

日本語における単語の長さの分布と文章の著者^(注1)

金 明哲

In research on Literary style and establishment of authorship, one of the most important points is which kind of elements abstracting from author's texts are used in analysis. In the present study, the statistical analysis of relationship between the word-length distributions and authors was used for 21 Japanese texts of 3 writers. The statistical analysis is based on the word-length distributions classified by parts of speech and the other is not. This analysis showed that the word-length distributions classified by parts of speech brought more accurate establishment. So, it is a better approach to use the word-length distributions classified by parts of speech in determining the authorship. The same analysis method is predicted to be adaptable in other languages. The author's characteristic included in word-length distributions is evaluated by use of the variance ratio and correlation ratio of the average distance of the same author and the minimum average distance of the different authors.

1. はじめに

文章の著者を推定したり、著者別に文章を分類する場合、著者の特徴情報として文章に関するどのような要素を用いるかが問題解決の鍵である。文章の著者の推定や文体などの研究では、欧米文では文の長さ^{(1),(2),(9),(24)}、単語の長さ^{(3),(4),(5),(8),(22),(23)}、単語の使用頻度^{(6),(7),(19),(20)}、などに関する情報、日本語では文の長さの分布^{(21),(25),(26)}、単語の使用頻度⁽¹⁹⁾、品詞の使用率⁽¹⁷⁾などがよく用いられている。文章の著者の推定などを行う場合、文章に関するどのような要素に注目すべきかは言語の種類によって異なると考えられる。例えば、日本語では漢字と仮名が混じっているため漢字の使用率に著者の特徴が現れる可能性がある⁽¹²⁾。また、日本語では読点の付け方に明確

な規則がないため、読点に関する情報も著者の特徴になる^{(10),(11),(15)}。では、単語の長さはどうであろうか。残念ながら、日本文が英文のように“分かち書き”されていないためであろうか、単語の長さの情報に関する基礎的な研究はない。近年の日本語の情報処理の技術の進展から、遠くない将来かなり精度が良い単語の自動分割システムが開発されると予測される⁽¹⁸⁾。したがって、単語に関する情報の基礎的研究には十分な意味を持つことは言うまでもない。そこで、本研究では3人の21編の文章(合計87,352単語)を用いて、単語の長さの分布には著者の特徴が現れるか否か、どうすれば著者の特徴がもっと明確に現れる質のよい情報が得られるかについて統計分析を行った。

表1 分析に用いた文章のリスト

著 者	文 章 名	文章の記号	単語数	出版社	発表の年
井上 靖	結婚記念日	I 1	4749	角川文庫	1951
	石庭	I 2	4796	同上	1950
	死と恋と波と(前半)	I 3	4683	同上	1950
	死と恋と波と(後半)	I 4	4386	同上	同上
	帽子	I 5	3724	新潮文庫	1973
	魔法囃	I 6	3624	同上	同上
	滝へ降りる道	I 7	3727	同上	1952
	晩夏	I 8	4269	同上	同上
三島 由紀夫	遠乗会	M 1	4984	新潮文庫	1951
	卵	M 2	4004	同上	1955
	詩を書く少年	M 3	4502	同上	1955
	海と夕焼	M 4	3359	同上	1955
中島 敦	山月記	L 1	3226	新潮文庫	1942
	名人伝	L 2	3202	同上	1942
	弟子(前の1/3)	L 3	4078	同上	1943
	弟子(中の1/3)	L 4	4092	同上	同上
	弟子(後の1/3)	L 5	3727	同上	同上
	李陵(前の1/4)	L 6	4563	同上	1944
	李陵(中の1/4)	L 7	4561	同上	同上
	李陵(中の1/4)	L 8	4638	同上	同上
	李陵(後の1/4)	L 9	4458	同上	同上

2. 分析に用いた文章

統計分析に用いたのは井上靖, 三島由紀夫, 中島敦の短篇小説である. 単語の長さ等の分析に用いた情報の安定性を見るため, これらの中で比較的長い文章はいくつかに分割した. 例えば, 井上の「恋と死と波と」は二つに, 中島の「弟子」は三つに, 「李陵」は四つに分割して用いることにした. 表1に, 用いた文章と発表年やサイズなどを示した. 単語の認定基準としては長い単位を用いた. 品詞の認定は「広辞苑」, 「国語辞典」^(註2)に基づいた. ただし, 複合語は複合された単語を1語と見なした.

