

文字冗長度による日本文学分析

Analysis of Japanese literature by redundancy

福田 宏、 山下泰弘*、勝矢光昭

Hiroshi Fukuda, Yasuhiro Yamashita, Mitsuaki Katsuya

静岡県立大学経営情報学部、〒422-8526 静岡市谷田 52-1

The university of Shizuoka, Yada 52-1, Shizuoka 422-8526

* 科学技術振興事業団、〒332-0012 川口市本町 4-1-8 川口センタービル

Japan Science and Technology Corporation, Honmachi4-1-8, Kawaguchi 332-0012

あらまし：コンピュータの発達により、統計処理に基づく計量分析の道が文学作品に対して切り開かれた。そこでは、品詞の出現確率や、異なり語の情報を用いて、わが国の代表的な古典文学、「源氏物語」と「手枕」、「雲隠れ六帖」、「山路の霧」等の文体比較が検討されている。更に「源氏物語」における助詞や助動詞の種類とそれらの出現確率が統計数理的に解析され、文学作品に対して計量分析研究分野が確立しようとしている。

本論文では文学作品に対し、情報理論の基本的な概念、エントロピーあるいは冗長度による分析を試みる。これらエントロピーあるいは冗長度等の経年変化や作家による比較等を議論し、これらが文学作品分析の上で興味深い指標となる可能性を指摘する。

Summary: The study of the quantitative analysis of Japanese literature by using the appearance probabilities of the verbs and the auxiliary verbs has been established. In this paper, we point out that the entropy and the redundancy are the interesting indices of the literature in

order to study the secular changes of the Japanese used in the literature and compare the literature works written by the different writers.

キーワード：エントロピー、冗長度、日本文学、計量分析、源氏物語

Keywords: entropy, redundancy, quantitative analysis, Japanese literature, Genji Monogatari

1 はじめに

シャノン¹によるエントロピー、相対エントロピー、冗長度等の概念の認識以来、人間の情報伝達手段として用いられる自然言語のエントロピー（自然言語一文字が持つ平均情報量）を評価する試みが多くの研究者によりなされてきている。自然言語を以下の観点から分析することは、自然科学の観点から

1 種々の文化圏で独自に発達した自然言語の一文字が持つ情報量や冗長度の比較

2 種々の作家により書かれた文学作品中の一文字が持つ情報量や冗長度の比較

3 自然言語の一字が持つ情報量や冗長さの経年変化

自然言語は典型的な記憶情報源である。この性質を考慮して、上記の観点を議論するには、一定長の文字列の情報量や冗長さから一字のそれらを求めなければならない。具体的には、N 続きの文字列の平均情報量と (N-1) 続きの文字列の平均情報量を計算し、始めの量から、後の量を引くことにより、一字当たりの平均情報量 (エントロピー) を求めることになる。そして、この作業を種々の N について実行し、慎重に結論を下さなければならない。

このようにして、自然言語における一字当たりの平均情報量を評価しようとすれば、膨大な量のデータが必要となる。特に日本語で使われる文字のように、その種類が多い自然言語では、量的に定まっている過去の文学作品や、一人の作家の文学作品から信頼できる一字当たりの平均情報量 (エントロピー) を求めることは容易ではない。

日本語事象系に対して英語事象系では、もしも大文字と小文字を区別しなければ、文字種は 26 で足りる。この理由のために、英語一字の持つ情報量や冗長度を求めるには日本語ほど、大量のデータを必要としない。

例えば、Brown 達²は 1.75bits を英語一字のもつ情報量の上限として 1992 年に発表している。Brown 達の計算方法を採用して、Han 達³は韓国語について同様の分析をおこない、6.01bits を韓国語一字のもつ情報量の上限として 1996 年に発表している。これらの値から、その自然言語データを圧縮する際の理論的限界が導かれる。

情報科学の立場では、一字のもつ情報量は非常に興味深い量であるが、人文科学の立場では、これらの値が単独に扱われるだけでは全く無意味であり、現在に至るまで、情報理論に基づく文学の分析は、注目されることはなかった。

