文学の統計的研究法について

安本, 美典
京都大学文学部

https://doi.org/10.15017/12729

出版情報：統計科学研究. 2 (2), pp.19-26, 1958-04. Research Association of Statistical Sciences
バージョン：
権利関係：

文学の統計的研究法について

安本美典（京大・文）

I

十九世紀以来の，めざましい自然科学の発展の影響をうけて，文学や言語学の領域においても，多分に自然科学的な研究法が，しだいにさかんになってきた．伝統的な文学・言語学の研究者から，あるときは排斥され，またあるときは受け入れられながらも，このような研究法が，さいきんは，ますます盛んになってきたように思われる．

自然科学的と思われる文学や言語学の研究は，一方では心理学と手をつなぎ，一方では，言語そのものの統計的性質を明らかにしていった．また，その内容も，音韻統計，語彙統計，文章統計，読者調査と，多方面にわかれている．しかし，研究者の数は，それほど多くはなく，このような研究を，専門にしている人は，日本全体でも，二十人とはいないだろう．

II

まず簡単に，外国および日本において，最近なされた，このような研究をながめてみよう．

1．波多野完治氏の研究．日本においては，波多野完治氏の「文章心理学」を頂点とする一連の研究がある．自然科学者が，一つの対象を，科学的な方法で分析して行くのと，ほとんど同じような態度で，文学作品という一つの対象を分析し，主に統計的な方法で整理をする．たとえば，谷崎潤一郎氏の文章が長いとか，志賀直哉氏の文章が簡潔だとかいうことを，従来は，たんに直観だけでいわれて来たのであるが，波多野完治氏は，実際に文章のいろいろな特徴を数量化し，谷崎氏の文章が長いとすれば，他の作家にくらべて，何倍ぐらい長いか，志賀直哉氏の文章が簡潔だとすれば，どこにその原因があるのかを，実際のデータをもって示された．

波多野完治氏と同じようなところに，基盤をもった研究に，小林英男氏の研究がある．

2．フレッシュの研究．アメリカのフレッシュ(Rudlf Flesch)は，文章の「読みやすさ」について研究した．どのような文章が読みやすいかを，沢山の被験者をつかい，実験心理学的に研究している．その結果は，文章が長いと，読みにくくなる．また，シラブルの長い単語が多いと，読みにくくなることを述べている．

日本文においても，朝日新聞社の，堀川直義氏が，日本文の「読みやすさ」についての，くわしい研究をされている．日本文においては，小説の文章は，文が短いほど読みやすいが，新聞の文章では，あまり短いと，読みにくくなることなどを，明らかにしておられる．

3．ジップの研究．アメリカのジップ(Zipf)の研究は，量的言語学ともいわれるべき内容をもったものである．従来の言語学は，言語構造の質的分析を問題として来たが，量的言語学は，言語の量的な法則性を明らかにしていった．

ジップは，昔から現在までの，各国の小説や文章の単語を分類し，使用回数の多いものから，順位をつけると，その頻度と順位との積は，一定であることを示した．

つまり，

$$fr = constant$$

という式がなりたつことを示した．(f=*frequence*, rは頻度数の順位 *rank*)．

これは，数千年前の，ホメロスのオディッセーでも，現代のT.S.エリオットの文章でもなりたっていると説いている．

この研究は，つぎにのべるギローなどに影響を与え，斬新な一連の言語統計を生みださせた．

4．ギローの研究．フランスのギロー(Giraud)は，ジップなどの研究を更に発展させ，ある書物の中に含まれている語彙の総量を計算する方法を，みちびき出している．

ギローはさらに，フランスの象徴詩人数人の作品をえらび，どのような単語がよく用いられているか，語彙の豊富さや，主要語彙の反復度はどうか，などを綿密にしらべ，それぞれの詩人の特徴を，うきぼりにしている．

日本においては，国立国語研究所の，水谷静夫氏が，ギローの方法の，数学的基礎に，根本的な反省を加え，日本語の書物に含まれている語彙の総量を推定する方式を，みちびきだしている．

