

文芸作品に見る日本語音節統計の歴史的特性

佐藤大和
東京外国語大学

本論文は、奈良時代の「万葉集」、平安時代の「源氏物語」、鎌倉・室町時代の「平家物語」、さらに近代の文芸資料である「夏目漱石の講演」等を素材として、1千年以上に渡る日本語の歴史における音節の統計的性質に関して述べたものである。分析に必要なテキストコーパスの作成、音韻・音節の分析法、およびその統計的特性が報告される。音節の分布特性は、4時代ともほぼ同一特性の指数分布に従うことが明らかとなり、これが日本語における普遍的特性であることが示される。また、指数分布の形状を決めるパラメータ（記号論的温度）と音節の平均情報量（エントロピー）との関連が、情報理論的側面から議論される。

Diachronic Characteristics of Syllable Statistics in Japanese Literature

Hirokazu SATO
Tokyo University of Foreign Studies

This paper reports diachronic characteristics of syllable statistics in Japanese language over a thousand years. Four classical or modern literary texts, that is, "Manyō-shū Anthology" in Nara era, "The Tale of Genji" in Heian era, "The Tale of the Heike" in Muromachi era, and "Collected Speeches of Soseki Natsume" as modern literature were analyzed. A unique characteristic based on the exponential distribution in syllable statistics through four periods is described. Moreover, relationship between "semiotic temperature" determining the characteristic and entropy of syllables is discussed from viewpoints of information theory.

1. はじめに

コーパスを利用した言語研究では、特定の対象言語において、語およびその連接、文法、運用、意味など、様々な位相における共時的諸性質を明らかにするのが一般的である。一方、文学や史的文献資料を用いて、対象言語における通時的変化を追跡する研究は、言語の時代時代の特徴と史的变化過程を浮かびあがらせる研究ではあるが、対象言語の時代を越えた普遍的な特質を明らかにしてくれる可能性をも秘めている。

筆者は、文字言語として継承されてきた日本の古典文学のうち、「源氏物語」と「平家物語」を素材として、文字の背後にある音声的側面の諸特性を明らかにする研究を進めてきた[1-3]。言葉は、その始まりにおいては「音声言語」であるが、文字の使用によって書き言葉（文字言語）が発達する。古典文学は、その当時の一般の人々が日常使用する話し言葉とは異なっており、文字によって書き留める形式のものであるとはいえ、「記録」「読み物」としての書き言葉風のものばかりでなく、「語る」、「詠う」など話し言葉風の形式のものが多い。「話し言葉」と「書き言葉」は相互に交渉があって、その結果としてそれぞれの時代の日本語の姿が定まっていたと考えられる。このような事情から、言語の表現形式としてその頂点を極める古典文学の分析からも、日本語の話し言葉の諸性質の通時的変遷を窺い知ることが

できるであろうと思われる。

本論文は、「万葉集」「源氏物語」「平家物語」等の古典文学から、夏目漱石の近代文学に至るテキストを素材として、日本語の1千年以上にわたる歴史のなかで、特に発話の基本的単位である音節の統計的性質について、各時代の分析結果を述べたものである。音節統計上から見た日本語の普遍的性質と、各時代の個別的特徴が述べられる。また、音節の平均情報量（エントロピー）の数理的表現が示され、その性質が情報理論的観点から議論される。著者は、既に「源氏物語」や「平家物語」の音韻・音節分析に関して報告をしているが[1, 2]、本報告は各時代の音韻と音節の設定に関して従来の不備を修正し、内容の詳細化を図ったものである。また近代の文学テキストの分析結果を追加し、それらを比較検討した結果を述べている。

2. 分析用テキスト・コーパス

分析用のコーパスとして、「源氏物語」「平家物語」および近代のテキストとして、夏目漱石の講演資料のテキストを作成した。「万葉集」に関しては、電子化コーパスを作成せず、大野による『万葉集』の音韻表・音節別使用度数の結果をそのまま利用した[4]。「源氏物語」は、村上が作成した中央公論社刊「源氏物語大成」[5]に基づく電子化コーパスを利用し[6]、「平家物語」は、元和九年刊行の流布本に基

づく[7]。また、夏目漱石のテキストは、「私の個人主義」等の文字化された10講演¹を用いた。

上記の分析に用いるテキストは、各種境界記号(区切り記号)付「かな」テキストである。これら境界記号は、文、句、語、形態素の認定のため、また音節の同定のために漢字や造語成分の区切りとして設定された。以下、テキストに使われた境界記号を示す。

