Statistical Law Related to Shape of Kanji

漢字の形における統計則

太田 守洋 (琉球大学大学院 理工学研究科) *
山本 健 (琉球大学 理学部)

Statistical Law Related to Shape of Kanji

Morihiro Ohta (Graduate Sch. of Eng. and Sci., Univ. of Ryukyus)
Ken Yamamoto (Fac. of Sci., Univ. of Ryukyus)

要旨

本研究では漢字の形を統計的に分析する。漢字のサイズを特徴づける基本的な指標として画数があり、画数が多い漢字ほど複雑な形である傾向がある。一方、漢字を構成する線の長さが形の複雑性を表すとみなすことができる。本研究では、漢字の線長をコンピュータのフォントを用いて計測し、線長と画数の関係を調べた。その結果、漢字の線長は画数に対しておおむねベキ乗則にしたがって増加することがわかった。さらに、フラクタル図形を基にした数学的なモデルを導入し、ベキ乗則の指数とフラクタル次元の関係を理論的に導出した。この関係を漢字のベキ指数に適用すると、フラクタル次元はおよそ 2 次元となった。すなわち、漢字の形は画数の増加とともに平面充填的に複雑化するといえる。

1. 言語とベキ乗則

言語についての様々な統計則が**ベキ乗則**(1)で表される。その中でも有名なものとして **Zipf の法則**がある。ある文章において出現する単語を出現頻度によって順位付けしたとき、順位 r の単語の出現頻度 f が r に反比例するという経験則が Zipf の法則である。つまり出現頻度と順位の間に $f \propto r^{-1}$ というベキ乗則が成り立つ (Zipf 2013)。他にも、ある文章において文書量が増えるとともに、語彙量がベキ乗則に従って増えることを表した **Heaps の法則**や、文書を一定の単語数ごとに区切ったとき、ある単語が各々の区間に現れる数の平均と標準偏差にベキ乗則が現れるという **Taylor の法則**などがある (田中久美子 2018)。

一般に、ベキ乗則とは 2 つの変数 x および y の間に、

$$y \propto x^{\alpha} \quad (\alpha \text{ は定数})$$

という関係が成り立つということである。すなわち、y が x の α 乗に比例するということである。ベキ乗則に従うデータを両対数グラフにプロットすると、

$$\ln y = \alpha \ln x + C \quad (C \text{ は定数})$$

となるので、データは直線上に並び、その傾きは α である。また、x を測る物差しの目盛を m

* k188323@eve.u-ryukyu.ac.jp
(1) 助動詞の「べき」と区別しやすくするため、ベキ乗則の「ベキ」はカタカナで表記する。

倍して mx としても

$$y \propto (mx)^{\alpha} \propto x^{\alpha}$$

となって、比例定数を除いて関数形は変わらない。つまり、ベキ乗則に従う現象は拡大や縮小に対して不変な性質、**スケール不変性**を持っている。

2. 漢字の画数と線長の関係

漢字には形・音・義という三要素があると言われているが、本研究では漢字の “**形**” の複雑性に注目する。漢字にはサイズを特徴づける基本的な指標として画数があり、画数が多い漢字ほど複雑な形である傾向がある。一方で漢字は全て線で構成されているため、線の長さの和（以下では単に線長とよぶ）はその漢字の形の複雑性を表すとみることができる。そこで本研究では、漢字の線長と画数の関係について分析する。なお、本研究の詳細は Ohta and Yamamoto (2019) で述べられている。

2.1 常用漢字と JIS 漢字

本研究では、**常用漢字** 2136 字と JIS 第 1 水準および第 2 水準漢字 (以下、**JIS 漢字**とよぶ) 6355 字を分析に用いた。各漢字の画数はオンラインのデータベース『Joyo_Kanji』(KeitarouNakayama 2015) および『漢字辞典オンライン』を利用した。

常用漢字とは、「法令、公用文書、新聞、雑誌、放送など、一般の社会生活において、現代の国語を書き表す場合の漢字使用の目安」(文化庁 2010) として内閣告示によって定められた漢字である。現行の常用漢字は 2010 年に改定され、2136 字からなる。

JIS 漢字コードはコンピュータ等のデジタル機器でデータを通信するために、日本工業規格 (JIS; Japanese Industrial Standards) で定められた文字コードである。現在では、JIS 第 1 水準から第 4 水準まで定められている。本研究では、JIS 第 1 水準および第 2 水準の 6355 字[(2)] を JIS 漢字と呼ぶことにする。