このほか，フランスのGrammonfの音韻の統計的な研究などがある．

Ⅲ

つぎに，私自身の研究を，簡単にまとめてみよう．

私の研究は，波多野完治氏的な方法を，統計理論的な面から，さらに数歩，おしすすめてみようとしたものである．

波多野完治氏の方法は，慎重な考察と，綿密な実証とによって，うらづけられているのであるが，結局は，大量の観察と，その結果の記述とに終っている．

いわゆる推測統計学的な考え方は，その数学的なとりあつかいの困難さの故か，文学の方面では，まったく未開拓のままで，とり残されている現状なのである．

私は，まず最初に，波多野完治氏がとりあげておられる，文章を規定するさまざまな要素の，統計的な性質を，考えることから始めてみた．たとえば，「文の長さ」「比喩の使用度」「名詞や動詞の使用度」などが，統計的に整理すると，どのような型の分布に従うかを考えてみた．

文　の　長　さ

波多野完治氏も調べておられるのであるが，文章の長さを，「字数」ではかり，その度数分布をしらべてみると，ちよっと，F分布などに似たような形をしている．この分布型が，理論的には，どのような式で表わされると考えたらよいのか，いろいろ考えてみたが，現在のところ，見当がつかない．

そこで，考え方を変え，つぎのような方法で，文章の長さをはかることにした．それは，全集本などで，書物の各頁に含まれている，句点「。」の数を数え，その句点の度数分布について考えるのである．各頁に存在する字数は，ほぼ一定している．したがって，一頁に存在する総字数を，句点の数で割れば，一つの句点あたりの字数，つまり一文章の平均の字数が算出される．そこで，各頁に存在する文章を，標本と考えれば，「標本の平均値の分布は，正規分布に近づく」というまったくの数学上の理由から，各頁に存在する句点の数の分布は，正規分布に近づく傾向があるのではなかろうか．

そこで，実際に，志賀直哉氏の「暗夜行路」と，有島武郎の，「或る女」とをとりあげ，その各頁に存在する句点の度数分布を，調べてみた．適合度の検定では，一応，正規分布とみなして，さしつかえないという結果が得られた．

実際には，一頁あたりの字数を，一頁あたりの句点の数で割って，文章の平均

字数を算出することも可能であるが，同一の全集本などによって，一頁あたりの句点の数が，多いか少いかさえを調べれば，文章の長さが，長いか短いかが，わかるわけである．また，二人の作家の，文章の長さの比較も，句点の数によって比較すれば，F検定やt検定を，用いることが出来るわけである．

比喩の使用度

比喩については，私は，直喩と声喩とについて考えてみた．直喩というのは「……のような」を用いた比喩であり，声喩というのは，「はらはら」「ほろほろ」などの擬態語をさしている．

各頁に存在する直喩や声喩の数を数え，その度数分布を考えてみると，その使用度が小さいため，だいたいポアソン分布に近い分布をしているようである．適合度の検定を行ってみても，直喩の場合も，声喩の場合も，一応ポアソン分布とみなして，さしつかえないという結果がえられた．

色彩語の使用度

「赤」「青」「黒」「白」など，色彩に関しての言葉を，比喩の使用度などと，同じようにして，しらべてみた．色彩語の使用度も，かなり小さいものであるが，検定を行ってみると，ポアソン分布とはいえない．

色彩語は，人物描写や，情景描写によく用いられ，心理描写や，会話文のところでは，ほとんど用いられない．また「黒い眼」「黒い髪」「赤い唇」など，作者の眼が，一度色彩にすえられると，たてつづけに色彩語があらわれる．つまり，色彩語は，かたまってあらわれる場合が多いのである．そのため，その度数分布の分散は，平均値をはるかに上まわり，とうていポアソン分布とは，いえなくなってくるのである．このように広い意味での，作者の連想作用といったものを考えると，自殺現象や，特定地域の伝染病患者発生数などと，問題がいくらか似ているように思われる．そこで，色彩語の分布に，ポリヤ・エゲンベルガー分布をあてはめてみた．その結果は，ポアソン分布よりも，はるかにうまく適合することが，いえるのである．