3. 単語の長さの分布

表2に, 表1の21編の文章における1文字から6文字までの長さ別の単語の使用頻度を示した.

いま, 文章*i*の長さ*j*の単語の使用頻度を x_{ij} と表すと, *I*編の文章における長さ*J*までの単語の使用頻度と相対使用頻度のマトリッ

表2 単語の長さ別の使用頻度

著者名	文章の記号	単語の文字数(長さ)					
		1	2	3	4	5	6
井上	I 1	2387	1536	534	215	62	14
	I 2	2409	1692	456	187	41	10
	I 3	2395	1492	488	227	60	14
	I 4	2432	1367	371	154	53	6
	I 5	2068	1164	345	126	17	3
	I 6	1977	1081	352	175	23	13
	I 7	2051	1175	322	147	30	2
	I 8	2390	1248	397	178	34	15
三島	M 1	2437	1579	595	281	72	18
	M 2	2053	1219	443	207	60	14
	M 3	2308	1442	509	197	35	4
	M 4	1814	1045	300	163	30	4
中島	N 1	1755	1112	266	70	21	2
	N 2	1761	1074	266	86	11	3
	N 3	2099	1546	294	109	19	8
	N 4	2117	1518	295	121	31	6
	N 5	2010	1357	254	84	13	7
	N 6	2375	1666	372	112	33	3
	N 7	2405	1609	398	105	39	5
	N 8	2466	1610	391	141	25	5
	N 9	2411	1565	344	109	23	3

クスはそれぞれ

$$X_{I \times J} = [x_{ij}]$$

$$P_{I \times J} = [p_{ij}]$$

$$p_{ij} = \frac{x_{ij}}{\sum_{v=1}^J x_{iv}}, \quad \sum_{j=1}^J p_{ij} = 1$$

に表示できる。

本文では相対使用頻度に関するデータを分布と呼ぶことにする。

3.1 群内の距離と群間の距離

もし、同じ著者の任意の二つの文章における単語の長さの分布の間の距離（以下群内の距離と呼ぶ）が、異なる著者の間の任意の二つの文章における単語の長さの分布の間の距離（以下群間の距離と呼ぶ）より小さい場合、単語の長さの分布には著者の特徴が現れていると考えられる。そこで、単語の長さの分布について、群内（同じ著者）、群間（異なる著者の間）の任意の二つの文章における単語の長さの分布の間の距離について分析することにする。さて、文章 i の単語の長さの分布と文章 l の単語の長さの分布の間の距離 d_{ij} を次のように定義する。

$$d_{ij} = \frac{1}{2} \sum_{j=1}^J \left(p_{ij} \log \frac{2p_{ij}}{p_{ij} + p_{lj}} + p_{lj} \log \frac{2p_{lj}}{p_{ij} + p_{lj}} \right)$$

ただし

$$p_{ij} = 0 \text{ なら } p_{ij} \log \frac{2p_{ij}}{p_{ij} + p_{lj}} = 0$$

$$p_{lj} = 0 \text{ なら } p_{lj} \log \frac{2p_{lj}}{p_{ij} + p_{lj}} = 0$$

とする。

上式で求められた分布の間の距離マトリクスを

$$D_{I \times I} = \begin{bmatrix} 0 & d_{ji} \\ d_{ij} & 0 \end{bmatrix}$$

で表記する。

分類の観点からは、群内での任意の二つの分布の間の距離の平均値が、群間での任意の

二つの分布の間の距離の平均値より小さいことが望ましい。いま、著者 k と h のそれぞれ k_n, h_m 編の文章があったときに、それぞれ著者に関し、群内、群間での任意の二つの文章のすべての組み合わせに関して、単語の長さの分布の間の距離の平均をそれぞれ

$$\overline{d(k)} = \frac{2 \sum_{k_i=k_1}^{k_n-1} \sum_{k_{i+1}=k_i+1}^{k_n} d_{k_i, k_{i+1}}}{(k_n - 1) k_n} \times 100$$

$$\overline{d(h)} = \frac{2 \sum_{h_i=h_1}^{h_m-1} \sum_{h_{i+1}=h_i+1}^{h_m} d_{h_i, h_{i+1}}}{(h_m - 1) h_m} \times 100$$

$$\overline{d(k, h)} = \frac{2 \sum_{k_i=k_1}^{k_n} \sum_{h_j=h_1}^{h_m} d_{k_i, h_j}}{k_n h_m} \times 100$$