しかしながら、日本の古典文学に対して、助詞や助動詞の種類とそれらの出現確率が統計数理的に

解析され、計量分析研究分野が確立しようとしている現在、情報理論に基づく指標を用いて日本文学の分析を試みることは興味深く思われる。

2 文字エントロピーと言語エントロピー

本論文では無記憶情報源の極限として日本語事象系を考察する。日本語に限らず自然言語は典型的な記憶情報源 (マルコフ過程) である。従って無記憶情報源の極限で得られるエントロピーは言語の正しいエントロピーとは言えない。そのためここではそれを文字エントロピーと呼び、言語エントロピーと区別する。そして文字エントロピーから導かれる相対エントロピー、あるいは冗長さの経年変化や作家による比較等を議論する。

文字エントロピーの定義式は以下の通りであり、以下で述べるように言語エントロピーの第一次近似とみなせる。

$$H_0 = -\sum_i p_i \log p_i \quad (1)$$

ここで p_i は i 番目の事象の発生確率を示す。対数の底は 2 を採用した場合、bit 単位の値になる。情報理論的には「文字エントロピーは自然言語エントロピーの十分安全な上限値である」ことを主張できる。著者達はその上更に文字エントロピーから導かれる相対エントロピー、あるいは冗長度が人文科学における新しい現実的な計量指標となる可能性を期待している。著者達は本論文でこのことを実際に分析し確認したい。

コンピュータの発達により、統計数理の立場から文学作品の計量分析の道が切り開かれた。⁵ そこでは、品詞の出現確率や、異なり語の情報を用いて、わが国の代表的な古典文学、「源氏物語」と「手枕」「雲隠六帖」「山路の露」の文体の比較が検討されている。さらに、「源氏物語」における助詞や助動詞の種類とその出現確率が統計数理的に解析され、「源氏物語」のような文学作品の計量分析研究分野が確立しようとしている。計量分析では統計学に基

づいて注目している品詞の出現頻度等の比較・分析・検定をしている。しかしながら、注目する統計量の取捨選択は、多分に主観的になることは避けられない。

文字エントロピーは情報理論において、明確に定義された平均情報量と密接に関係している統計量である。文字エントロピーは文字の意味は問わないで、単純に文字の出現確率にだけ注目し、文字の情報量の平均値として抽出される客観的な量である。文字エントロピーや、それから導かれる相対エントロピーあるいは冗長度等の量は、現実的な客観的なデータ量であり冒頭に述べた三つの観点の分析を可能にする。このため、これまで進められてきた文学作品に対する計量分析研究の相補的な側面を持つものであると考えられる。

次に言語のエントロピーについて解説する。言語のエントロピーを計算する方法はシャノンにより以下のように定式化された。先ず N-gram エントロピーを以下のように定義する。

$$F_N = - \sum_{i,j} p(b_i, j) \log p_{b_i}(j) \quad (2)$$

ここで b_i は N-1 続きの文字列を意味する。 j は b_i に続く任意の文字である。 $p(b_i, j)$ は、N 続きの文字列 (b_i, j) の出現確率を表わす。 $p_{b_i}(j)$ は

b_i 文字列の次に文字 j が出現する条件付確率である。条件付確率には以下の (3) 式が成立する。

$$p_{b_i}(j) = \frac{p(b_i, j)}{p(b_i)} \quad (3)$$

(2) 式は (3) 式の性質を使えば、以下のように表される。

$$F_N = - \sum_{i,j} p(b_i, j) \log p_{b_i}(j) + \sum_i p(b_i) \log p(b_i) \quad (4)$$

上記 (4) 式は、一文字の平均情報量は、N 続きの文字列の平均情報量から、(N-1) 続きの文字列の

平均情報量を引いたもので定義されることを示している。

小さい N に対しては、この方法で一文字の持つ平均情報量の計算は可能であるが、大きい N に対しては、特に日本語のように文字種が多い言語ではこれを評価するには膨大なデータを必要とし、現実的ではない。

本論文で著者達が提案している文字エントロピーは一般式 (2) の $N=1$ に相当するものであり、自然言語のエントロピーとは言えないが、これと密接に関係した客観的な計量分析に利用できる量であり、著者達はこの論文で「文字エントロピー」と呼ぶことにする。

