百人一首のスペトル解析およびヒストグラム解析

小林恒夫

福島県立医科大学医学部物理学講座

Spectrum and Histogram Analysis for Hyakunin Isshu (Single Songs of a Hundred Poets)

Tsuneo KOBAYASHI

Department of Physics, Fukushima Medical University School of Medicine

1. はじめに

筆者は、和歌等の日本語のひらがな表記に番号付けを行い、得られる数値列を時系列、すなわち時間的に刻々と変化してゆくデータとみなし、解析を行う手法を思い付いた [1]。いったん数値に変換してしまえば、統計解析・時系列解析など既存の数理的解析方法が使える。日本語研究の一方法を標榜する次第である。

前論文 [1] では、時系列解析の一種であるスペクトル解析 [2] を行い、1/f ゆらぎ [3]（エフぶんのいちゆらぎと読む、英語では one–over–f fluctuation）となっている和歌等が多いことを報告した。

前論文で、百人一首として無作為に選んだ 10 首のうちの 100% が 1/f ゆらぎを示していた。そこで、今回は百人一首のすべてについて解析を行い、まず、1/f ゆらぎかどうかの判定を行った。また、現れる文字に周期性があるかどうかを、自己回帰モデルにより探索した。次に、一首一首についてヒストグラムをつくり、正規分布とみなせるかどうかの正規性検定を試みた。最後に、百人一首の歌詞すべてを並べてみると 3134 文字となるが、この 3134 文字について、ヒストグラム解析等を行った。

2. 方法

百人一首の文献は極めて多いが、ここでは、大岡信著「百人一首 [4]」（講談社文庫）および新国語例解辞典付録「全訳小倉百人一首 [5]」（小学館）を参考にし、ひらがなデータを作成した。いろは歌および五十音図の 2 種類を基準として和歌の歌詞のひらがな読みに番号付けを行い [1]、得られる数値列を時系列とみなし、スペクトル解析およびヒストグラム解析を行った。解析には S-PLUS 6.2J を使用した [6]。

一首一首の和歌については、標本自己相関関数、標本偏自己相関関数により、数値列が定常時系列とみなせるかどうかをしらべた。また、標本自己相関関数をフーリエ変換したピリオドグラムを平滑化して得られるパワースペクトルを、1/f ゆらぎの判定に使用した。周期性を探索するために、自己回帰モデルによるパワースペクトルも求めた。さらに、使われているひらがなの異なるものの数をしらべた。最後に、

ヒストグラムを書き、シャピロ・ウィルクの正規性検定[7]を行った。

百人一首の歌詞すべてを並べた3134文字については、使われているひらがなの数と、ヒストグラム解析を行った。

なお、奈良時代には、現在の読みとは異なる上代特殊仮名遣いが使用されていたとのことであるが[1]、本論文では現代の人々の愛好しているひらがな読み―ただし、歴史的仮名遣いによる―について解析を行った。

3. 結果と考察

3.1. 定常性

標本自己相関関数と標本偏自己相関関数をプロットしてみると、いずれにおいても95%信頼区間に収まっている場合が多く、ほとんどの和歌について、番号付けによる時系列が定常時系列とみなせることがわかった。

3.2. 1/f ゆらぎ

1/f ゆらぎの可能性をみるために、ピリオドグラムを平滑化したパワースペクトル密度と周波数の両対数グラフの傾きを求めた[1]。ここでは、傾きが –0.8～–1.2 となるものを 1/f ゆらぎと判定した。

例として、百人一首第一番「あきのたのかりほのいほのとまをあらみわがころもではつゆにぬれつつ」を解析してみると、最小2乗法でフィットした直線の傾きは、いろは歌基準で –0.70 ± 0.27、五十音図基準で –0.93 ± 0.23 となった。したがって、この和歌は誤差を考慮すれば 1/f ゆらぎを示しているといえる。同様な解析を100首について行った。