で求めた。

表3に、表2に示した3人の21編の文章における単語の長さの分布を用いて求めた、群内の距離、群間の距離の平均値を示した。表3から、3人の群内の距離の平均値が最小の群間の距離の平均値より小さいことがわかる。この結果から、日本文の場合でも、単語の長さの分布にも著者の特徴が現れると判断する。

欧米の研究では、文章の著者の推定や文章の分類を行う際、単語の長さの分布を品詞別に分けずに用いている。しかし、単語の長さは文章の内容によって変わりやすいのもあれば、書く内容にあまり影響を受けないのもあると考えられる。文章の著者の推定に用いる情報としては、文章の内容に依存しないことが望まれる。単語の長さの分布から見た場合、どのような単語に著者の特徴が現れ、どのような単語には著者の特徴が現れないかを実証するため、単語を品詞別に分けて比較分析を行うことにする。表4に使用頻度が割に高い

表3 単語の長さの分布における距離の平均値

著者名	群内	群 井上	三島	間 中島	最小の群間
井上	0.1433		0.1713	0.3150	0.1713
三島	0.1360	0.1713		0.5305	0.1713
中島	0.0641	0.3150	0.5304		0.3150

名詞、動詞、形容詞、形容動詞、助詞、助動詞、副詞の品詞別の単語の長さの分布の群内、群間の距離の平均値を示した。ただし、名詞は長さ1から6までの6変数、動詞は長さ1から5までの5変数、助詞、助動詞は長さ1から3までの3変数、形容詞、形容動詞、副詞は長さ2から5までの4変数を用いた。

品詞別に分けていない（以下は“すべての単語”と呼ぶ）場合と、名詞、動詞、形容動詞の場合は、群内の距離の平均値が最小の群間の距離の平均値より小さいが、形容詞、助詞、助動詞、副詞の場合は群内の距離の平均値が最小の群間の距離の平均値より大きいものもある。著者の特徴情報としては、すべての単語を用いた方がよいであろうか、それともどれかの品詞の単語の長さの分布を用いた方がよいであろうか。つまり、すべての単語

と品詞別に分けた場合の単語の長さの分布のなか、どの情報に著者の特徴がもっとも明確に現れているかについてを評価する必要がある。

3.2 特徴の明確さの評価

複数の情報の中でどの情報に著者の特徴がもっとも明確に現されているであろうか？著者別に文章を分類する観点からは、すべての著者に対して群内の距離の平均値が最小の群間の距離の平均値より小さければ小さいほど分類がうまくいくと考えられるため、群内の距離の平均値と最小の群間の距離の平均値との差が一つの目安となる。そこで、群内の距離の平均値と最小の群間の距離の平均値との差を以下の分散比および相関比で評価することにする。

いま、著者 k の群内での距離の平均値を

表4 品詞別の単語の長さの分布における距離の平均値

品詞名	著者名	群内	群	間	最小の群間	
名詞	井上	0.9191		1.0072	1.3229	1.0072
	三島	0.9730	1.0072		1.8449	1.0072
	中島	0.4000	1.3229	1.8449		1.3229
動詞	井上	0.1535		0.6525	0.4794	0.4794
	三島	0.4597	0.6525		0.7551	0.6525
	中島	0.1942	0.4794	0.7551		0.4794
形容詞	井上	0.2556		2.2920	2.8630	2.2920
	三島	2.1359	2.2920		1.5212	1.5212
	中島	0.9958	2.8629	1.5212		1.5212
形容動詞	井上	1.1148		2.8472	2.3553	2.3553
	三島	1.5683	2.8472		5.3052	2.8472
	中島	1.7552	2.3553	5.3052		2.3553
助詞	井上	0.0727		0.1454	0.0701	0.0701
	三島	0.2271	0.1454		0.1215	0.1215
	中島	0.0669	0.0701	0.1215		0.0701
助動詞	井上	0.2147		0.1917	0.7710	0.1917
	三島	0.1986	0.1917		0.7446	0.1917
	中島	0.3074	0.7710	0.7446		0.7446
副詞	井上	1.8567		1.6211	3.2165	1.6211
	三島	0.9355	1.6211		3.8723	1.6211
	中島	0.4450	3.2165	3.8723		3.2165

