国立国語研究所学術情報リポジトリ

“文の長さ”の統計学的一考察：時系列的解釈

メタデータ	言語: Japanese 出版者: 公開日: 2017-03-31 キーワード (Ja): キーワード (En): 作成者: 米田, 正人, YONEDA, Masato メールアドレス: 所属:
URL	https://doi.org/10.15084/00001039

"文の長さ" の統計学的一考察

——時系列的解釈——

米　田　正　人

一人の作者によって書かれた文章の特徴を数量的に知ることは興味のあることである。作者はおそらく意識していないであろうが，書かれた作品がある周期を持ったとしても不思議なことではない。多くの作品について調査すれば，共通の傾向，特徴を統計的に検出できるかもしれない。ここでは "文の長さ" について，一つの作品全体を通して，どの様な流れを持ち，どの様な特徴を持つのか少数の例について統計学的手法をかりて調査してみることにする。

詳細については後節でふれるが，統計学的手法を当てはめる際データのとり方に多少問題があるので，その方法の当てはめの可能性を主として問題にし，以下の四編の作品について調査を行なうことにする。

森　鷗外	『高瀬舟』	鷗外全集（岩波書店）による
	『寒山捨得』	鷗外全集（岩波書店）による
志賀直哉	『城の崎にて』	現代日本文学大系（筑摩書房）による
	『焚火』	現代日本文学大系（筑摩書房）による

§1　単　　位

電子計算機で言語処理をする際，単位は非常に重要な問題である。文の長さを測定する際にも如何なる単位を用いるかは重要な問題となる。用いる単位によって測定された長さにかなりの差が出るからである。また単位切りの際に，単位の安定性ということも少なからず結果に影響を与えるであろう。しかしながら，単位の選定にあたっては，まだ解決されねばならない幾らかの問題を含んでいる。どの単位を選んだとしても問題は残るであろう。とりあえず，この

調査では，文節単位（国語研究所C単位）を用いることにする。文体にあまり影響されない単位という理由でこの単位を選んだ。

なお，原文を文節に切る際の作業上の問題点を以下に挙げておく。

i）一文中に引用されている対話部分は，対話部分全体を一つの文節タイプに相当するものとして扱った。

〔例〕

／「まあ，一寸」と／閭が／呼び留めた／。

この例の場合，文の長さは〝3″となる。

ii）対話部分が単独で出て来る場合は，その対話部分は文節測定の対象とは考えなかった。

〔例〕

「寸志のお禮がいたしたいのですが。」

この様な対話部分は省いて測定を行なった。

iii）本文中では切れていないが，文の性質上そこで切った方が適当と思われるものについては，そこを切れ目として扱った。

〔例〕

山の手線の電車に跳ね飛ばされて怪我をした，その後養生に一人で但馬の城の崎温泉へ出かけた。

上の例文の場合，「怪我をした」と「その後養生に」の間にある「，」を「。」と考えて，二文として扱った。

iv）記号類は測定の対象としなかった。

v）その他，補助用言の扱い等，細部においては説によって定義が異なるものがあるが，それらについてはある一貫した定義に基づいて，作業規則を作り，それに従って作業を進め，単位切りによりデータに乱れが生じないよう注意した。

§2 分　　布

文字によって測定された文の長さの度数分布は対数正規分布で近似出来るということが知られている。文節で測定したものについてはどうであろうか。そのことを対数正規確率紙を使って調べてみよう。fig. 1, fig. 2 にそのグラフを示す。グラフからわかるように，プロットされた点はほぼ一直線上にならぶ。このことから文節単位で測定した文の長さの分布は，ほぼ対数正規分布をしていることがわかる。ためしに「城の崎にて」のデータを正規確率紙にプロットすると fig. 3 の様になり直線から大きくはずれていることがわかる。

Table 1 は「城の崎にて」の"文の長さ"の累積度数表である。表の右側の数値は累積度数を総文数で割り 100 倍したものである。例えば"文の長さ"1 から 7 までの累積度数は全体の 58.85% を占めることを意味する。この数値を対数正規確率紙にプロットしたのが fig. 1, fiig. 2 である。