百人一首100首のうち 1/f ゆらぎと判定されたものは、いろは歌を基準とした場合は88首、五十音図を基準とした場合は95首と、大部分のものが 1/f ゆらぎを示した。

万人に数百年もの長きにわたり語り継がれ、愛唱されてきた百人一首である。無意識のうちに集められた和歌たちが、1/f ゆらぎという、人にとって心地よいリズムを多く持っているというのはまことに理に叶った結論といえる。

3.3. 自己回帰モデル

ひらがなとして出現する文字の周期性に何らかの特徴があるかどうかをみるため、自己回帰モデル(AR)によりパワースペクトルを計算した(前節のピリオドグラムではスペクトルの特徴的なピークを見いだすのが困難であることが多い)。

自己回帰モデルとは、時刻 t における時系列のデータ X_t として自分自身の過去からの回帰を考えるモデルである。いちばん簡単な1次の自己回帰モデルは、$X_t = A_1X_{t-1} + E_t$ と表される。ここに、X_{t-1} はひとつ前の時系列データ、A_1 は X_{t-1} からの寄与を表すパラメータ(定数)、E_t は時刻 t におけ

る予測誤差（ランダム変動）である。すなわち、現在の時刻 t におけるデータ X_t はひとつ前のデータ X_{t-1} に大いに関係があって、その度合いを A_1 で表す、また、必ずしも過去の自分に全面的に支配されるのではなく、少しは気まぐれに変化するであろうから、それを E_t で表す。同様に、ふたつ前のデータにさかのぼって、現在のデータを考えるのが 2 次の自己回帰モデルで、$X_t = A_1X_{t-1} + A_2X_{t-2} + E_t$ と表される。p 個前までの自分にさかのぼって考えるのが最も一般的な p 次の自己回帰モデルとよばれ、$X_t = A_1X_{t-1} + A_2X_{t-2} + \cdots + A_pX_{t-p} + E_t$ と表される。ここに、A_i は時刻 t – i におけるデータ X_{t-i} からの寄与を表すパラメータ（定数）、また、E_t は時刻 t における予測誤差（ランダム変動）である。現在の観測値が、過去のいくつかの自身の観測値による帰結（回帰）として表されるとの発想である。

自己回帰モデルによるスペクトル推定法は 1969 年に赤池 [8] により考案されたが、これは 1967 年に Burg [9] によって提案された最大エントロピー法と同等であることがすぐに証明された。自己回帰モデルの次数 p は赤池の情報量基準（Akaike's Information Criterion, AIC）を最小とするものが採用されることが多い [2]。 赤池氏自身は An Information Criterion と提唱したが、現在ではどの論文でも Akaike's Information Criterion で引用されている。

AR の具体的な計算には Burg 法と Yule-Walker 法という 2 つのアルゴリズム [2] があるが、データが少ないときは Burg 法が有利とされる。

3．4． 自己回帰モデルによる周期性探索

百人一首第一番「あきのたのかりほのいほのとまをあらみわがころもではつゆにぬれつつ」について、自己回帰モデルによるパワースペクトルを求めてみると、Burg 法では 10.7 字、5.3 字、4.0 字、2.5 字、という周期性が検出されたが、Yule-Walker 法ではピークは検出されなかった。他の多くの和歌でも Yule-Walker 法ではピークが検出されない傾向がみられたため、ここでは Burg 法を採用した。

周期性の探索結果をどうまとめるかは難しい問題であるが、本論文では、重みをつけた頻度分布を求めてみた。例えば一首に 4 つのピークがあるときはそれぞれのピークの頻度を 1／4 = 0.25 のように一首に現れるピーク数で割った値をそのピークの頻度とした。100 首についての結果をまとめると、いろは歌基準でも五十音図基準でも、2.7 文字の周期が最多となり、以下、いろは歌基準では 3.0 字、3.8 字と続き、五十音図基準では 2.5 字、3.0 字と続いた。以下に、五十音図基準で 2.7 文字周期の見いだされた和歌を紹介する。