なお、常用漢字のうち 2102 字は JIS 第 1 水準であるが、“鬱” や “丼” など 30 字は JIS 第 2 水準であり、“叱”, “塡”, “剝”, “頬” の 4 字は JIS 第 3 水準である[(3)](安岡孝一・安岡素子 2017)。そのため、本研究の JIS 漢字は常用漢字を完全に包含しているわけではない。

2.2 線長の測定法

本研究では、コンピュータフォント MS ゴシックを用いて線長を測定する。コンピュータでは線長を直接測定することが難しい。そこで、以下の方法で線長を計算した。まず漢字を縦横 100 ピクセル四方の 2 値画像として出力し、黒いピクセルの数を数えることでその漢字の面積を測定する。図 1 (a) は 20×20 で “太” を出力した例である。次に、各行各列で黒いピクセルが連続していくつ並んでいるのかを数えていく。例えば、図 1 (a) の下から 3 段目では “4, 1, 3” となる。線の長さにあたる黒ピクセルの長さは、筆画ごとに値が変わる。一方で、MS ゴシックは一つの文字の中で線幅がほぼ一定であるため、その線幅にあたる黒ピクセルの長さは

(2) この 6355 字は JIS X 0208 という規格に当たる。のちに JIS 第 3 水準および第 4 水準を追加した JIS X 0213 という規格が定められた。

(3) JIS 第 1 水準に代わりとして “叱”, “塡”, “剝”, “頬” の 4 字が示されている。

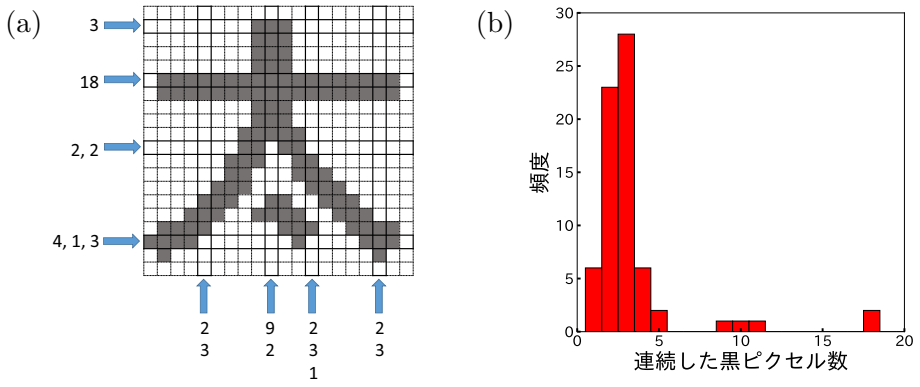


図 1　漢字の線幅の測定法の例。(a) 20 × 20 で出力した “太” と連続して並んだ黒いピクセル数の例。(b) (a) の連続した黒いピクセルの長さのヒストグラム。

ほとんど一定になる。よって、線幅にあたる長さは線長にあたる長さより圧倒的に多く出現するはずである。したがって、図 1 (b) のように連続した黒いピクセルの長さのヒストグラムを描くと、最頻値がその漢字の線幅として推定される。例えば、図 1 の “太” では、線幅は 3 と推定することができる。漢字の線長は、面積を線幅で割れば求めることができる。なお、線幅は字によって異なるため、線幅の推定と線長の計算はそれぞれの漢字について行う必要がある。

2.3　結果

図 2 は (a) 常用漢字 2136 字および (b) JIS 漢字 6355 字について、線長と画数の関係を両対数目盛でプロットしたグラフである。線長は 2 値画像の 1 辺の長さが 1 になるように規格化している。各画数の平均線長 l は画数 s が増えるに従って、おおむねベキ乗則で増加していることがわかる:

$$l \propto s^{\beta}. \tag{1}$$

そのベキ指数 β は、常用漢字で $\beta = 0.47$, JIS 漢字で $\beta = 0.52$ と 0.5 に近い値をとる。

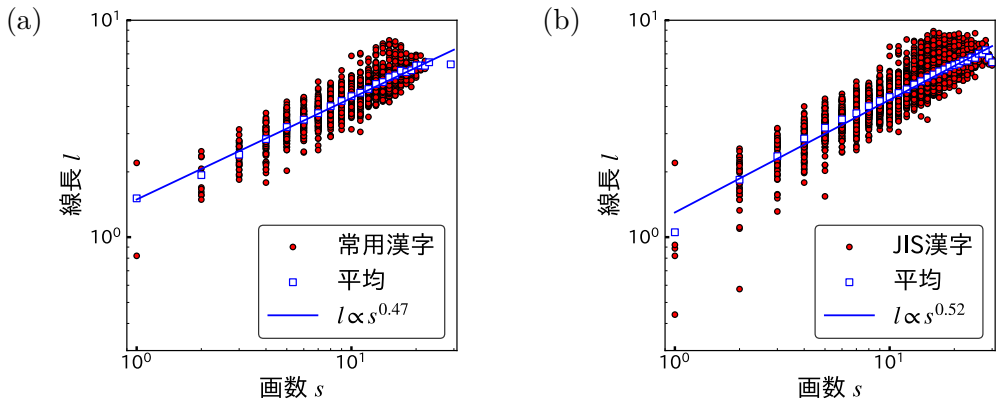


図 2　(a) 常用漢字 2136 字および (b)JIS 漢字 6355 字についての線長と画数の関係。丸い点は各漢字のデータを示し、四角い点は各画数での平均線長を示している。実線は平均線長のベキ乗則を表したものであり、そのベキ指数は (a) 常用漢字で $\beta = 0.47$, (b)JIS 漢字で $\beta = 0.52$ である。

3. フラクタルを基にしたモデル化

3.1 フラクタルとは

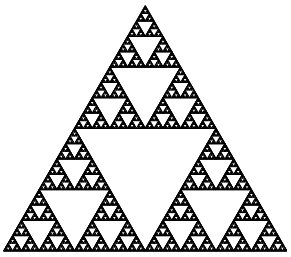


図 3 Sierpinski gasket. フラクタル次元は約 1.59 である。

フラクタルとは図形の一部を拡大すると図形全体と一致するという性質（**自己相似性**）を持つ図形のことである (詳しくは本田勝也 (2002) 等を参照)。例えば、図 3 の Sierpinski gasket では、上半分の正三角形を 2 倍に拡大すると全体と一致する。フラクタルはフラクタル次元という非整数の次元を持つ。フラクタル次元の定義は何通りかあるが、ここでは相似次元という次元を紹介する。正方形の 1 辺の長さを 1/2 倍にすると、その 1/2 倍の正方形を 4 $(= 2^2)$ 個並べるともとの正方形と一致する。また、立方体の 1 辺の長さを 1/2 倍にすると、その 1/2 倍の立方体を 8 $(= 2^3)$ 個並べるともとの立方体と一致する。正方形は 2 次元、立方体は 3 次元の図形であり、これらの次元が指数に現れている。この結果を拡張し、ある図形の 1 辺の長さを r 倍にしたとき、その縮小した図形を N 個並べるともとの大きさに戻るのであれば、

$$N = \left(\frac{1}{r}\right)^D$$

によって相似次元 D を定義する。すなわち、図形の縮小倍率 r と個数 N の間にベキ乗則が成り立ち、そのベキ指数が次元 D に対応する。これを D について解くと、

$$D = \frac{\ln N}{\ln (1/r)}$$

となる。一般に、この次元 D は整数とは限らない。例えば、図 3 の Sierpinski gasket は 1/2 倍にした図形を 3 つ並べるともとの大きさに戻るので、そのフラクタル次元は $D = \ln 3 / \ln 2 \approx 1.59$ である。

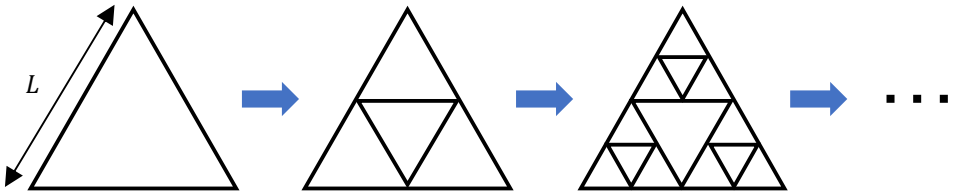
3.2 モデル化

式 (1) のベキ乗則を説明するために、フラクタルを基にしたモデルを導入する。このモデルは 1 ステップごとに三角形を追加していく。まず $n = 0$ のとき, 図 4 左のように 1 辺の長さが L の正三角形を描く。このとき、各線分を 1 画で描くことにすると画数は $s_0 = 3$, 線長は $l_0 = 3L$ である。次に $n = 1$ で、図 4 中央のように正三角形の中に 1 辺の長さが $L/2$ の正三角形を追加する。画数と線長はそれぞれ $s_1 = s_0 + 3^1 = 6, l_1 = l_0 + (3/2)^1 L = 9L/2$ になる。$n = 2$ では、さらに 1 辺の長さが $L/4$ の正三角形が 3 つ加えられ、画数は $s_2 = s_1 + 3^2 = 15$, 線長は $l_2 = l_1 + (3/2)^2 L = 27L/4$ となる。この操作を n 回繰り返したとき、画数 s_n と線長 l_n はそれぞれ

$$s_n = s_{n-1} + 3^n, \quad l_n = l_{n-1} + \left(\frac{3}{2}\right)^n L$$

という漸化式で表すことができ、その解は

$$s_n = 3 + \sum_{i=1}^{n} 3^i = \frac{3}{2}\left(3^n + 1\right) \tag{2}$$



$n=0$ $s_0=3$ $l_0=3L$

$n=1$ $s_1=3+3^1$ $l_1=3L+\left(\frac{3}{2}\right)^1 L$

$n=2$ $s_2=3+3^1+3^2$ $l_2=3L+\left(\frac{3}{2}\right)^1 L+\left(\frac{3}{2}\right)^2 L$