品詞の使用度

文章から，「百字づつ」の標本をたくさん抽出し，その百字中に含まれている，名詞や動詞の数を数え，その分布をしらべてみると，正規分布に近い．これは，

百字中に含まれる品詞の数は，ほぼ一定し，名詞などの個数も，大体一定しているためと思われる．

なお，J.R.Thomsonなどは，1915年に，すでに，語の分布が，「ポアソン分布」にしたがうことを，明らかにしている．この，Thomsonなどの研究を，イギリスの統計学者，G.U.Yuleが，壺から珠をとりだす確率の問題として明示してから，その応用がなされるようになった．

IV

さて，このような，文体統計の立場から，作者未詳の古典の作者や，執筆時期を推定する一助とすることはできないものであろうか．

日本の文献学では，このような作家の文体の特徴を考えるということは，これまで，まだほとんど行われていない．しかし，外国の文献学では，このような文体統計の方法も，作者の推定や，執筆時期の推定に，使用されている．たとえば，先にのべたギローは，ランボオの一作品の，執筆時期の推定を，記述統計的な立場からではあるが，行っている．

私も，ほんの一つの試みにすぎないのであるが，むかしから，しばしば論争の種になった，源氏物語の宇治十帖の作者について考えてみた．方法としては，源氏物語の文章のさまざまな特徴を数量化し，統計的に整理し，検定を行ってその差を示す．さらに，その差をいろいろな面から考えてみることにより，宇治十帖の作者が，果して他の四十四帖と同じ筆者，つまり紫式部によって書かれたものであるかどうかを考えてみた．

さて，武田宗俊氏は，源氏物語の成立過程について，次のような説をたてておられる．源氏物語は，現在並んでいる順序で執筆されたものではなく，仮に「紫上系」と名づけられる巻々が先に書かれ，「たまかつら系」と名づけられる巻々が後で書かれた．つまり，後で書かれた「たまかつら系」が，先に書かれた「紫上系」の間に，やや無理なかたちで，挿入されているというわけである．

武田氏は，いろいろな理由から推定して，このような説をたてておられるのであるが，私は，コントロール的な意味で，この「紫上系」と「たまかつら系」との文体の比較も行ってみた．

また，私は別に，宇治十帖を除いた他の四十四帖について，その最初の十帖(

これを仮に「桐壷十帖」と名づけよう）と，最後の十帖（これを仮に「梅枝十帖」と名づけよう）とをえらび，文体統計による比較を行ってみた．これは，宇治十帖を除いた四十四帖でも，かなり長篇なので，その最初と最後とでは，かなりの執筆時期の差が考えられる．そこで，この最初の十帖と最後の十帖とを比較してみることにより，執筆時期の差が，どれぐらい文体の差を生みだすかをみてみようと思ったわけである．

以上すべての調査の結果をまとめると，次のような表がえられた．

	紫上系と玉鬘系	桐壷十帖と梅枝十帖	宇治十帖と他の四十四帖
頁数	—	97	99
和歌	95	—	93
直喩	—	—	95
声喩	—	—	97
心理描写	—	93	91
文の長短	99	—	—
色彩語	—	—	95
名詞	—	97	99
用言	—	—	91
助詞	—	—	96
助動詞	90	—	—
品詞数	—	—	94

この表の，頁数，和歌の使用度，直喩，声喩の使用度，心理描写の数，文の長短，色彩語の使用度などの項目は，文章のいろいろな特徴を数量化したもので，別にくわしく調査を行ったものである．表のうち，斜線を引いたものは，検定を行って，有意の差のみられなかったものである．また，表中の数字は，100から危険率を引いた値である．たとえば，宇治十帖と他の四十四帖とでは，その各巻の頁数の平均的な値において，一パーセント以下の危険率で差がみられたので，100-1=99で，99と記してある．したがって，表中の数字は，大きければ大きいほど，差の激しいことを示している．なお，この検定には，分布の型の不明なものが多かったので，すべてノンパラメトリック検定を行っている．ノンパラメトリック検定では，もし分布が正規であるような場合は，従来の検定にくらべて，一般に効率が低いと考えられるので，10パーセント以下の危険率で差のみられ