2.7 文字ピークのみを持つ和歌は、

第三十番「ありあけのつれなくみえしわかれよりあかつきばかりうきものはなし」

第三十七番「しらつゆにかぜのふきしくあきののはつらぬきとめぬたまぞちりける」

第三十八番「わすらるるみをばおもはずちかひてしひとのいのちのをしくもあるかな」

第四十三番「あひみてののちのこころにくらぶればむかしはものをおもはざりけり」

第八十六番「なげけとてつきやはものをおもはするかこちがほなるわがなみだかな」

以上の 5 首であった。

2.7 文字ピークともう一つのピークを持つ和歌は、

第七番「あまのはらふりさけみればかすがなるみかさのやまにいでしつきかも」

第十四番「みちのくのしのぶもぢずりたれゆえにみだれそめにしわれならなくに」

第十八番「すみのえのきしによるなみよるさへやゆめのかよひぢひとめよくらむ」

第二十七番「みかのはらわきてながるるいづみがはいつみきとてかこひしかるらむ」

第四十六番「ゆらのとをわたるふなびとかぢをたえゆくへもしらぬこひのみちかな」

第六十番「おほえやまいくののみちのとほければまだふみもみずあまのはしだて」

第七十七番「せをはやみいはにせかるるたきがはのわれてもすゑにあはむとぞおもふ」

第七十八番「あはぢしまかよふちどりのなくこゑにいくよねざめぬすまのせきもり」

第八十四番「ながらへばまたこのごろやしのばれむうしとみしよぞいまはこひしき」

第九十八番「かぜそよぐならのをがはのゆふぐれはみそぎぞなつのしるしなりける」

の 10 首であった。

2.7 文字ピークともう二つのピークを持つ和歌は、

第十三番「つくばねのみねよりおつるみなのがはこひぞつもりてふちとなりぬる」

第三十三番「ひさかたのひかりのどけきはるのひにしづこころなくはなのちるらむ」

第五十四番「わすれじのゆくすゑまではかたければけふをかぎりのいのちともがな」

の 3 首であった。

2.7 文字ピークともう三つのピークを持つ和歌は、

第三十二番「やまがはにかぜのかけたるしがらみはながれもあへぬもみぢなりけり」

第四十五番「あはれともいふべきひとはおもほえでみのいたづらになりぬべきかな」

第六十四番「あさぼらけうぢのかはぎりたえだえにあらはれわたるせぜのあじろぎ」

第八十九番「たまのをよたえなばたえねながらへばしのぶることのよはりもぞする」

の 4 首であった。

2.7 文字ピークともう四つのピークを持つ和歌は、

第二十三番「つきみればちぢにものこそかなしけれわがみひとつのあきにはあらねど」

第二十八番「やまざとはふゆぞさびしさまさりけるひとめもくさもかれぬとおもへば」

第六十二番「よをこめてとりのそらねははかるともよにあふさかのせきはゆるさじ」

第七十四番「うかりけるひとをはつせのやまおろしはげしかれとはいのらぬものを」

第九十九番「ひともをしひともうらめしあぢきなくよをおもふゆゑにものおもふみは」

の 5 首であった。

2.7 文字ピークともう五つのピークを持つ和歌は、

第三十九番「あさぢふのをののしのはらしのぶれどあまりてなどかひとのこひしき」

第八十番「ながからむこころもしらずくろかみのみだれてけさはものをこそおもへ」

第八十三番「よのなかよみちこそなけれおもひいるやまのおくにもしかぞなくなる」

の 3 首であった。

以上を合計してみると、100 首のうち 30 首が 2.7 文字ピークを示していた。

図 1 に、五十音図基準で、2.7 文字のピーク一つを持つ和歌の例として、百人一首第四十三番「あひみてののちのこころにくらぶればむかしはものをおもはざりけり」の時系列（上段）とパワースペクトル（下段）を示す。図 1 のパワースペクトルでは、周波数 0.37〔1/文字〕のところにピークが顕著で、このピークに該当する周期は 2.7〔文字〕である。この図では、AR の次数は AIC 最小の基準により、 $p = 3$ を採用している。