図 4　フラクタルを基にしたモデル。1 辺の長さが L の三角形から始め、各ステップごとに 1 辺の長さを半分にした三角形を加えていく。n 回目のステップの画数 s_n と線長 l_n をそれぞれの三角形の下に示す。

および

$$l_n = 3L + \sum_{i=1}^{n} \left(\frac{3}{2}\right)^i L = 3L\left(\frac{3}{2}\right)^n \tag{3}$$

である。式 (2) および (3) から n を消去すると

$$l_n = 3L\left(\frac{2}{3}s_n - 1\right)^{1-\ln 2/\ln 3}$$

となり、十分に n が大きいとすると

$$l \propto s^{1-\ln 2/\ln 3} \tag{4}$$

というベキ乗則が得られる。このモデルは $n \to \infty$ で Sierpinski gasket (図 3) に収束する。Sierpinski gasket のフラクタル次元が $D = \ln 3/\ln 2$ であることから、式 (4) は

$$l \propto s^{1-1/D} \tag{5}$$

と表すことができる。式 (1) および (5) の指数を比べると

$$\beta = 1 - \frac{1}{D},$$

つまり、

$$D = \frac{1}{1-\beta} \tag{6}$$

である。漢字の線長と画数の関係から得られた指数 $\beta_{\text{常用}} = 0.47$, $\beta_{\text{JIS}} = 0.52$ を式 (6) に代入すると、常用漢字の場合には $D = 1.89$, JIS 漢字の場合には $D = 2.08$ [(4)] となる。次元 D が 2 に近いことから、漢字の形は画数が増えるにしたがって平面充填的に複雑化していくことが示唆される。この結果は、画数が多い漢字を小さく印刷すると真っ黒に塗りつぶされたように見えるという日常的な経験と対応していると思われる。

(4) 漢字は平面に描かれた図形なので、次元 D が 2 を超えることはない。JIS 漢字の場合に $D > 2$ となったのは、単に統計的な誤差が原因と考えられる。

4. まとめ

本研究によって、漢字の線長と画数におおむねベキ乗則の関係があることがわかった。このベキ乗則を説明するために、フラクタルを基にしたモデルを考案し解析した。実データとモデルの解析結果から、漢字の形は画数が増えるにしたがって平面充填的に複雑化していくという結果が得られた。

今回の結果は MS ゴシックを用いたものであったが、他のフォント・書体でもほぼ同様の結果が得られる。さらに、中国における常用漢字 “通用規範漢字表” や台湾における常用漢字 “常用国字標準字体表” および “次常用国字標準字体表” を用いても $\beta \approx 0.5$ のベキ乗則を得ることができる。今後は他のベキ乗則との関係（スケーリング関係）について分析し、漢字の形におけるベキ乗則についてさらに理解を深めたい。

謝　辞

本研究は、科研費 基盤研究 (C) (18K06406) の助成を受けたものである。

文　献

George K. Zipf (2013). *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. London: Routledge.

田中久美子 (2018).「言語の数理的普遍 (1) 言語の経験則」　数学セミナー, 57:8, pp. 68–73.

Morihiro Ohta, and Ken Yamamoto (2019). “Power-law Relation and Complexity in the Shape of Chinese Character (Kanji).” *Journal of the Physical Society of Japan*, 88:6, p. 064803.

KeitarouNakayama (2015). *Joyo_Kanji*. `http://linkdata.org/work/rdf1s3597i`.

文化庁 (2010).『常用漢字表』, `http://www.bunka.go.jp/kokugo_nihongo/sisaku/joho/joho/kijun/naikaku/kanji/`.

安岡孝一・安岡素子 (2017).「日本・中国・台湾・香港・韓国の常用漢字と漢字コード」　京都大学学知創生ユニット報告書, pp. 1–146.

本田勝也 (2002).『フラクタル』　朝倉書店, 東京.

関連 URL

『漢字辞典オンライン』　`https://kanji.jitenon.jp/`