$y_{k1} = \overline{d(k)}$, 最小の群間の距離の平均値を
 $y_{k2} = \min\{\overline{d(k,1)}, \overline{d(k,2)}, \dots, \overline{d(k,h)} \dots, \overline{d(k,g)}\}_{k+h}$ で表すと, g 人の著者の群内の距離の平均値, 最小の群間の距離の平均値はマトリックス $Y_{g \times 2}$ で表示できる.

$$Y_{g \times 2} = \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \\ \vdots & \vdots \\ y_{k1} & y_{k2} \\ \vdots & \vdots \\ y_{g1} & y_{g2} \end{bmatrix}$$

$Y_{g \times 2}$ の総変動

$$S = \sum_{k=1}^g \sum_{b=1}^2 (y_{kb} - \bar{y})^2$$

は, 以下のように分類できる.

$$S = 2 \sum_{k=1}^g (y_{k\cdot} - \bar{y})^2 + g \sum_{b=1}^2 (y_{\cdot b} - \bar{y})^2 + \sum_{k=1}^g \sum_{b=1}^2 (y_{kb} - y_{k\cdot} - y_{\cdot b} + \bar{y})^2$$

ただし

$$y_{\cdot b} = \frac{\sum_{k=1}^g y_{kb}}{g}, \quad y_{k\cdot} = \frac{\sum_{b=1}^2 y_{kb}}{2}$$

$$\bar{y} = \frac{\sum_{k=1}^g \sum_{b=1}^2 y_{kb}}{2g}$$

である.

右辺の第一項は行間 (著者間) の変動で, 第二項は列間 (群内の距離と最小の群間の距離との間) の変動で, 第三項は誤差変動である. それぞれを SS_A, SS_B, SS_E で表記する.

この行間の変動 SS_A , 列間の変動 SS_B , 誤差変動 SS_E の自由度はそれぞれ $(g-1), (2-1), (g-1)(2-1)$ であるので, 自由度で基準化した変動をそれぞれ S_A^2, S_B^2, S_E^2 (分散) で表す.

$$S_A^2 = \frac{SS_A}{g-1} = \frac{2 \sum_{k=1}^g (y_{k\cdot} - \bar{y})^2}{g-1}$$

$$S_B^2 = SS_B = g \sum_{b=1}^2 (y_{\cdot b} + \bar{y})^2$$

$$S_E^2 = \frac{SS_E}{(g-1)} = \frac{\sum_{k=1}^g \sum_{b=1}^2 (y_{kb} - y_{k\cdot} - y_{\cdot b} + \bar{y})^2}{(g-1)}$$

$$S_E^2 = S_A^2 + S_B^2 + S_E^2$$

行間の平均の差, 列間の平均の差はそれぞれの分散比

$$\frac{S_A^2}{S_E^2}, \quad \frac{S_B^2}{S_E^2}$$

および相関比

$$\frac{S_A^2}{S_T^2}, \quad \frac{S_B^2}{S_T^2}$$

で評価すれば良い. 分散比および相関比の値が大きければ大きいほど行間, 列間の差が大きいことを意味する.

我々が興味をもっているのは群内の距離の平均値と最小の群間の距離の平均値との差であるため, 分散比 S_B^2/S_E^2 および相関比 S_B^2/S_T^2 で評価すればよい. 上記の分散比・相関比が大きければ著者の特徴が明確であることを意味する. 表5にすべての単語を用いた場合と品詞別に分けた場合の S_B^2/S_E^2 および S_B^2/S_T^2 を示した.

分散比 S_B^2/S_E^2 および相関比 S_B^2/S_T^2 がともに最も大きいのは動詞である. したがって, 動詞の長さの分布に著者の特徴が最も明確に現れていると考える.