周期性の全く見いだせなかった和歌は、いろは歌基準で 6 首、五十音図基準で 3 首であった。五十音図基準で周期性のなかった和歌を挙げると、

第三番「あしひきのやまどりのをのしだりをのながながしよをひとりかもねむ」

第五番「おくやまにもみぢふみわけなくしかのこゑきくときぞあきはかなしき」

第六番「かささぎのわたせるはしにおくしものしろきをみればよぞふけにける」

以上の 3 首であった。

3．5． 文字周期の意味

スペクトル解析で得られる文字周期、例えば、重みつきのピーク出現頻度が最多となった 2.7 文字の周期にはどのような意味があるのであろうか。例えば“の”の字が 2.7 文字おきに出現することが考えられるが、前節で示した 2.7 文字周期の和歌の歌詞を眺めてみても、そのような傾向はみられない。

図 1 上段の時系列のグラフを見ながら考えると、和歌の歌詞で、五十音図のはじめの方から引用される文字と終わりの方から引用される文字の出現の有様が 2.7 文字くらいの間に周期的に変わる、といえるのではなかろうか。そのようなことが、いろは歌基準でも同様に 2.7 文字くらいの周期で変わるのが多くみられるというのも興味深いところである。

3．6． 異なるひらがなの文字数

ひらがな読みにしたときに、一首一首で使用しているひらがなの総数の 100 首についての平均値は 31.3 文字であった。和歌はやはり三十一文字のものが多い。

一首一首で使用している異なるひらがなの文字を数えてみると、第一番「あきのたの･･･」については濁点を区別して 25 文字、濁点を区別しないで 24 文字となった。

同様に 100 首についてしらべた結果の平均値は、濁点を区別して 23.2 文字（最大 28 文字、最小 18 文字）、濁点を区別しないで 21.8 文字（最大 27 文字、最小 17 文字）となった。濁点を区別しない場合、ひらがな全 47 文字のうち平均として 46%のものが使われていることになる。案外限られた文字が使われていることがわかる。

3．7． 正規性検定

一首一首について番号付けによって得られた数値のヒストグラムを書き、シャピロ･ウィルクの正規性検定を行ってみた。 p 値（有意確率）が 0.05 より大きい場合は有意水準 1%で、0.01 より大きい場合は有意水準 5%で、正規分布とみなせることになる。

図 2 に、百人一首第四十三番「あひみての･･･」を五十音図基準で番号付けして得られた数値のヒストグラムを示す。図では、確率分布に変換してあり、ヒストグラムの総面積は 1 である。図中の曲線は標本標準偏差と平均値から計算した正規分布曲線である。正規性検定では $p = 0.2731 > 0.05$、すなわち、有意水準 1%で正規分布といえる結果となった。また、最も出現頻度の高い文字は“の”であった。この和歌で使われていた異なるひらがなの文字数は、濁点を区別して 24 文字、濁点を区別しないで 23 文字であった。また、最も出現頻度の高い文字は“の”であった。

100 首についての解析の結果、$p > 0.05$ だったものは、いろは歌基準で 50 首、五十音図基準で 74 首であった。また、$p > 0.01$ だったものは、いろは歌基準で 86 首、五十音図基準で 94 首であった。ここでも、五十音図基準の方が正規分布とみなせる和歌の数が多いという結果となった。

3．8 全 3134 文字についての解析

100 首すべての歌詞をあわせると 3134 文字となる。ヒストグラムを作ってみると、出現頻度の最も高いひらがなは “の” であった。以下、“か”、“は”、“し”、“な”の順となった。ただし、“か”、“は”、“し”は濁点の付いたものも含めて数えている。

