句点のゆらぎの解析による文学作品の作家推定

牧野浩典，吉田一美
東海大情報理工

1. はじめに

人は文字を使った主張や交流を繰り返してきた。文字のまとまりとしては、論文、文学、手紙、詩などが挙げられる。これらには、どんな形であれ書き手それぞれの異なった特徴が表れる。人間の指紋のように一人一人、書き手ごとに固有の特徴を含んでいると言ってよい。

文章は長い文と短い文で構成されており、文章を句読点や単語で区切ると、それら構成要素の長さには、「長い」「短い」・・・といったリズムが表れる。本研究では、このような文の構成要素の長さに生じるゆらぎを統計的に定量化し、著者を識別する新しい方法を提案する。文章のゆらぎを分布関数を用いて解析し、著者の特徴を捉えることが目標とした。

文章が持つゆらぎの法則として最もよく知られているものに、ジップの法則[1]がある。この法則は文章を単語単位で区切り、単語の出現頻度とその順位を求めると、一定の傾きを持つ比例関係で結ばれるという経験則である。ジップの法則は、個々の書き手の特徴には全く無関係に成り立つ。これに対し、我々の研究では個々の書き手の特徴が意識とは無関係に、文章の中に秘められてしまう可能性を明らかにする。

2. 文章のゆらぎ

対象となる文章を句点ごとに分割し、句点に挟まれた領域の文字数を、句点間距離 S と定義する。本研究ではこの距離 S によって表される文章の長さのゆらぎを調べる。S を横軸、その出現頻度を縦軸にとり、分布関数 $P(S)$ として表したものが図１である。ただし、句点間距離 S は、平均値が１になるように横軸の数値を規格化している。また、分布関数 $P(S)$ は面積が１、すなわち $\int_0^\infty P(S)dS = 1$ となるように縦軸の数値を規格化している。分布関数 $P(S)$ はゆらぎの統計的性質を与える。

もしも、各句点が互いに無相関でランダムな配置をとるならば、距離 S のゆらぎは指数分布 $P(S) = \exp(-S)$ で表される（ポアソンの少数の法則）。これに対し、人が書いた文章のゆらぎは一般にランダムなゆらぎよりも極端に短い文と極端に長い文の発生頻度が少ないため、指数分布から大きく外れる。句点の配置には相関があり、相関の強弱には著者の個性が反映しているのではないかと考えられる。

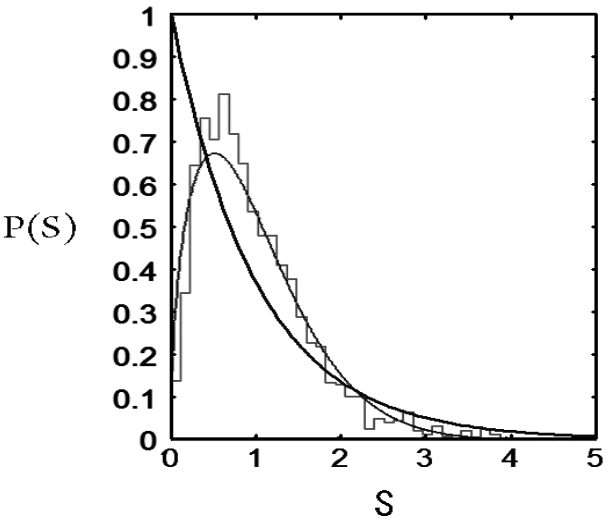

図１　森鴎外の小説「雁」に対する句点の最近接間隔分布 P(S)。太実線はポアソン分布、細実線はブロディパラメータ β=0.487 で与えたブロディ分布

3. ゆらぎの定量化

最近接間隔分布に表れる句点列の相関の大き

さをブロディ分布を用いて定量化する。ブロディ分布は単一パラメータ β ($0 \leqq \beta \leqq 1$)によって句点列に生じる相関の大きさを定量化することのできる分布関数であり、下記の定義で与えられる。

$$P_\beta(S) = \alpha(1+\beta)S^\beta \exp\left(\alpha S^{1+\beta}\right)$$

$$\alpha = \left[\Gamma\left(\frac{2+\beta}{1+\beta}\right)\right]^{1+\beta}$$

β=0 のときには最大ゆらぎのポアソン分布を与え、$0<\beta \leqq 1$ では図１に示した句点の最近接間隔分布に近い概形を与える。$\Gamma(x)$はガンマ関数である。個々の文学作品から得られた最近接間隔分布に対し、ブロディ分布関数を最小二乗法でフィッティングし、最適なブロディパラメータ β を導出する。図１には森鴎外の小説から得られた最近接間隔分布に最もよく一致するブロディ分布を示した。このときのブロディパラメータは β=0.487 である。

3. ブロディパラメータの測定結果

図２には同様の手続きにより個々の作品に対して導出したブロディパラメータ β を、作家別に表記した。興味深いことに、ブロディパラメータは作家ごとに異なる固有の集積点を持っており、最近接間隔分布で表された句点間隔のゆらぎには、個々の作家の特徴が明確に表れている。

また、β の値が０付近に集まっている太宰、１付近に集まっている芥川に関して、二人とも自殺する直前に発表した数作品が特に極端な数値（芥川＝０、太宰＝１）を示していることがわかる。両名の自殺直前の作品が他の作品に劣ることは決してないが、彼らの生々しい感情がむき出しになった表現が文章の所々に登場し、それが極端な数値として表れているのではないだろうか。そういった表現を含め、作家の感情面の振幅が観測結果に反映していると考えられる。

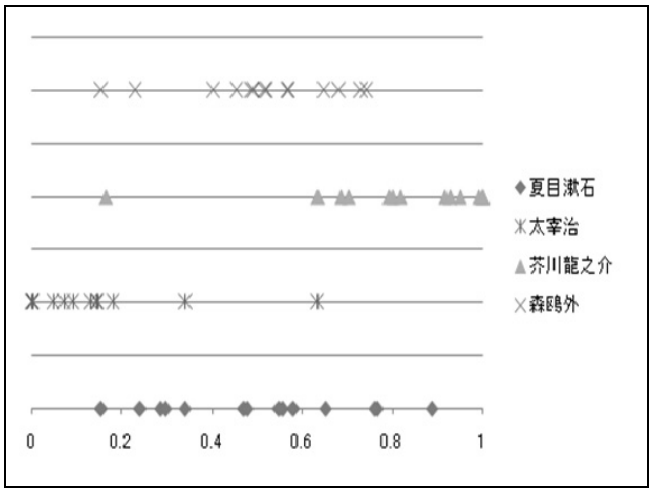

図 2　作家別にまとめたブロディパラメータ β の測定結果。１つのプロットが１作品に対応している

4. まとめ

本研究では日本語の文章の句点間距離に生じるゆらぎの統計的性質を、分布関数を用いて分析した。分布関数に表れるゆらぎの特徴をブロディパラメータを用いて定量化し、個々の作家の特徴を取り出すことに成功した。本稿で紹介した方法は、例えば作者不明の文学作品について、作家の過去の作品を用いた作者の同定に利用出来る可能性がある。また、一般の文章についても、著者の性格やその時の精神状態（例えば自殺願望者の感知など)を推測できるようになるかもしれない。

6. 参考文献

[1] G. K. Zipf, *Human Behavior and the Principle of Least Effort*, Addison-Wesley (1949).

[2]金明哲・村上征勝他（2003）：言語と心理の統計：岩波出版

[3]インターネット電子図書館・青空文庫：http://www.aozora.gr.jp/