3.3 文章の分類

前節ではミクロの観点から分散比・相関比を用いてどのような情報に著者の特徴がもっとも明確に現れているかについて分析を行っ

表5 単語の長さの分布における分散比・相関比

	すべての単語	名詞	動詞	形容詞
S_B^2/S_E^2	2.06	1.47	46.25	0.42
S_B^2/S_T^2	0.65	0.58	0.75	0.21
	形動	助詞	助動詞	副詞
S_B^2/S_E^2	22.32	0.98	0.81	1.46
S_B^2/S_T^2	0.55	0.17	0.21	0.56

た。本節では、その分析結果を検証するため、文章における単語の長さの分布を用いて文章を分類する場合、文章が著者別に分類されるかどうかの視点から（マクロの観点から）著者の特徴について比較分析を行う。

分類分析にはいくつかの方法があるが、本研究では主成分分析方法を用いる。分類分析を行う際の主成分分析としては、データ

$$X_{I \times J} = [x_{ij}]$$

$$P_{I \times J} = [p_{ij}]$$

$$p_{ij} = \frac{x_{ij}}{\sum_{j=1}^J x_{ij}}, \quad \sum_{j=1}^J p_{ij} = 1.$$

を用いて主成分分析を行うのが一般的であるが、本研究では、 $P_{I \times J}$ を用いて分布の間の距離を求め、距離のマトリックスについて主成分分析を行う方法を用いる。具体的な手順としては、求められた距離のマトリックス

$$D_{I \times I} = \begin{bmatrix} 0 & d_{ij} \\ d_{ij} & 0 \end{bmatrix}$$

を以下のように標準化し、

$$\widehat{d}_{ij} = d_{ij} - \overline{d_{i.}} - \overline{d_{.j}} + \overline{d_{..}}$$

$$\overline{d_{i.}} = \frac{1}{I} \sum_{v=1}^I d_{iv}, \quad \overline{d_{.j}} = \frac{1}{I} \sum_{v=1}^I d_{vj},$$

$$\overline{d_{..}} = \frac{1}{I \times (I-1)} \sum_{v=1}^I \sum_{v_2=1}^I d_{vv_2}$$

$$\widehat{D}_{I \times I} = [\widehat{d}_{ij}]$$

$\widehat{D}_{I \times I}$ の分散共分散の行列を用いて主成分分析を行う。まず、すべての単語の長さの分析を用いた主成分分析を行った。第1主成分、第2主成分の寄与率はそれぞれ91.19%、7.49%で、第2主成分までの累積寄与率は97.68%である。図1に主成分得点のプロットを示した。比較のため分散比および相関比とともに最も大きい動詞の長さの分布を用いた場合の主成分分析を行った。第1主成分、第2主成分の寄与率はそれぞれ57.04%、34.04%で、第2主成分までの累積寄与率は91.08%である。第2主成分までの主成分得点のプロットを図2に示した。主成分分析の結果では、すべての単語の長さの分布を用いた場合は文章が著者別にきれいに分けられない

