのはかなり難しい。

結局、ジョッカーズのこの分析の範囲では、ジャンルや年代は確かに文体に影響しており、文学において重要な役割を果たしているのだが、作者のほうが文体に対する影響が遥かに強く、より重要であるということになっている。では、「作者は存在しない」という考えは統計分析の結果間違っていて、やはり「文は人なり」が正しいということになるだろうか？　ジャンル分析は作家の文体分析より意味がないのだろうか？

必ずしもそのような結論にはならない。そもそも頻出語を用いて分類を行う手法自体が著者推定で高い精度を達成するために発展した方法であり、それを流用して分析を行っているため、作者が重要だという結果が出るのは当然といえば当然である。

つまり、他の分析方法なら作者の重要性が落ちることもあると考えられる。例えば、作品の話題やテーマという観点から分析するなら、同じ作者でも作品のテーマごとに大きく

異なる結果が出るだろうと推測することができる。しかし、例えばミステリというジャンルに属する作品群において、「殺人」「犯人」「被害者」という言葉が多く使われているという分析結果が得られたとして、これはあまりにも自明な結論であり、意味があるのか疑わしい。プロットやキャラクターとの関連の中での文体分析などを統計的に工夫して行うにしても同様に、精読によって簡単にわかることの確認になってしまうことも多いだろう。言い換えれば、統計分析が精読とは異なる独創的な見解を導き出したとして、それを人間が受け入れることが可能なのかという問題がある。著者が誰であるかを判断する以上の価値判断をしないという意味で文学研究に応用することを考えれば物足りなさがあるとはいえ、著者推定が方法として優れているのは、ジョッカーズが示唆しているように、それが人間が見逃すわずかな言葉の出現率を精査する、精読以上の精読だからである。マクロ分析には豊かな可能性があるが、それを充分に発揮するためには克服すべき困難があるのである。

7 終わりに

ここまで、いくつかの研究を検討してきたことから導き出される統計的文体分析についての見解は以下のようなものである。

情報技術の発展を受けて、統計的に洗練された著者推定が主流になっているが、それはあくまでも著者が誰かという問題に注力するため、文学研究に利用するためにはより多様な、そして文学的解釈に利用しやすい分析方法が望まれる。そのような分析の一例としてマクロ分析があり、これは社会学的・歴史学的な分析と特に相性がよい。ただ、マクロ分析にしても、グラフや地図を利用することで分析の見栄えをよくすることはできても、すでに知られていることを裏付けるだけの結果になる（もちろん、それはそれで意義があるが）危険性がある。文学についての独創的な新しい知見を統計分析によってもたらそうとすれば、著者推定とは違う方法を見出さなければならない。新しい切り口を見つけるために従来の文学研究や文学理論から発想を借りて統計分析に応用を試みるということを例えば『遠読』や Stanford Literary Lab のサイトに掲載されている論文が行っているが、アイディアが先行しているとやはり予想された結果の確認になりがちであったり、特定の場合ではいいが広範に利用しにくいものになってしまったりというような問題を抱えている。

著者推定を応用するか、社会学・歴史学との接続を試みるか、あるいは新しい分析手法の開発に挑戦するか、いずれにせよデジタル・ヒューマニティーズや文学の統計分析は広大な未踏の領域を残していることは疑いない。