- 「。」: 文末、
 - 「,」: 句末境界
 - 「𐀀」: 文節境界(文節の頭につける)、
 - 「-」: 語境界、補助動詞境界
 - 「=」: 接辞境界(「漱石」テキストでは接尾辞境界、接頭辞境界は「≠」)
 - 「/」: 付属語境界(助詞および助動詞)
- この他、会話部分はかぎ括弧「」、和歌は「W」、院宣等の文書は「&」で囲うなど、他の記号もいくつか使用している。
- かな表記は、古典文学に関しては原テキストの表記に従っているが、特に「平家物語」の場合、適切な発音を導出するため、以下の場合などではカタカナ表記を取り入れた。
- ・促音の「つ」→「ツ」
 - ・拗音化処理をしないなどの「や」「よ」の一部
「つひやす(費やす)」→「つひやす」
 - ・ハ行音がワ行音化せず、ハ行音に止まる場合
「もののふ(武士)」→「もののフ」
 - ・その他、表記と異なる発音となるもの
「てんわう(天皇)」→「てん・ワウ」

なお、漱石コーパスは現代文のテキストであるため、促音や拗音ではひらがなの小文字を使用している。こうして作成されたテキスト・コーパスの冒頭の例を以下に示す。

『源氏物語』桐壺

𐀀いづれ/の𐀀おほん=とき/に/か。𐀀によ・う・ごーかう・い𐀀あまた𐀀さぶらひーたまひ/ける𐀀なか/に、𐀀いと𐀀やんごと・な・き𐀀きは/に/は𐀀あら/ぬ/が、𐀀すぐれ/て𐀀とき・めきーたまふ𐀀あり/けり。

『平家物語』祇園精舎

𐀀ぎ・をんーしやう・じや/の𐀀かね/の𐀀こゑ、𐀀しよ・ぎやうーむ・じやう/の𐀀ひびき𐀀あり。𐀀しや・らーさう・じゆ/の𐀀ハな/の𐀀いろ、𐀀じやう・しやーヒツ・すぬ/の𐀀ことわり/を𐀀あらはす。

漱石『私の個人主義』

𐀀わたくし/は𐀀きよう𐀀ははじめ・て𐀀この𐀀が

く・しゅう=いん/とーいうーもの/の𐀀なか/に𐀀はいり/まし/た。𐀀もつとも𐀀い・ぜん/から𐀀がく・しゅう=いん/は𐀀た・ぶん𐀀この𐀀けん・とう/だろ/う/ぐらい/に𐀀かんがえ/てーい/た/に/は𐀀そう・いーあり/ませ・ん/が、𐀀はつきり/とは𐀀ぞん・じ/ませ・ん/でし/た。

3. 分析方法

音節や音韻等の分析は、「仮名遣い音節表」を参照しながら、境界記号付き仮名テキストから、抽出を行うものである。また、文、句、文節、語、形態素の拍数とそのパターンも抽出することができるが、本論では扱わないので内容は省略する。

音節の種類や表記は時代によって異なるため、分析に使用する「仮名遣い音節表」は、各テキストごとに異なったものを用いる。音節表は、(音節)(表記)(子音部)(母音部)(拍数)(拗音、長音の別)等で表記されており、これを用いて音韻の統計分析も可能となる。

なお、本論では、促音、撥音は、便宜上音節の一種として扱うが、長音に関しては、これを含む長音節をひとつの音節として扱うことにする。

4. 各時代の音節

日本語における音韻・音節の種類は、文献によってその姿が明らかになる7世紀から現代に至るまで、各時代を通じて安定して存続しているものもあるが、時代によって種々変貌したり、消失や生成を経てきている。音節の分析に際しては、「仮名遣い音節表」設定のため、各時代の音韻と音節の姿を定めておくが必要になる。以下、各時代における音節をどのように定めたかについて述べる。これらは、馬淵、中田等の著作に基づいてはいるが[8, 9]、分析上一定の線引きが必要になるため、著者の判断で定めた部分もある。

(1) 奈良時代:「万葉集」

奈良時代(7~8世紀)においては、イ列、エ列、オ列で、母音に甲類、乙類の区別があり、従って母音の種類は8個であるという説に従う。子音は、ローマ字で表記すれば、/k, s, t, n, f, m, y, r, w, g, z, d, b/の13種であるが、その音価に関してはここで問題としない。乙類の母音については、/i/, /e/, /o/等で表す。ア行の/i/とワ行の/wi/, およびア行の/e/, ヤ行の/ye/, ワ行の/we/は区別されており、音節の種類数は全部で87音節となる。乙類母音は、次の時代では消滅する。

¹ 10講演とは以下を指す。「教育と文芸」「現代日本の開化」「創作家の態度」「道楽と職業」「中味と形式」「文芸と道徳」「文芸の哲学的基礎」「無題:大正3年東京高等工業学校において」「模倣と独立」「私の個人主義」