3 相対エントロピーと冗長度

事象系の事象数が増加すれば、文字エントロピーは増加し、事象に変化がない同じ事象系を長年にわたり使い続けられれば、文字エントロピーは経年により減少することは十分予想ができる。事象系の事象数は、経年により変化する可能性があるし、同じ時代の文学作品であっても、その中で使われる文字種は必ずしも同じではない。従って、文字エントロピーを指標とするよりも、以下で定義される相対エントロピーあるいは冗長度の方が指標として適しているとも考えられる。

$$R = \frac{- \sum_i p_i \log p_i}{\log n} \quad (5)$$

ここで、 n は事象の数を意味し、 $\log n$ はこの事象系の最大エントロピーである。この量 R は現実の文字エントロピーの事象系が本来持つことができる最大文字エントロピーに対する割合である。情報科学の観点からは、この量が経年によりどのように変化するか、興味あるところである。

情報理論的に全く無駄のない事象系では、相対エントロピー R は 1 である。無駄が多ければ R は 0 に近づく。そこで、冗長度として以下を定義する。

$$\eta = 1 - R \quad (6)$$

この冗長度は、無駄が多ければ1に、無駄が少なければ0に近づく量である。本論文では文字エントロピー、相対エントロピーあるいは冗長度を指標として約1000年昔の文学作品、「源氏物語」と近・現代文学作品の分析を試みる。

4 データベース

冒頭に述べた3つの観点を議論するための研究対象は、できるだけ古い時代から最近迄の文学作品が望ましい。そこで研究の手始めとして、著者達は古い文学作品の代表として、紫式部が書いたとされる「源氏物語」、そして近・現代文学作品の代表として、芥川龍之介の「羅生門」、川端康成の「雪国」、夏目漱石の「こころ」、松本清長の「点と線」、三島由起夫の「金閣寺」等を選んでデータベースを構築した。本論文に対する評価が得られれば、更に、途中の年代の文学作品をデータベース化し、本研究を進展させたいと考えているが、研究の緒についた現段階では、最古の作品と現代の作品の比較・分析に留めた。

「源氏物語」に出現する文字種や文字数を統計的に分析するため、「源氏物語」54巻すべてをパソコンで統計分析が可能のようにテキストファイルとしてデータベース化し、これを用いて本論文では分析をしている。データベースの作成は、静岡県立大学経営情報学部・情報科学研究室が独自に行った。データベースの作成に用いた本文は、池田亀鑑編著「源氏物語大成」(中央公論社)である。これの分析に際しては、村上・今西による「源氏物語の助動詞の計量分析」⁶でなされたように54巻を4つのグループに分けて検討しなかった。データベースには空白データがたくさん存在するが、本研究では改行データはカウントし、段落が終わり新しい段落が始まるまでの空白はカウントしないことにした。

近・現代文学作品として、芥川龍之介の「羅生門」、川端康成の「雪国」、夏目漱石の「こころ」、松本清

長の「点と線」、三島由起夫の「金閣寺」等を選んで、統計的な分析が可能のようにテキストファイルによるデータベースを静岡県立大学経営情報学部・情報科学研究室で構築した。これに用いた本文は、CD-ROM版の「新潮文庫の100冊」である。現在はインターネット上で入手できる文学作品のデータベース化作業を進めている。

5 情報理論に基づく日本文学分析

表1に近・現代文学作品に対する文字エントロピー、冗長度等データを示す。

表1 近・現代作品の文字エントロピーと冗長度

作品名	H_0	η
こころ	7.33	0.335
金閣寺	7.60	0.331
人間失格	7.32	0.321
雪国	7.27	0.322
点と線	7.36	0.313

文字エントロピーでは、例えば三島由起夫の金閣寺のように突出した値も存在するが、冗長度(相対エントロピー)で比較すると、はほぼ他の作品と同じ値になることは興味深い。

表2には、源氏物語における各帖毎の同様の量を示す。図1、2は、これらのデータを図に表したもので、これから分かるように、文字エントロピーの推移は殆ど無秩序であり、所々に出現する極端な文字エントロピーの値は気になる所であるが、源氏物語の各帖の文字数が文字エントロピーの計算には不十分なために生じたものと考えられる。この図から、源氏物語の文字エントロピーは大体5.5から5.8ビットという非常に狭い範囲に収まることが分かる。上記各帖の文字エントロピー、冗長度を独立量とし、箱髭図で示したのが、図3、4である。その右に、近・現代文学作品の代表作(こころ、金閣寺、人間失格、雪国、点と線)によるデータも示した。源氏物語と近・現代文学作品の間には約1000年の経年があり、図から明らかなように、文字エントロピーはこの間、約1.3倍程度増加している。