たものは，一応表中に記しておいた．

このようにしてしらべてみると，宇治十帖と他の四十四帖とでは，かなり文体が異なっているようである．検定の結果では，調べた十二項目のうち，十項目について差がみられている．

同一作家でも，たとえば森欧外などでは，初期と晩年とでは，大変文体が異なる．「舞姫」と「寒山拾得」とを並べてみると，その持っている雰囲気がとても違う．ところが，志賀直哉氏では，その初期の作品も後期の作品も，それほど大きな文体の相異を持っていないようである．「暗夜行路」など，その前篇と後篇とでは，執筆時期に十年のへだたりがあるにもかかわらず，文体はほとんど変っていない．読んだ感じが違わないばかりでなく，統計的にも差がみとめられないのである．

このように，作家によって，その文体の変化には違いがある．いちじるしく文体の変動しやすい作家もあれば，生涯を通じて，ほとんど変化しない人もある．

したがって，源氏物語においては，源氏物語の場合について，その文体が変動しやすいものであるかどうかを，調べてみなければならない．

その為に私は，「紫上系」と「たまかつら系」との比較や，「桐壷十帖」と「梅枝十帖」との比較を行ってみたのである．たとえば，「桐壷十帖」と「梅枝十帖」とでは，執筆時期が異ると考えられるばかりでなく，題材もかなり差がある．「桐壷十帖」は，「夕顔」「若紫」などの巻を含む，夢多い少・青年時代の，はなやかな恋の遍歴を描いた絢爛たる絵巻であるのに対し，「梅枝十帖」は，「御法」「幻」などの巻に象徴される，暗いどうすることも出来ない人間的な苦悩を湛えた巻々である．

このような題材や執筆時期のかなりの差が考えられるにもかかわらず，これらは，十二項目のうち，わずか三項目においてしか差はみられなかった．

源氏物語は，それほど変化しない文体であるように思われる．それでは，宇治十帖と他の四十四帖との，十項目にわたる大きな文体の差は，どこにその原因があるのであろう．

これだけの調査から，早急な結論を出すのは，大変危険であると思うが，あるいは，宇治十帖の作者は，他の四十四帖と異なるのではなかろうか．

なお，調べた十二項目では，従来の文体統計で，文章の特徴を示すものとして考えられたものを，ほとんど網羅している．これは，現在の段階では，どのよう

な項目が，作者の推定に役立ち，どのような項目が役立たないか，全く不明である．それで，なるべく項目を多くして，全体的に文体が異なっているか，異っていないかを考えてみたわけである．少くとも，一項目か二項目だけで作者の推定を行うのは，とても危険なことと思う．いずれにしても，将来は，もっと有効な，妥当性の高い方法が考えだされることと思う．

なお，本居宣長は，源氏物語の筋の欠けているところを補うつもりで，「手枕」一巻を書いている．これは，後人が源氏物語の文体をまねて，代作する場合，どんな文章が出来上るかを示す意味で，面白い研究材料になると思う．私は，「手枕」などの分析も行ってみたが，たとえば，「手枕」の文章の平均の長さは，源氏物語の文章の平均の長さの，二倍以上もあるという，途方もないものである．

最後に，このような文学・言語学の研究水準の向上にしたがって，文学と統計学との，双方の財産がより豊かになる日が，一日も早く来ることを願って，この筆をおきたいと思う．

参　考　文　献

つぎの書物には，その筆者の書いた，他の書物も紹介されている．

「文章心理学入門」波多野完治著　新潮文庫

「How to Test Readability」　Rudolf Flesch.

「分りやすさの分析」堀川直義著　朝日新聞社

「The Principle of the Least Effort」　G. K. Zipf.

「Les Caractéres Statistiques du Vocabulaire」　Pierre Giraud.

参　考　論　文

「語彙の量的構造の諸問題」水谷静夫，　「計量国語学」ヤ二号掲載予定．

「文の長さの分布型」拙稿，　「計量国語学」創刊号．

「みだれ髪と一握の砂の音韻統計による比較」拙稿，　「解釈」ヤ三巻ヤ六号．

「宇治十帖の作者」拙稿，　「文学・語学」ヤ四巻．