(2) 平安・室町時代：「源氏物語」「平家物語」

「源氏物語」では平安時代後半11世紀を、「平家物語」では室町時代末16世紀を想定する。「平家物語」に関しては、南北朝(14世紀末)には確立しているが、その後のキリシタン文献などで音価に至るまで明らかになっており、ここでは音声記号に関してその結果に従う。「源氏物語」に関しては、音声化の試み、すなわち当時の時代の発音とアクセントによって朗読する試みもなされているが[10]、漢語の読みに関しては未だ確実なことが分かっておらず推定の域を出ないため、今回の分析においては、漢語と混種語に関してそれを含む文節を分析対象から除いた(漢語と混種語の選定は、『古典対照語い表』[11]に基づいた)。

「源氏物語」と「平家物語」の分析における、音韻・音節上の両者に共通な内容および違いは以下のとおりである。

(i) ア行とヤ行の「エ」は、「源氏」「平家」ともに統合し、/ye/とする。

(ii) ア行とワ行の「オ」は、「源氏」では/o/と/wo/が区別されるが、「平家」では/wo/に統合。

(iii) 「平家」の時代になり、才段の長母音二種、合音/o:/、と開音/ɔ:/、およびウ段の長母音/u:/が発生した。

(iv) ハ行転呼は、「源氏」「平家」ともにあり、ハ行音は語中でワ行音となる。

(v) 両時代とも、サ行音、ザ行音は/sa, ʃi, su, ʃe, so/, /za, ʃi, zu, ʃe, zo/とする。

タ行音とダ行音は、「源氏」では/ta, ti, tu, te, to/, /da, di, du, de, do/であるが、「平家」では/ta, tʃi, tsu, te, to/, /da, dʃi, dzu, de, do/とする。

(vi) 「平家」では、拗音が確立し、撥音、促音も確立する。「源氏」では、撥音は認めるが、促音は表記文字もないことから認めないこととした。

(vii) 「平家」ではカ行、ガ行の合拗音とその長音/kwa, gwa/, /kwo:, gwo:/が発生した。

(viii) 「平家」では、「エウ(フ)」は長音とする。「源氏」では長音としない(割って読む)。

以上の音韻・音節上の状況から、用いた音節の数は、「源氏」では撥音を含めて67種、「平家」では実際に抽出された157種とした(文献[8]では室町時代の音節として169種が記述されている)。

(3) 近代：「夏目漱石」

漱石の時代の音韻・音節は、現代の発音と同じであるととした。室町時代からの主な違い(変化)をまとめると以下ようになる。

(i) ア行音/a, i, u, ye, wo/→/a, i, u, e, o/

(ii) 5母音すべてに長母音形/a:, i:, u:, e:, o:/がある。開音/ɔ:/は消失。

(iii) 四つ仮名「ジ、ヂ、ズ、ヅ」は、

/zi, dzi, zu, dzu/ → /zi, zi, zu, zu/ とする。

(iv) 外来語からの新たな音節が使われる。

(ヴァ/va/, ヴィ/vi/, ファ/fa/, フィ/fi/ 等)

現代語の場合、外来語で使われる音節も含めて、日本語の音節のすべての可能性を数えあげるのは難しい面がある。ここで扱った「漱石」テキストに現れた音節の種類数は、撥音と促音を含めて全部で185種類であった。

上記4時代において使用されている音韻の種類をまとめて表1に示す。

5. 音節の出現分布とその性質

4時代の文芸コーパスにおける音節の分析結果を図1に示す。図中、横軸は生起数の多い音節の順位(ランク)であり、縦軸は生起数の対数である。生起率0.1%未満の出現頻度の低い音節は、グラフには表示していない。図1から分かるように、音節ランクと生起数の対数はいずれのテキストでも直線関係を示している。その傾きは、「漱石テキスト」で傾きの絶対値がやや大きいものの、テキストによってほとんど同じ傾き(≈-0.07)であると言ってよいであろう。縦軸が対数軸で直線関係になるということは、音節の出現分布は生起ランクに対して指数分布をなすことを意味している。同じ特性を生起確率の対数で示した図2の場合、4つのテキストの音節の出現確率分布特性は、ほぼ重なる。つまり、1千年以上にわたる日本語の歴史のなかで、発音の基本となる音節の出現分布は安定していて、時代やテキストに依らないほとんど同一の分布特性となる。

分布特性の直線の傾きの絶対値は、後に述べるようにその逆数βで表すと便利である。このとき、βの値は14~16程度の値となる。βに関しては、6.3節であらためて触れることにする。