ヒストグラムの形を見ると、いろは歌基準では平坦な一様分布的な形で、五十音図基準の方はやや正規分布的な形となっており、番号付けの基準としては五十音図の方が優れている可能性を感じさせた。しかしながら、正規性検定を行ってみると、いろは歌基準では $p = 2\times10^{-30}$、五十音基準では $p = 1\times10^{-26}$で、いずれも全く正規分布とはいえない結果となった。

使用文字数は濁点を区別して 67 文字、濁点は区別しないとすると 47 文字となって、100 首ではすべての文字が使われていた。

4． おわりに

百人一首のひらがなよみに番号付けを行い得られる時系列について、1/f ゆらぎとなるかどうか、特徴的な周期性があるかどうか、正規分布となるかどうか、全 3134 文字についてはどうなるか等をしらべた。

五十音図を基準とした場合は 95 首と、大部分のものが 1/f ゆらぎを示した。また、文字の現れる周期性は 2.7 文字の周期が最多で、100 首中 30 首にみられた。100 首のうちで正規性を示したものは 94 首であった。一首一首に着目すると、必ずしも 1/f ゆらぎである和歌が正規性分布をしているわけではなかった。ただし、1/f ゆらぎでもなく、正規分布でもない和歌は一首もなかった。100 首すべての 3134 文字のうち、出現頻度の最も高いひらがなは “の” であった。

いろは歌はあまり国文法とは関係なしに、たまたまひらがな全部を表す一方法にすぎないであろう。一方、五十音図は母音と、それぞれの母音に関係する子音が整然と並べられているわけで、やはり番号付けの基準としては五十音図の方が優れているものと推察される。

日本語のひらがな読みを番号付け･解析して、というこの研究はいったい何をやっているのですか、という質問をよく受ける。スペクトル解析の方は、日本語の文字の出現順序を解析している。すなわち、ある文字が使われて、その次に使われる文字が47文字のうちどれになるかの確率をしらべているといえるのではなかろうか。今回紹介したヒストグラム解析の方は、単にひらがなを数値に置き換えて、47 種類のひらがなの出現頻度をしらべているわけであるから、統計学で使われている通常のヒストグラムとまったく同じ意味あいである。話の順序としては、より基本的なヒストグラム解析の方を先にすべきであったかもしれないが、今回は前論文からの続きということもあり、筆者の研究途上で思いついたままの順序とさせていただいた。

筆者は理系の研究が本業であるが、このように文系に関係した研究も始めることができ、幸運を感じている。しかしながら、文系の読者からも理系の読者からも不評を買っているのが実態かもしれない。これも多文化共生のひとつの形と御笑覧いただければ幸いである。

最後ながら、このような研究に対し励ましをいただいた故芳賀 馨先生、森 一先生、引地岳雄先生に感謝いたします。また、和歌の入力、論文校正などに協力してくれた、妻の裕子に感謝します。

参考文献

1) 小林恒夫、「和歌等における 1/f ゆらぎ」 比較文化研究, No.70, 13-22, 2005.
2) 北川源四郎、「時系列解析入門」 岩波書店、2005. ブロックウェル･デービス、逸見他訳、「入門時系列解析と予測」 シーエーピー出版、2004.
3) 武者利光、「ゆらぎの世界」 講談社ブルーバックス、1980. 武者利光、「ゆらぎの発想」 日本放送出版協会、1998. 武者利光、「人が快不快を感じる理由」 河出書房新社、1999.
4) 大岡 信、「百人一首」 講談社文庫、1970.
5) 新国語例解辞典付録「全訳小倉百人一首」(小学館).
6) (株)数理システム http://www.msi.co.jp/splus/
7) Shapiro SS and Wilk MB. An analysis of variance test for normality (complete samples). Biometrika 52: 591-611; 1965.
8) Akaike H. Fitting autoregressive models for prediction. Annals Inst Statist Mathem 21:243–247; 1969.
9) Burg JP. Maximum entropy spectral analysis, paper presented at the 37th Annual International Meeting. Oklahoma City, OK: Society of Explor Geophys; 1967.