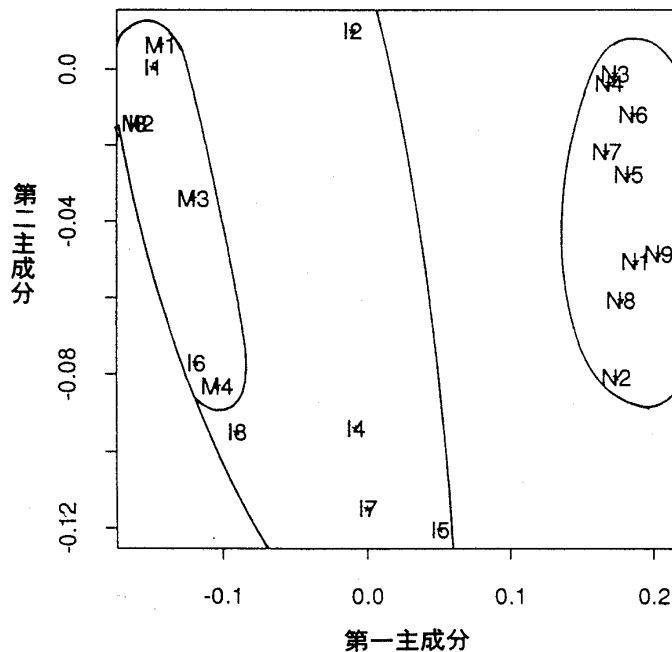


図1 すべての単語の長さの分布を用いた主成分得点のプロット

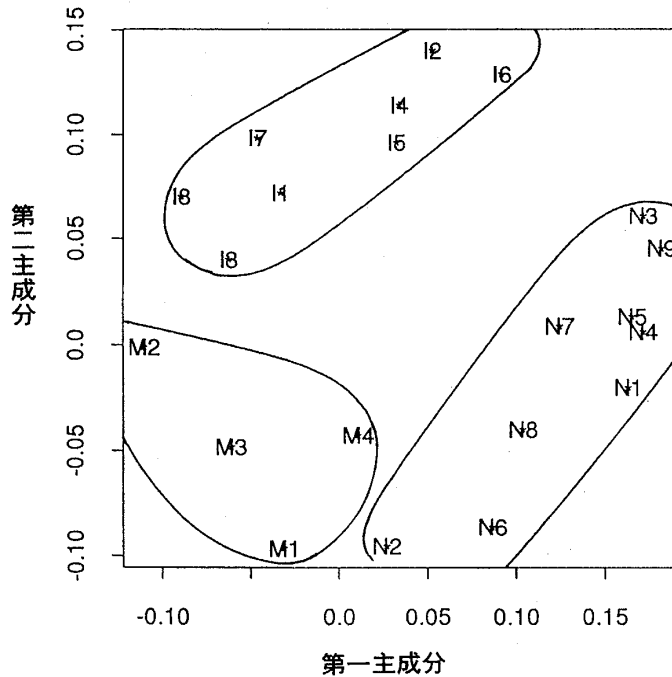


図2 動詞の長さの分布を用いた主成分得点のプロット

が、動詞の単語の長さの分布を用いた場合はきれいに分けられることがわかる。この結果は、前節の分散比・相関比の分析結果と一致する。したがって、単語の長さの分布を用いて文章の著者の推定や著者別に文章を分類するなどの研究では、単語を品詞別に分け、著者の特徴が出やすい品詞の単語の長さの分布を用いるべきである。

4. むすびにかえて

本研究では三人の作家の21編の文章を用いて、日本文における単語の長さの分布には著者の特徴が現れるか否か、その情報を用いて、著者別に文章を分類する場合、文章が著者別に分類できるかどうかについての分析を行った。また、より質のよい情報の抽出の方法を提案するため、単語の長さの分布を品詞別に分けて比較分析も行った。その結果、日本語の場合でも、単語の長さの分布に著者の特徴が現れることが明らかにされた。また、単語を品詞別に分けていない場合には、単語の長さを書き手の個性が出にくい助詞、文章の内容への依存性が高い名詞など（ノイズ）

が含まれてしまうため、著者の特徴が明確に現れないが、品詞別に分けることによって著者の特徴がより明確に現れることがわかった。最も著者の特徴が明確に現れるのは、動詞の長さの分布である。

動詞の長さの分布に著者の特徴が現れることについては、和語・漢語、合成語・非合成語の相対使用頻度の影響ではないかとも考えられる。そこで、動詞のなかの和語・漢語、合成語・非合成語の相対使用頻度と動詞の長さの分布との関係について分析を行った。結果としては、分析に用いた文章においては、動詞の長さの分布に著者の特徴が現れるのは、和語・漢語、合成語・非合成語の使用率によるものではないことが確認された。参考文献(16)にその詳細を述べた。

単語を品詞別に分けることによって、より質のよい情報が得られることは、日本語だけではなく、他の言語も同様な性質を持っていると推測される。

本研究の延長として、より多くの著者と文章について分析を行い、本研究の結果の普遍性を確認することと、動詞の長さの分布に著

者の特徴が最も明確に現れる現象の裏付けとして、言語学、心理学の立場からの調査分析が必要であると考えられる。

謝辞 本研究に用いたデータベースは、統計数理研究所及び総合研究大学院大学の村上征勝教授の研究費で作成したものである。本研究をご指導及びご支援くださった村上教授に心より感謝致します。