表2に、各テキストの上位20位までの音節の生起数と生起率を示した。奈良時代以降、日本語の音韻・音節は、乙類母音の消滅、音便や漢語の流入に伴うモーラ音素と拗音の発生、連濁現象、合音の発生と消滅、唇音退化や四つ仮名混同など、様々な変化や消長を経てきた。表2を見ると、このような発音の変化・発展を経てきたにもかかわらず、各時代を通じて一貫して頻度の高い音節(/nō/, /no/の系統、/si/, /ʃi/の系統、/ka/等)がある一方、「平家物語」以降では、撥音/N/や母音音節/i/の頻度が著しく高いなど、時代による違いも明らかである。このような音節の生起順位に時代による異同があるにも関わらず、日本語の音節系としてその出現分布が一定の特性を示すということは大変興味深い。

表1 4時代における音韻の種類

	万葉集	源氏物語	平家物語	漱石作品
母音	8 母音 (a, i, u, e, o, ɪ, ɛ, ɔ)	5 母音 (a, i, u, e, o)	5 母音+3 長母音 (a, i, u, e, o) (u:, o:, o:)	5 母音+5 長母音 (a, i, u, e, o) (a:, i:, u:, e:, o:)
単音節母音	5 (a, i, u, e, ɔ)	4 (a, i, u, o)	3 (a, i, u)	10
無声子音	k, t, s, f	k, t, s, f, ɸ	k, t, p, s, f, tʃ, ts, ɸ	k, t, p, s, f, tʃ, ts, h, ɸ
有声子音	g, d, b, z m, n, r	g, d, b, z, ʒ m, n, r	g, d, b, dz, z, dʒ, ʒ m, n, r	g, d, b, z, ʒ m, n, r
半母音	y, w	y, w	y, w	y, w
その他				外来語音節: ti, di, fa, fi, va, vi 等
拗音節子音	なし	なし	ky, j, tʃ, py, ɸy gy, dʒ, by, ʒ, my, ny, ry 合拗音: kw, gw	ky, j, tʃ, hy, py, gy, ʒ, by, my, ny, ry
撥音	なし	あり N	あり N	あり N
促音	なし	なし	あり Q	あり Q

ɪ, ɛ, ɔ は乙類の母音を表す。

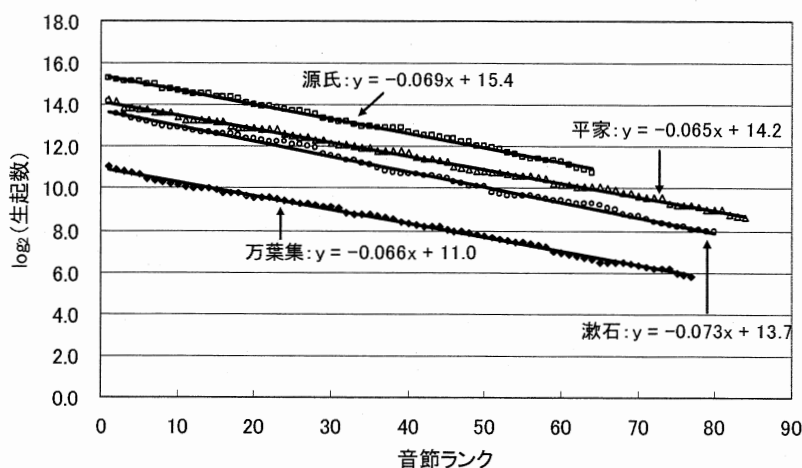


図1 日本語文芸コーパスにおける音節生起ランクと生起数の関係

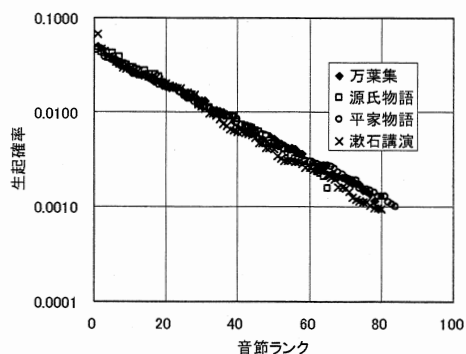


図2 日本語文芸コーパスの音節生起確率分布

6. 音節のエントロピー

6.1. 各時代の音節エントロピー比較

音節の生起確率が知られていれば、音節エントロピー H (平均情報量) が求まる。

$$H = -\sum p_r \log_2 p_r \quad \dots (1)$$

(p_r : r 番目の音節の生起確率)

エントロピーの値が大きい場合、出現する音節の予測はむづかしくなるが、言語系としての表現の多様性は大きく、逆にエントロピーが小さい場合は、音節の予測はたやすくなる分、表現の多様性に乏しい系となる。また、エントロピー H を使って、言語系