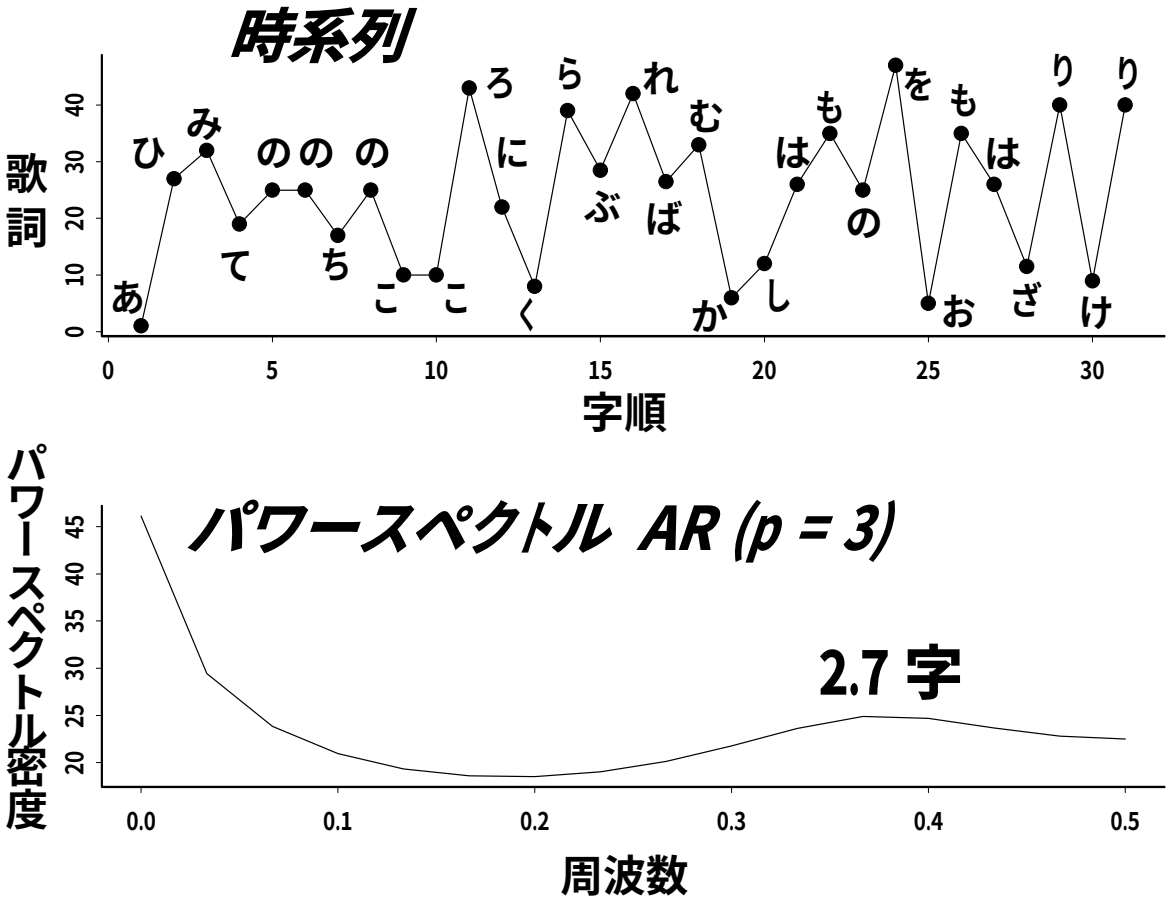

図 1 百人一首第四十三番「あひみてののちのこころにくらぶればむかしはものをおもはざりけり」を五十音図基準で番号付けを行った時系列（上段）と、自己回帰モデル（AR）によって求めたパワースペクトル（下段）。AR の次数は AIC 最小の基準により、 $p = 3$ を選択している。下段でスペクトルの横軸の周波数の単位は〔1/文字〕、縦軸のパワースペクトル密度の単位は〔（字番）^{2}（字）〕である。周波数 0.37 〔1/文字〕 のところにピークが顕著で、このピークに該当する文字周期は 2.7 〔文字〕 である。