表1 「源氏物語」における文字エントロピーと冗長度一覧

帖番号	帖名	H_0	η
1	きりつほ	5.78	0.238
2	はは木	5.65	0.244
3	うつせみ	5.55	0.161
4	夕かほ	5.65	0.254
5	わかむらさき	5.67	0.236
6	すえつむ木	5.60	0.195
7	もみちの賀	5.71	0.242
8	花のえん	5.76	0.189
9	あふひ	5.74	0.246
10	さか木	5.75	0.263
11	花ちるさと	5.45	0.125
12	すま	5.72	0.260
13	あかし	5.84	0.265
14	みをつくし	5.72	0.240
15	よもきふ	5.60	0.205
16	せきや	5.59	0.142
17	えあはせ	5.71	0.212
18	松かせ	5.63	0.213
19	うす雲	5.69	0.219
20	あさかほ	5.57	0.181
21	をとめ	5.74	0.265
22	玉かつら	5.69	0.257
23	初音	5.61	0.187
24	こてふ	5.59	0.177
25	ほたる	5.56	0.182
26	とこなつ	5.56	0.178
27	かかり火	5.61	0.139

帖番号	帖名	H_0	η
28	野わき	5.63	0.179
29	みゆき	5.62	0.225
30	藤はかま	5.68	0.194
31	まきはしら	5.68	0.238
32	梅かえ	5.71	0.237
33	藤のうらは	5.81	0.254
34	わかな上	5.77	0.297
35	わかな下	5.71	0.275
36	かしは木	5.60	0.196
37	よこ笛	5.65	0.193
38	すすむし	5.74	0.220
39	夕きり	5.63	0.244
40	みのり	5.69	0.207
41	まほろし	5.77	0.239
42	にほふ兵部卿	5.93	0.226
43	こうはい	5.79	0.213
44	たけ川	5.79	0.246
45	はし姫	5.66	0.239
46	しいかもと	5.73	0.246
47	あけまき	5.64	0.251
48	さわらひ	5.60	0.163
49	やとり木	5.72	0.279
50	あつまや	5.73	0.268
51	うき舟	5.61	0.235
52	かけろふ	5.62	0.243
53	てならひ	5.78	0.284
54	夢のうきはし	5.56	0.168