表2 高頻度音節とその生起率

順位	万葉集		源氏物語		平家物語		漱石講演	
	音節	生起率%	音節	生起率%	音節	生起率%	音節	生起率%
1	nō	5.0	to	4.5	no	4.9	i	6.8
2	si	4.6	ta	4.3	i	4.7	N	4.7
3	ni	4.4	ka	4.2	N	3.8	no	4.4
4	mo	4.3	fi	4.2	wo	3.8	ka	3.8
5	ka	3.9	ma	4.2	fi	3.6	to	3.8
6	na	3.3	ko	3.8	ni	3.5	fi	3.5
7	tō	3.2	na	3.8	to	3.2	na	3.3
8	fa	3.1	no	3.2	ta	3.2	ta	3.0
9	ma	3.0	ni	3.2	ka	3.1	ni	2.9
10	ki	2.8	ri	3.0	te	2.8	te	2.9
11	ta	2.7	mo	3.0	ri	2.6	o	2.8
12	a	2.5	o	2.8	wa	2.5	ku	2.7
13	tu	2.5	ki	2.8	ma	2.5	ru	2.5
14	ku	2.5	wa	2.8	mo	2.4	de	2.4
15	ra	2.3	wo	2.6	na	2.3	ma	2.4
16	wo	2.2	te	2.5	ko	2.3	ga	2.4
17	ri	2.2	ru	2.5	ra	2.1	mo	2.3
18	mi	2.2	u	2.4	ki	2.0	su	2.2
19	i	1.9	i	2.1	re	2.0	ha	2.0
20	ru	1.9	ra	2.0	sa	1.9	ko	2.0

の冗長度Rを定義することができる。

$$R = 1 - (H/H_{\max}) \quad \dots (2)$$

ここで、 $H_{\max}$ はすべての音節が等確率で出現する場合のエントロピーであり、その系の最大のエントロピー値である。

4 時代におけるテキストから求めたエントロピーH、最大エントロピー $H_{\max}$ 、冗長度R、および文字(仮名)エントロピー等を表3に示す。すべての音節が均等に使用されている場合は、音節数が2倍になればエントロピーは1ビット増加する。しかし、表3が示すように音節種類数はテキストによって様々だが(「漱石講演」の音節数は「源氏物語」の3倍である)、4つのテキスト間でエントロピーに大きな違いは見られない。「源氏物語」→「万葉集」→「平家物語」と音節数が増加するにつれてエントロピーも幾分増加する傾向も見られるが(0.3ビット程度)、「漱石講演」ではエントロピーの顕著な増加は見られない。これらの結果は、エントロピーという量が示す音節記号系の不確かさが、音節数の増加にもかかわらずほぼ一定に保たれていることを示している。音節の種類数の増加は、エントロピーではなく冗長度Rに寄与している。つまり、時代が下るにつれて発音が多様化して音節数が増加しても、音節の出

現を予測することをむづかしくするエントロピーの増大には働かず、冗長性を増すことに利用されていると考えることができる。音節系の冗長性が増せば、不確かな発音や文字の回復に寄与することになる。

表3には、発音上の音節エントロピーのほかに、仮名文字のエントロピーも示した。「万葉集」では音節と仮名が一对一に対応するので音節と仮名文字のエントロピーは等しい。また「源氏物語」でもほとんど同様である。一方、「平家物語」「漱石講演」では、拗音や長音の発生のため、仮名文字と音節エントロピーとの間に0.2ビット程度の乖離が生じている。この乖離は、文字言語の変化はないが、音声言語は多様に変化することを示した値である。

6.2. 音節のbigram/trigramとエントロピー

6.1節で述べた音節のエントロピーは、各音節が独立に生起する場合のものであった(unigram)。実際の言葉では、単語の生起頻度、単語内の音韻配列制約(phonotactics)、付属語の接続制約などによって、音節の生起は条件づけられている。一般に、音節の種類が知られていれば、次に生起する音節のエントロピーは減少する。

先行1音節が分かっているときのエントロピーを H_2 、先行2音節が分かっているときのエントロピーを H_3 とすると、これらは音節の二つ組(bigram)と三つ組(trigram)の生起確率を用いて、以下のように表現できる[12]。

$$\begin{aligned} H_2 &= -\sum p(i, j) \log_2 p(i, j) \\ &= -\sum p(i, j) \log_2 p(i, j) + \sum p(i) \log_2 p(i) \end{aligned} \quad \dots (3)$$

同様に

$$\begin{aligned} H_3 &= -\sum p(i, j, k) \log_2 p(i, j, k) \\ &\quad + \sum p(i, j) \log_2 p(i, j) \end{aligned} \quad \dots (4)$$

H_2 、 H_3 、をunigramのエントロピー H_1 とともに、表4に示す。表4では、音節エントロピーのほか、かな文字のエントロピーも示した。なお、万葉集に関しては、利用したデータはunigramしかないため、ここでは取り上げていない。また表4では、文節内および単語(自立語)内におけるそれぞれのエントロピーが求められている。なお、「源氏物語」では、かな文字と発音(音節)はほぼそのまま対応するため、かな文字のエントロピーのみを示したが、音節エントロピーと読み替えても差し支えない。音節内のunigramエントロピーは、表3の結果の再掲となるが、単語内の結果は付属語が分析から除かれるため、表3とは異なる結果となる。