Table 1 「城の崎にて」累積度数表

文の長さ	度数	累積度数	累積度数 / 総文数× 100	文の長さ	度数	累積度数	累積度数 / 総文数× 100
1	2	2	$\frac{2}{209}\times 100=0.957$	12	5	172	82.30
2	8	10	4.78	13	12	184	88.04
3	17	27	12.92	14	7	191	91.39
4	26	53	25.36	15	1	192	91.87
5	31	84	40.19	16	7	199	95.22
6	22	106	50.72	17	1	200	95.69
7	17	123	58.85	18	2	202	96.65
8	16	139	66.51	19	3	205	98.09
9	10	149	71.29	23	1	206	98.56
10		157	75.12	24	3	209	100.00
11	10	167	79.90				

fig. 1　城の崎にて

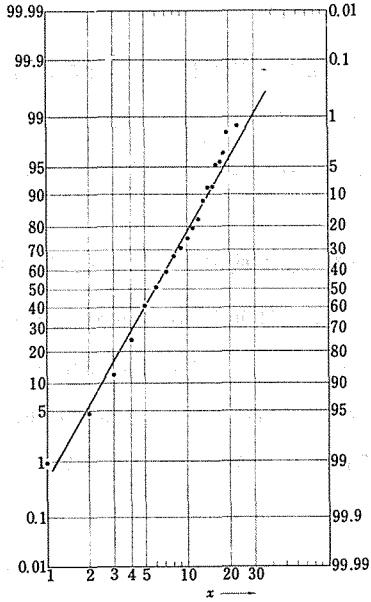


fig. 2　寒山拾得

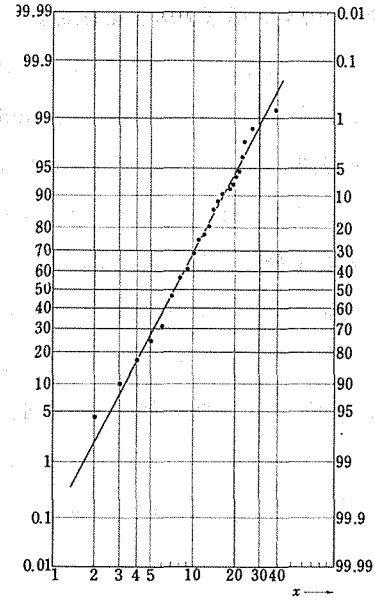
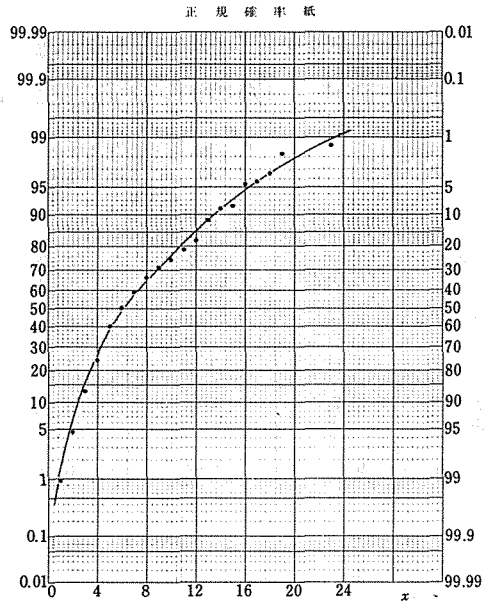


fig. 3　城の崎にて



§3　平均，標準偏差，変動係数

四編の作品について，一文中に含まれる文節数（文の長さ）の平均，標準偏差，変動係数を計算したものをtable 2にまとめておく。分布が対数正規なので対数変換後のデータのほうが分類を試みるには適していると思われるが，生のデータの表も示しておくことにする。

対数変換後のデータの平均について平均値の差を検定してみると，同一作者の作品の間には有意差はないが，異なった作者の作品の間には危険率5%で差が生じた。分散に関しては，等分散性の検定をしてみたが，有意水準5%では差を検出できなかった。また，変動係数が安定した値になっているのも面白い。この表だけからなら，志賀直哉のほうが森鷗外より僅かではあるが変化のある文章を書くが，文の長さは全体として短い文章を書くということが言えそうである。

Table 2　平均，標準偏差，変動係数

作　　者		森　鷗　外		志　賀　直　哉	
作　　品		高瀬舟	寒山捨得	城の崎にて	焚　火
対数変換後のデータ	総文数	126	130	209	211
	平　均	0.96	0.90	0.82	0.79
	標準偏差（分散）	0.29（0.082）	0.27（0.071）	0.26（0.067）	0.28（0.078）
	変動係数	0.30	0.30	0.32	0.35
生のデータ	平　均	11.21	9.49	7.82	7.59
	標準偏差（分散）	6.94（48.20）	6.31（39.82）	4.64（21.52）	5.35（28.59）
	変動係数	0.62	0.67	0.59	0.70