図1：源氏物語におけるエントロピーの移り変わり

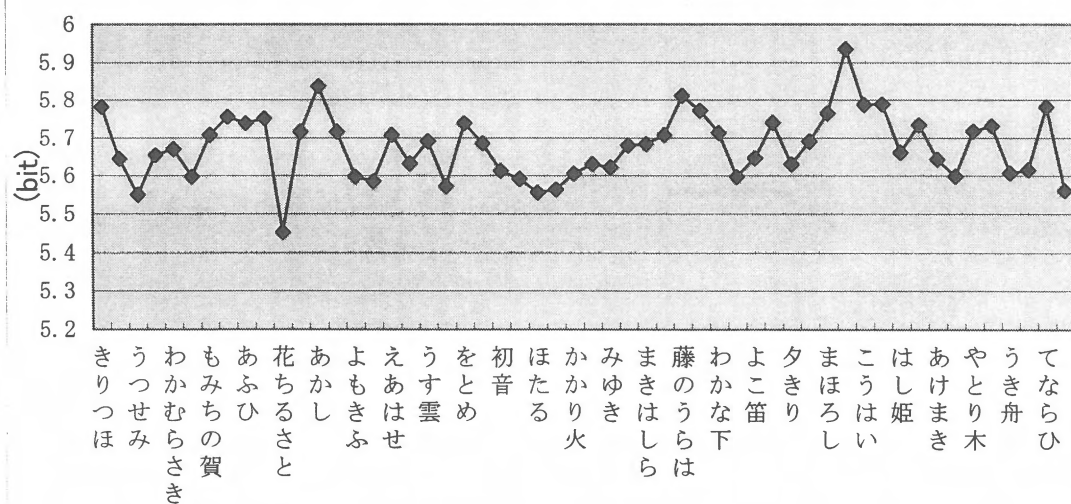
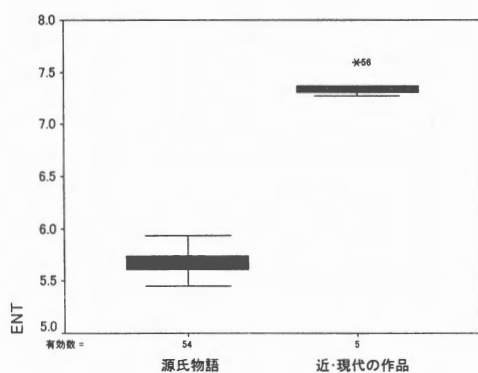
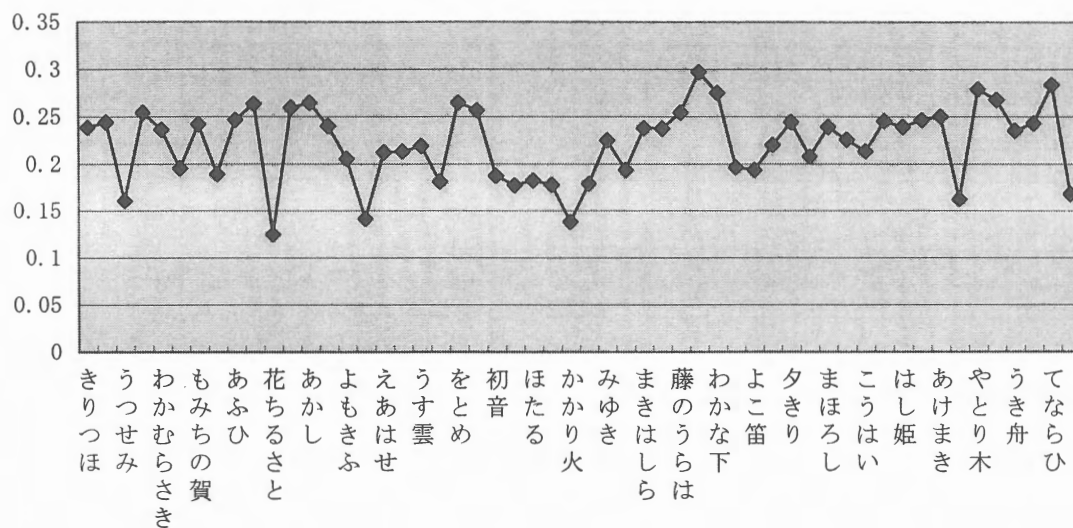
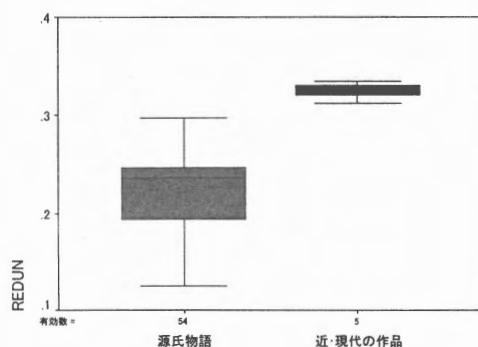


図2：源氏物語における冗長さの移り変わり



作品

図3 文字エントロピー箱髭図



作品

図4 冗長さ箱髭図

参考文献

- 1 C.E.Shannon. The Mathematical Theory of Communication Bell System Technical Journal, Vol.27(1948),pp.379-423; ibid pp.623-656.
- 2 P.F.Brown, S.A.D.Pietra, V.J.D. Pietra, J.C. Lai, and R.L.MercerAn. Estimation of an Upper Bound for the Entropy of English. Computational Linguistics, Vol.18 (1992),pp. 31-40,
- 3 Y. S. Hanm H. R. Park, J. H. Shin and K. S. Choi. An Upper Bound Estimate for the Entropy of Korean. Literary and Lingustic Computing Vol.1(1996), pp. 141-146.
- 4 村上征勝、今西祐一郎 源氏物語の助動詞の計量分析 情報処理学会論文誌 第40巻 (1999)pp.774-782.