かな文字および音節とも、unigramのエントロピー H_1 は文節内と単語内でその値に大きな差はない。しかし、 H_2 、 H_3 になると両者の値の差は大きくな

表3 文芸コーパスにおける音節の統計的諸性質

	万葉集	源氏物語	平家物語	漱石講演
音節種類数	87	68	157	185
分析総音節数 (×1000)	42	854	398	266
音節エントロピー (bit) H	5.74	5.58	5.89	5.74
最大エントロピー (bit) $H_{\max}$	6.44	6.09	7.29	7.54
冗長度 R	0.11	0.08	0.19	0.24
仮名文字エントロピー (bit)	5.74	5.59	5.70	5.56
指数分布パラメータ β	15.2	14.4	15.5	13.7
$\log_2 \beta$ (bit)	3.9	3.8	4.0	3.8

表4 音節の bigram, trigram を考慮したエントロピー

(1)「源氏物語」 (bit)

種類	かな文字	
	文節内	単語内
H_1	5.59	5.54
H_2	4.14	3.68
H_3	2.44	1.56

(2)「平家物語」 (bit)

種類	かな文字		音節	
	文節内	単語内	文節内	単語内
H_1	5.70	5.66	5.89	5.95
H_2	4.58	4.31	4.65	4.31
H_3	2.70	1.84	2.66	1.72

(3)「漱石講演」 (bit)

種類	かな文字		音節	
	文節内	単語内	文節内	単語内
H_1	5.56	5.50	5.74	5.80
H_2	4.16	3.92	4.30	3.93
H_3	2.54	1.81	2.41	1.54

っていく。 H_3 の場合、その差は1ビットに近い。これは、文節内の場合、語の複合や付属語連続の連接によって、音節系列の多様性が増し、これが後続音節の予測をむつかしくしているためエントロピーが大きくなると考えられる。単語内の H_3 の場合のエントロピーは、1ビット台と小さな値となっている。

すでに触れたように、「平家物語」と「漱石講演」

における Unigram のエントロピー H_1 は、音節の多様性を反映して、音節エントロピーの方がかな文字エントロピーよりもやや大きい。しかし、 H_2 と H_3 になると、むしろ音節エントロピーの方の値が小さくなる傾向が認められる。これは、文字の並びよりも音節の並びの分布に偏りがあること、あるいは配列の制約の影響であることが考えられる。

6.3. 音節エントロピーの定式化

前節では、音節の生起分布特性が指数分布となることを述べた。生起ランク r の音節の生起数 n_r を、2 を底とする指数表現で表すと以下のように書ける。

$$n_r = A \cdot 2^{-r/\beta} \quad \dots (5)$$

このような指数分布特性は、統計力学や1因子情報路モデル等の理論でよく知られているように、エントロピー $H = -\sum p_r \log_2 p_r$ を最大に、またエネルギーなどの平均特性値（ここでは平均ランク値） $A R = \sum r \cdot p_r$ を最小にするという条件で解かれた分布の形となっている。

この式を用いると、音節エントロピー H は以下のような数式で表現できる。

$$H = \log_2 Z + (1/\beta) (\sum r 2^{-r/\beta} / Z) \quad \dots (6)$$

ここで、 $Z = \sum 2^{-r/\beta}$: 分配関数

β は、音節の生起分布の形をきめるパラメータであり、(ランク出現数の対数) 平面上での直線特性の傾きを表す。 β が大きいとき、直線の傾き(絶対値)は小さくなるので、分布は平坦になって各音節の生起頻度の差は小さくなるが、 β が小さいときは直線の傾きが急峻になり、音節分布の偏りが大きくなる。またよく知られているように、 β は、統計物理学における理論とのアナロジーから、“温度”に相当するパラメータになっている。すなわち“温度”が高いときは、ランクの大きい音節への生起頻度が増し、

各音節の生起率の差が小さくなって一様分布に近づく。逆に“温度”が低くなると、小さいランクの音節に分布が低ランク領域に片寄ることがイメージされる。ここでは分布パラメータ β を“記号論的温度 (semiotic temperature)”と呼ぶことにする。式(6)から分かるように、エントロピー H は分配関数 Z の対数と、ランクの平均値を β で除したものの和で表される。