§4　文の長さの推移

平均，標準偏差等については前のセクションで述べたが，それらは一つの値に情報が集約されていて，全体としてどのような流れを持つのかはわからない。そこで，その推移の経過をグラフにまとめて眺めてみることにしよう。

この系列は非常に凹凸が激しくて，見ただけでは特徴をつかむことが出来ない。そこで補助手段として9項移動平均の系列を前の系列の各点に対応させてみる。いくら滑らかになり全体の大きな流れをつかむのに便利であると思う。

fig. 4 「寒山拾得」文長の推移と9項移動平均

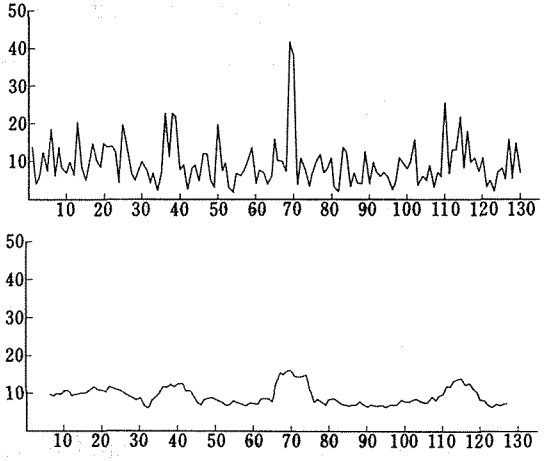
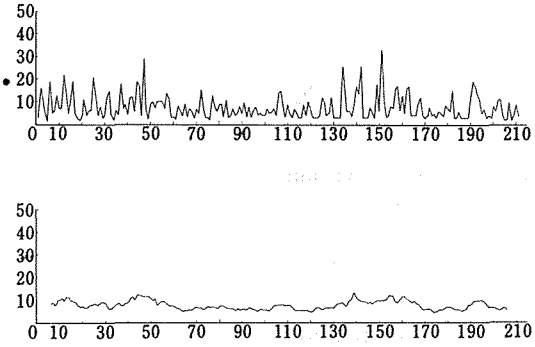


fig. 5 「焚」文長の推移と9項移動平均



しかし，移動平均や推移のグラフを見ても，人間の視覚には限りがあるので系列に小さな周期があったとしても，それを発見することは困難である。そこで次の節では，コレログラムにより，周期について考えてみたい。

§5　コレログラム分析

文の長さの系列はどの様な周期を持つのであろうか。そのことを知る為に，コレログラム（Correlogram）と呼ばれる系列相関係数のグラフを用いて，分析を試みてみることにする。

系列相関係数とは，時系列で一定の距離だけ離れた，2項の値の間の相間を意味している。そして2項間の隔たりをLag（遅れ）という。いま系列を x_1, x_2, ……, x_n とすると，遅れ l の系列相関係数は次の様になる。いま，系列 x_s と，各 x_s に l だけ隔たった系列 x_{s+l} を考える。

系列 x_s	系列 x_{s+l}
x_1	x_{1+l}
x_2	x_{2+l}
x_3	x_{3+l}
…	…

この x_s と x_{s+l} の間の相関係数が，遅れ l の系列相関係数と呼ばれるものである。そして，この系列相関係数の値を遅れの関数として，グラフに表わしたものをコレログラムという。数式で表現すると，

$$C(l)=\frac{1}{n-l}\sum_{s=1}^{n-l}(x_s-\bar{x}_1)(x_{s-l}-\bar{x}_2)$$

を系列共分散（Serial covariances）

$$r(l)=\frac{C(l)}{C(O)}$$

を系列相関（serial correlations）という。

ここに

$$\bar{x}_1=\frac{1}{n-l}\sum_{s=1}^{n-l}x_s$$

$$\bar{x}_2=\frac{1}{n-l}\sum_{s=l+1}^{n}x_s$$

r(l)（l=1,2,……, n−l）が標本系列のコレログラムである。

r(l)が統計学的に有意であるとき，遅れlで，その系列は何らかの周期を持つと考えられる。

さて，ここで文の長さの推移の系列について，このコレログラムを作るとどうなるであろうか。計算した結果を以下のグラフにまとめてみる。(fig. 6—fig. 9)