註

1. 金明哲「データサイエンスで文章を科学する」、『文化情報学入門』村上征勝編、勉誠出版、2006 年、10 ページ。
2. 他の歴史記述としては奥総一郎「文学言語の計量化とその展望」(『シリーズ朝倉 < 言語の可能性 > 10 言語と文学』斉藤兆史編、朝倉書店、2005 年）や村上征勝「計量文献学―文献の新たな研究法」(『計量文献学の射程』村上征勝他著、勉誠出版、2016 年）などがある。
3. 波多野は、「文章心理学」という言葉そのものを作ったのは自身ではなかったと書いている（波多野完治『最近の文章心理学』大日本図書、1965 年、293-4 ページ)。
4. 中村明『日本語文体論』岩波書店、2016 年、128 ページ。
5. 安本美典『文章心理学入門』誠信書房、1965 年、34 ページ。
6. 『最近の文章心理学』、285-6 ページ。
7. 『文章心理学入門』、45 ページ。
8. 波多野完治『波多野完治全集第 1 巻 文章心理学』小学館、1990 年、104 ページ。
9. 同上、112 ページ。
10. 『最近の文章心理学』参照。
11. 『日本語文体論』、193-4 ページ。
12. 『文章心理学入門』、83 ページ。
13. 同上、『文章心理学入門』、53 ページ。
14. 同上、205 ページ。
15. 村上征勝・伊藤瑞叡「三大秘法禀承事の計量文献学的研究」『大崎学報』、立正大学仏教学会、148、1-52 ページ。
16. 『波多野完治全集第 1 巻　文章心理学』、27 ページ。
17. 小林英夫『文体論の建設』育英書院、1943 年、7 ページ。
18. 同上、7-9 ページ。
19. 村上征勝『文化を測る　文化計量学序説』朝倉書店、2002 年、50 ページ。
20. 波多野完治『文章心理学入門』小学館、1988 年、143 ページ。
21. 同上、143-4 ページ。
22. 近代化にあたって、日本だけではなく世界各国で自国語を近代文学が可能になるように改造する運動が行われたことは、パスカル・カザノヴァ『世界文学空間―文学資本と文学革命』(岩切正一郎訳、藤原書店、2002 年）に詳しい。
23. 村上征勝『シェークスピアは誰ですか？　計量文献学の世界』(文藝春秋、2004 年) を参照。
24. レーモン・クノー『文体練習』朝比奈弘治訳、朝日出版社、1996 年、目次。
25. ロラン・バルト「作者の死」『物語の構造分析』花輪光訳、みすず書房、1979 年。
26. 同上、89 ページ。
27. フランコ・モレッティ『遠読 〈世界文学システム〉への挑戦』秋草俊一郎他訳、みすず書房、2016 年。
28. 同上、72-3 ページ。
29. 同上、72 ページ。
30. フランコ・モレッティ『ドラキュラ・ホームズ・ジョイス　文学と社会』植松みどり他訳、新評論、327 ページ。
31. 同上、77 ページ。
32. https://litlab.stanford.edu/
33. Matthew L. Jockers *Macroanalysis :Digital Methods & Literary History* University of Illinois Press,2013,pp27.
34. Ibid.,pp28.

Potentiality of Digital Humanities in Literary studies.

-The Psychology of Textual Composition, Computational Authorship Attribution, and Macroanalysis-

SUGIURA Kiyoto

This paper examines the application of digital techniques and methods to literary research.

There has been a fairly long history behind the attempts to utilize numerical methods in the interpretation of literary texts. In Japan, Kanji Hatano's work is the origin of quantitative approach to literary text. In his research, which he called the Psychology of Textual Composition, he speculated the characteristics of texts by close reading in advance, then selected some countable features such as the number of nouns or the length of sentences, and counted them from the text by hand to support his observations and classify texts into certain stylistical groups. His examination and classifications of literary texts based on psychology had keen insight and his far-reaching theory of the Psychology of Textual Composition is fascinating, but had limitations because of the immaturity of the statistical technique he used.

A computer-based and statistically more strict method is Authorship Attribution; in Japan, it is represented by Masakatsu Murakami's work.The objective of Authorship Attribution is to identify the writer of a certain text. A problem of this method is the difficulty in applying it on literary interpretations and readings.

Therefore, methods that have statistical rigidity and also could be used for broader interpretation are in great demand. One of the example of such an approach is Macroanalysis by Franco Moretti and Matthew L. Jockers. Moretti pointed towards social factors such as the history of literary production, and made use of them to interpret results of quantitative analysis. With his statistically and theoretically detailed logic, Jockers skillfully approached the problem of what determines the styles of literary texts.

While there are still some technical and theoretical obstacles to go over, Digital Humanities have abundant possibilities in literary studies.