(6) 式に、数列の和の公式、

$$\sum x^r = x(1-x^M)/(1-x)$$

$$\sum r x^r = x(1-x^M)/(1-x)^2 - M x^{(M+1)}/(1-x)$$

$$(x=2^{(-1/\beta)}, M: \text{音節の種類数})$$

を適用して、音節数 M が十分大きいときの音節エントロピー H_s を求めると、以下ようになる。

$$H_s = -\log_2 (1 - 2^{(-1/\beta)}) + 2^{(-1/\beta)} / \beta (1 - 2^{(-1/\beta)}) \dots (7)$$

これは、 β の値だけで決まる一定値であり、ランク r が増加するとエントロピーは (7) 式で決まる値に飽和することが分かる。 $\beta = 1$ のときは、 $H_s = 2$ (ビット) となるが、 $\beta > 1$ の大きな値のときは、 $2^{(1/\beta)} \approx 1 + \log 2 / \beta$ で近似すると、(7) 式は以下で表現される。

$$H_s = \log_2 \beta + 1.9715 \approx \log_2 \beta + 2 \dots (8)$$

すなわち、音節の飽和エントロピーは、 β の対数に約 2 ビットを加えた値となる。また“対数温度” $\log_2 \beta$ が 0 ビット度のとき、エントロピーは 2 ビットであるが、 $\log_2 \beta$ が 1 ビット増加する毎にエントロピーが 1 ビットづつ増加する特性となっている。

“対数温度” $\log_2(\beta) = 0$ ビット度は、ビット記号系としての音節を考えたときのある種の基底的状態と考えることができる。

表 3 には、各テキストにおけるエントロピーの他、 β および $\log_2 \beta$ の値も示したが、エントロピーは“対数温度” $\log_2 \beta$ に 2 を加えた (8) 式の値となっていることが実際のデータからも分かるであろう。

7. 考察

7 世紀から現代までの日本語における音節の生起率は、時代によって頻度順位の種類に差はあるものの、いずれの時代も出現ランクに対して指数分布をなすことを述べてきた。「平家物語」の音節数は、「万葉集」や「源氏物語」に比べて 2~3 倍数が多いが、エントロピーは 0.2~0.3 ビット大きいにすぎない。これは音節の分布が指数分布となることと関係している。本節では、音節エントロピーがほぼ一定値に止まる理由を考えてみたい。

図 3 は、ここで分析したテキストの平均的 β の値、すなわち $\beta = 16$ のときの (7) 式が示す音節ランク

とエントロピーとの関係を示したものである。細線のカーブは、すべての音節が等確率の場合の最大エントロピー曲線である。図中、各文学資料の音節数の位置を示してある。図から分かるように、エントロピー曲線は、音節ランク、すなわち音節数が 100 以上のところで飽和していることが分かる。パラメータ β によって音節分布が定まっていれば、音節数が多くても音節系のエントロピーは増加せず、一定値に止まる ($\beta = 16$ ビットの場合は、(8) 式から $H = 6$ ビットとなる)。このことは、生起頻度の少ない音節は、エントロピーにほとんど寄与しないということをも意味している。一方この低頻度の音節は、(2) 式の冗長度の増大に寄与していることになり、従って音節の指数分布特性は、音節数が増加しても、音節系の曖昧度を一定に抑えつつ、冗長度を増加して言語の運用性を高める役割を果たしているとも考えることができる。

ここでもう少し音節エントロピーの特性について考えてみることにしよう。図 4 は、種々の β (こ

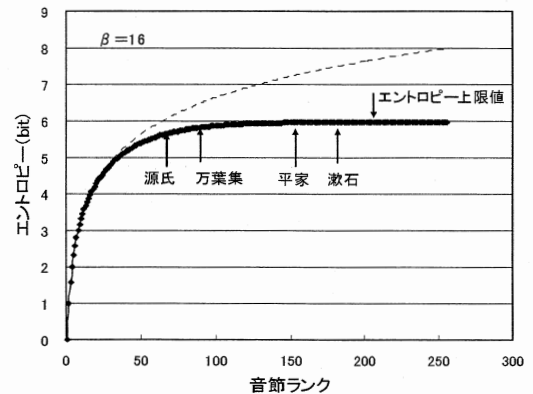


図3 音節ランクとエントロピーの関係
(理論値、 $\beta = 16$ の場合)