fig. 6 「焚火」コレログラム

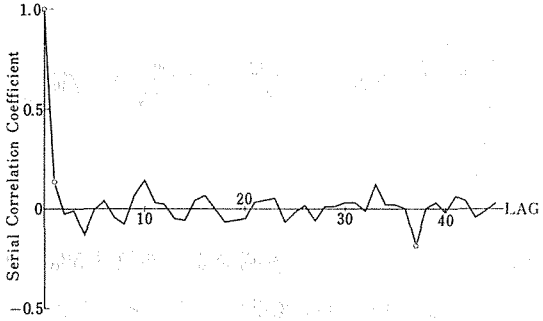


fig. 7 「城の崎にて」コレログラム

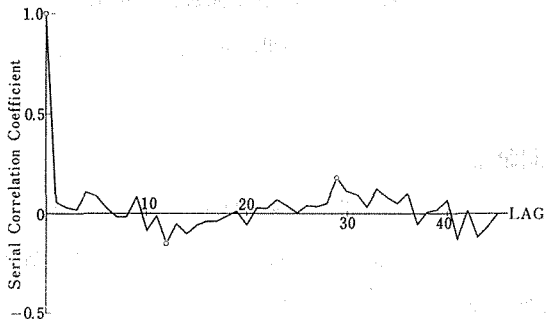


fig. 8 「寒山拾得」コレログラム

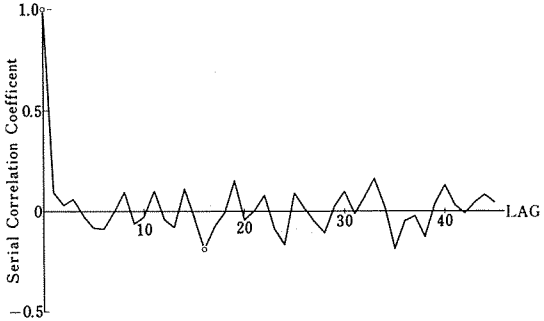
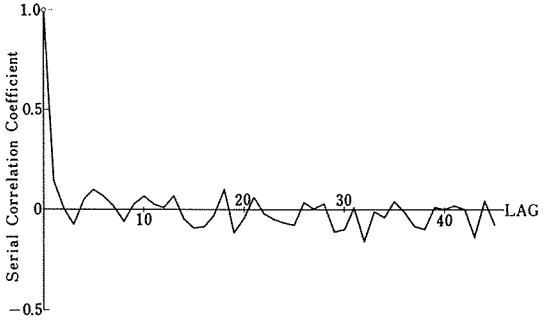


fig. 9 「高瀬舟」コレログラム



グラフで○印のついている点は，相関係数が危険率 5% で有意になっている点である。したがって，○印の点は相関があると考えられる。「焚火」では遅れ 1 と 37 で，「城の崎にて」では遅れ 12 と 29 で，「寒山捨得」では遅れ 16 でそれぞれ相関がありそうである。しかし相関係数が±0.15 から±0.20 位の値をとり，有意点ぎりぎりであるため，周期があったとしてもはっきりしたものではないだろう。

§6　今後の問題点

今まで述べてきたことは，僅か四編の作品についてであった。今後，より多くの作品についてデータを蓄積する必要がある。その為に解決せねばならないいくつかの問題があるので，それを以下に列記する。

i) 単位としては何が適当か。作業上の問題も含めて考えねばならない。

ii) 対話部分の扱いをどうするか。

iii) 系列相関係数を計算する際，もとの文長の系列が定常(stationary)(注)であるからか疑わしい。

これらの問題をふまえて，今後多くのデータを検討すれば，§3に記した変動係数も作品の分類に役立つであろう。

また，コレログラムについても，多量にデータがあれば，型による分類の可能性もより明確になるであろう。

これらについては，今後の調査，研究に負うところが大きい。

参考文献

中山伊知郎編 「現代統計学大辞典」東洋経済新報社 1962

山内二郎編 「統計数値表 JSA-1972」日本規格協会 1972

赤池弘次・中川東一郎 「ダイナミックシステムの統計的解析と制御」サイエンス社 1972

P. G. ホーエン 「初等統計学」培風館 (浅井 晃・村上正康共訳) 1970

注) 時間軸が移動しても，確率分布には変化がないような不規則変動現象の観測値の系列を定常時系列と呼ぶ。先に述べた系列相関係数は定常であることが前提となっている。