注

(注1) 本稿は、参考文献(13)、(14)を書き改めたものである。

(注2) 旺文社、第八版。

参考文献

- (1) A.Q. Morton: *The Aauthorship of Greek Prose*. J.R. Statist. Soc., A, 128 (2), pp.169-233 (1965).
- (2) C.B. Willams: *A Note on the Statistical Analysis of Sentence-Length as a Criterion of Literary style*, Biometrika, 31, pp.356-361 (1940).
- (3) C.B. Willams: *Studies in the History of Probability and Statistics*, Biometrika, 43, pp. 248-256 (1956).
- (4) C.B. Willams: *Mendenhall's Studies of Word-length Distrbution in the Works of Shakespeare and Bacon*, Biometrika, 62, pp. 207-211 (1975).
- (5) C.S. Brinegar: *Mark Twain and the Quintus Curtius Snodgrass Letter*, J. Amer. Stat. Ass., 58, pp.85-96, (1963).
- (6) D.R. Cox and L. Brandwood: *On a Discriminatory Problem Connected with the Words of Plato*. J.R. Statist. Soc. B, 21, pp. 195-200 (1958).
- (7) F. Mosteller, D.L. Wallace: *Inference in an Authorship Problem*. J. Amer. Stat. Ass., 58, pp.275-309 (1964).
- (8) G. Herdan: *The Relation Between the Dictionary Distribution and the Occrrence Distribution of Word Length and Its Importance for the Study of Quantitative Linguistics*, Biometrika, 45, pp.222-228 (1958).
- (9) G.U. Yule: *On Sentence-Length as a Statistical Charcteristic of Style in Prose: With Application to two Cases of Disputed Authorship*, Biometrika, 30, pp.363-390 (1939).
- (10) M.Z. Jin and M. Murakami (金・村上): *Authors' Characterristic Writing Styles as Seen Through Their Use of Commas*, Behaviormetrika, Vol.20, pp.63-76 (1993).
- (11) 金 明哲, 樺島忠夫, 村上征勝: 読点と書き手の個性, 計量国語学, Vol.18, No.8, pp.382-391 (1993).
- (12) 金 明哲, 樺島忠夫, 村上征勝: 手書きとワープロによる文章の計量分析, 計量国語学, Vol.19, No.3, pp.133-145 (1993).
- (13) 金 明哲: 単語の長さの分布による文章の分類と分類の良さの評価, Research Memorandum No.497, The Institute of Statistical Mathematics, Tokyo (1994).
- (14) 金 明哲: 単語の長さの分布による文章のパターン分類, 情報処理学会第48回全国大会講演論文集1 (1994).
- (15) 金 明哲: 読点の打ち方と文章の分類, 計量国語学, Vol.19, No.17, pp.317-383 (1994).
- (16) 金 明哲: 動詞の長さの分布に基づいた文章の分類と和語および合成語の比率, 自然言語処理, Vol.2, No.1, pp.57-75 (1995).
- (17) 村上征勝: 計量的文体研究の威力と成果, 言語, Vol.23, No.2, pp.30-37 (1994).
- (18) 中野 洋: パソコンによる語の認定処理, 国立国語研究所研究報告集, Vol.12, pp.83-131 (1991).
- (19) 葦沢 正: 由良物語の著者の統計的判別, 計量国語学, No.33, pp.21-28 (1965).
- (20) R. Thisted, B. Efron: *Did Shakespeare Write a Newly-Discovered Poem?*, Biometrika, 74, pp.445-455 (1987).
- (21) 佐々木和枝: 文の長さの分布型, 計量国語学, No.78, pp.13-22 (1976).

- (22) W. Fucks: *On mathematical Analysis of Style*, *Biometrika*, 39, pp.122-129 (1952).
- (23) W. Fucks: *On Nahordnung and Fernordnung in Samples of Literary Texts*, *Biometrika*, 41, pp.116-132 (1954).
- (24) W.C. Wake: *Sentence-Length Distribution of Greek Authors*, *J.R. Statist.Soc., A*, 20, pp. 331-346 (1957).
- (25) 安本美典：文の長さの分布型, 計量国語学, No.1, pp.20-30 (1957).
- (26) 安本美典：文体統計各種分布型, 計量国語学, No.2, pp.20-24 (1957).

1995年12月26日受付

1996年2月13日受理