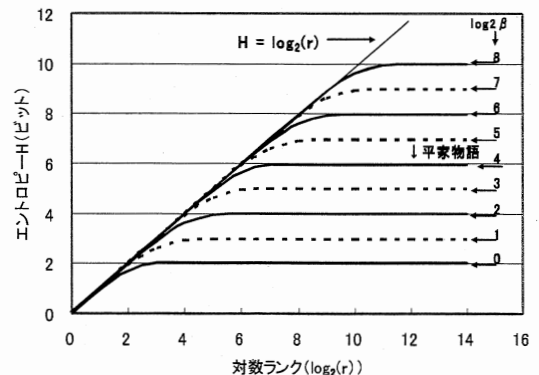


図4 音節の対数ランクとエントロピーの関係

では対数 β)のときの音節ランクの対数とエントロピーの関係を示したものである。 β が小さいときは、ランク数の増加とともにエントロピーはいち早く飽和するが、 β が大きくなると飽和エントロピーは大きくなる。対数ランクの小さいところでは、エントロピーは $H = \log_2(\text{rank})$ の直線関係で増加する(音節数の対数が1ビット増えれば、エントロピーは1ビット増加する)。すなわち、エントロピーは等確率生起の場合と同じ増加特性を示すが、(対数ランク+3)ビット程度の音節数のところで飽和値となる。図中、矢印は「平家物語」の場合の β の位置を示す。

以上示したように、音節のエントロピーは、音節数が少ない場合には音節数の対数に比例して増加するが、音節数が増えると指数分布のパラメータ β に依存して決まり、(β の対数+2)ビットの値をとる。通常音節の種類数は、100程度あるいはそれ以上であるから、音節のエントロピーは、音節の数よりはむしろ分布パラメータ β によって決まっていると考えられる。日本語の音節は、1千年以上にわたってこの「記号論的温度」 β がほぼ一定であるという性質に支配された分布特性になっている。

8. 結論と今後の研究

日本語の約1300年程度に渡る歴史の中で、「万葉集」「源氏物語」「平家物語」「夏目漱石講演」という4つの時代における文芸作品を取り上げ、その時代の言葉の諸特性のうち、特に音節の統計的性質に焦点を当てて分析した結果を述べた。日本語の音節は、この歴史的期間において様々な消長があったものの、それぞれの時代の音節記号系は一定の分布特性をもち、その特性は時代によって大きく変動していないことが明らかとなった。音節分布は指数分布形であり、その分布形を定めるパラメータ β は4時代でほぼ一定の値をもつ。また、音節のエントロピーはこの β によって決まるので、時代によって音節の種類が増加しても音節系の不確かさを表すエントロピーは大きく増加することはない安定している。このことは、人間が言葉の運用をしやすくし、かつコミュニケーションを円滑に進める上で有効に働いているのではないかと考えられるのである。

音節のbigramとtrigramの分析も試みたが、このような音節n-gramの分析は単語の生起頻度の分析に通じていくであろう。音節の統計的性質は、単語内のphonotacticな性質ばかりでなく、単語の生起頻度にも深く関連している。「源氏物語」と「平家物語」の語の生起分布は、世界の他の言語でも広く観測されているように、Zipfの法則に従うことが確認されている[3]。こうした語の頻度分布と音節の分布との関連も今後の課題である。

本研究は、21世紀COE「大規模知識資源の体系化と活用基盤構築」(東京工業大学)で作成されたコーパスに基づいて、グローバルCOE「コーパスに基づく言語学教育研究」(東京外国語大学)においてなされた。「源氏物語」の電子テキストを提供くださった同志社大学：村上征勝教授、また種々ご教示いただいた早稲田大学：上野和昭教授に深謝する。

参考文献

- [1] 佐藤大和, 貝沼諭, 山室恭子, 「古典文学テキストの音声学的分析—「平家物語」を素材として—」, 「人文科学とコンピュータシンポジウム」, 135-142 (2004)
- [2] Sato, H., K. Yamamuro and S. Kainuma, "Search for Spoken Japanese Characteristics in Classical Literature", *Proceedings of Symposium on Large-scale Knowledge Resources (LKR2005)*, 13-20 (2005)
- [3] Sato, H., S. Chida, and K. Yamamuro, "Statistical Properties of Classical Japanese Literature as Spoken Language", *Proceedings of Symposium on Large-scale Knowledge Resources (LKR2006)*, 95-98 (2006)
- [4] 大野晋, 「音韻の変遷(1)」『岩波講座 日本語 5 音韻』147-219, 岩波書店 (1977)
- [5] 池田亀鑑, 『源氏物語大成』, 中央公論社 (1984)
- [6] 金明哲, 村上征勝 (2003) 「文章の統計分析とは」, 『言語と心理の統計』3-57, 岩波書店
- [7] 高橋貞一, 『平家物語 (上下)』, 講談社文庫 (1972)
- [8] 馬淵和男, 『国語音韻論』第8版, 笠間書院 (1994)
- [9] 中田祝夫, 『講座国語史 第2巻 音韻史・文字史』, 大修館書店 (1972)
- [10] 金田一春彦, 「紫式部は「源氏物語」をどう発音したか」『日本語セミナー五 日本語のあゆみ』43-74, 筑摩書房 (1983)
- [11] 宮島達夫, 『古典対照語彙表』, 笠間書院 (1971)
- [12] 中川聖一, 『情報理論の基礎と応用』, 近代科学社 (1992)