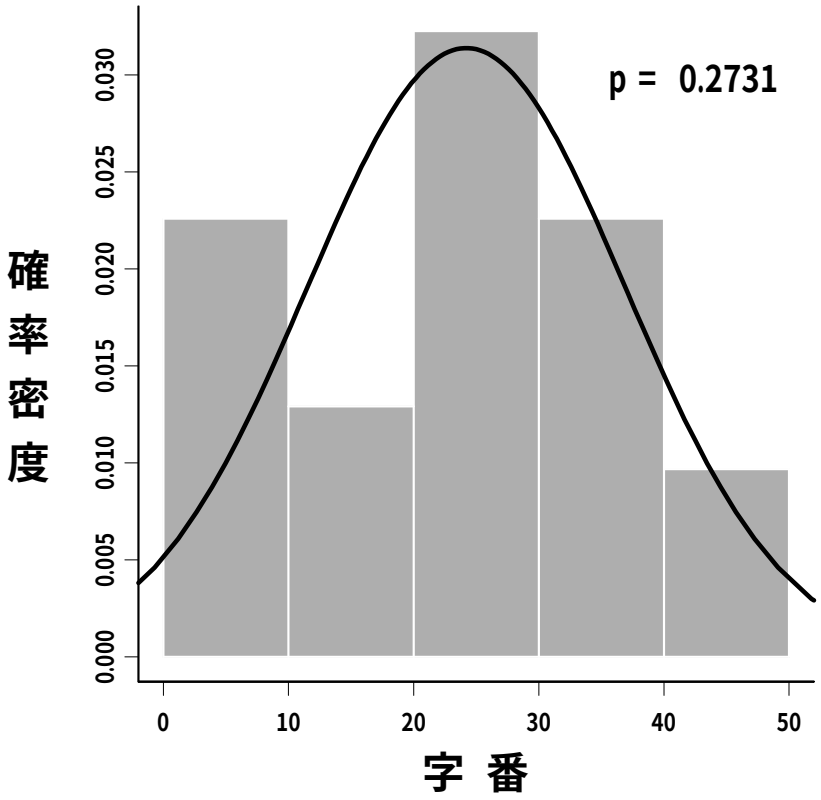

図 2 百人一首第四十三番「あひみてののちのこころにくらぶればむかしはものをおもはざりけり」を五十音図基準で番号付けをして得られた数値のヒストグラム。確率分布に変換してあり、ヒストグラムの総面積は 1 である。曲線は標本標準偏差と平均値から計算した正規分布曲線である。正規性検定では有意確率が $p = 0.2731 > 0.05$ となり、有意水準 1%で正規分布とみなせる結果となった。この和歌で使われていた異なるひらがなの文字数は、濁点を区別して 24 文字、濁点を区別しないで 23 文字であった。また、最も出現頻度の高い文字は“の”であった。

The author has been investigating various Japanese articles by expressing numerically the hiragana reading of them and applying several statistical analyses. In the previous paper, ten randomly selected poetries of the Hyakunin Isshu (single songs of a hundred poets) revealed 1/f (one–over–f) fluctuations. Thus, this paper researches all poetries of the Hyakunin Isshu with spectral and histogram analyses. Firstly, spectral analysis indicated that 95% of the poems showed 1/f fluctuations, and the most frequently appeared periodicity was 2.7 characters. Secondly, from the histogram analysis, 94% of poems passed the test of normality. Thirdly, among the 3,134 characters gathered from the 100 poetries, most frequently appeared character was “no” that corresponds to English character of “of”, e.g. Lastly, the author speculated on the meaning of this kind of numerical analysis. For the basis of making numerical expression, Gojuonzu (Japanese syllabary) might be preferable to Iroha Uta (a poetry that consists of different 47 characters).