第 3 回: ジップとジップの法則

George Kingsley Zipf
1902–1950 (CC 4.0)

有名なジップの法則について考える。関連して語彙の量的構造、分布の問題についても考える。

ジョージ・キングズリー・ジップ（George Kingsley Zipf、1902.01.07–1950.09.25）は、アメリカの言語学者、哲学者でさまざまな言語の統計的特徴に関する研究をおこなった。ジップはハーバード大学のドイツ語部門の代表を務め、University Lecturer となった。ジップは中国語や人口統計学などの研究も行った。ジップの法則は、ウェブのアクセス頻度、所得等の分布の一般的特徴を説明できると言われている。

1 ジップ曲線

『トムソーヤの探偵』 マークトゥエイン 1896 年
CHAPTER I. TOM SEEKS NEW ADVENTURES

DO you reckon Tom Sawyer was satisfied after all them adventures? I mean the adventures we had down the river, and the time we set the darky Jim free and Tom got shot in the leg. No, he wasn't. It only just p'isoned him for more. That was all the effect it had. You see, when we three came back up the river in glory, as you may say, from that long travel, and the village received us with a torchlight procession and speeches, and everybody hurrah'd and shouted, it made us heroes, and that was what Tom Sawyer had always been hankering to be.

For a while he WAS satisfied. Everybody made much of him, and he tilted up his nose and stepped around the town as though he owned it. Some called him Tom Sawyer the Traveler, and that just swelled him up fit to bust. You see he laid over me and Jim considerable, because we only went down the river on a raft and came back by the steamboat, but Tom went by the steamboat both ways. The boys envied me and Jim a good deal, but land! they just knuckled to the dirt before TOM.

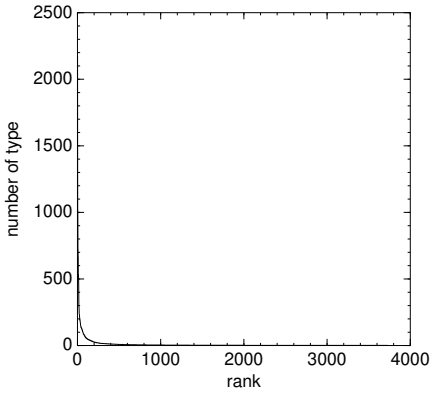

図 1: マークトゥエイン「トムソーヤの探偵」中の単語頻度・順位

表 1: 「トムソーヤの探偵」頻度上 10 位（左）、下 10 位（左）の一覧。

Top 10	freq.	word	bottom 10	freq.	word
1	2172	and	3721	1	yahoos
2	1603	the	3722	1	yanked
3	1027	a	3723	1	yanking
4	1006	it	3724	1	yart
5	880	to	3725	1	yelpers
6	730	t	3726	1	yers
7	718	was	3727	1	yit
8	671	he	3728	1	yourn
9	660	of	3729	1	yuther
10	538	we	3730	1	zip

問 1　表 1 を見て、上下 10 位にどんな語が出ているかを考えよ。

2　L 字型分布

「坊っちゃん」（夏目漱石, 1906）を単語に分割し、その頻度を計算し、頻度の多い方から順位をつけ、作図した。どんなテキストでも、単語の頻度・順位は L 字型 (図 2) のグラフになる (水谷 1975: 6)。

> **『坊っちゃん』 夏目漱石 1906 年（明治 39 年）**
> 親譲りの無鉄砲で小供の時から損ばかりしている。小学校に居る時分学校の二階から飛び降りて一週間程腰を抜かした事がある。なぜそんな無闇をしたと聞く人があるかも知れぬ。別段深い理由でもない。新築の二階から首を出していたら、同級生の一人が冗談に、いくら威張っても、そこから飛び降りる事は出来まい。弱虫やーい。と囃したからである。小使に負ぶさって帰って来た時、おやじが大きな眼をして二階位から飛び降りて腰を抜かす奴があるかと云ったから、この次は抜かさずに飛んで見せますと答えた。

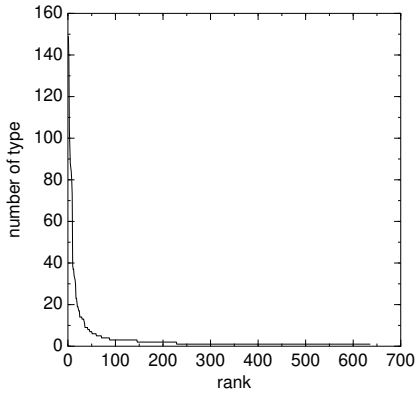

図 2: 夏目漱石「坊っちゃん」冒頭 15 行中の単語頻度・順位

リスト 1: 「坊っちゃん」中の単語頻度のコマンド

```
% cat bochan.txt| mecab | awk -F',' '{print␣$7}' | sort | uniq -c | sort -nr | nl |\
        awk '{print␣$1,$2}' | graph -Tps --x-label "rank" --y-label "number␣of␣type" > t.eps
```

表 2: 「坊っちゃん」冒頭の 15 行を形態素解析し、頻度上 10 位（左）、下 10 位（左）の一覧。

上 10 位	頻度	語	下 10 位	頻度	語
1	149	て	627	1	ざあざあ
2	138	た	628	1	さす
3	104	の	629	1	ごまかす
4	97	だ	630	1	ぐらぐら
5	88	は	631	1	ぐう
6	86	が	632	1	ぐいぐい
7	84	に	633	1	ぎゅうぎゅう
8	80	と	634	1	がた
9	73	を	635	1	うん
10	39	から	636	1	あんな

リスト 2: 「坊っちゃん」冒頭の 15 行を形態素解析し、上 10 位（上行）、下 10 位（下行）を抽出。

```
% head -15 bochan.txt| mecab | awk -F',' '{print␣$7}␣' | sort | uniq -c | sort -nr | nl | head -10
% head -15 bochan.txt| mecab | awk -F',' '{print␣$7}␣' | sort | uniq -c | sort -nr | nl | tail -10
```

問 2　表 2 を見て、上下 10 位にどんな語が出ているかを考えよ。また、表 1 と比較し、異同について話し合いなさい。

問 3　テキスト量が増えた場合、たとえば、15 行だけでなく、全文を対象としたときに、上下 10 位の頻度にどんな変化が出るかを考えよ。

リスト 3: 「坊っちゃん」のすべてを形態素解析し、上 10 位（上行）、下 10 位（下行）を抽出。

```
% cat bochan.txt| mecab | awk -F',' '{print␣$7}␣' | sort | uniq -c | sort -nr | nl | head -10
% cat bochan.txt| mecab | awk -F',' '{print␣$7}␣' | sort | uniq -c | sort -nr | nl | tail -10
```

問 4　単語ではなく、文字で計算した場合、プロットの形、上下 10 位の文字はどうなるかを考えよ。

問 5　「坊っちゃん」(1906 年) は 120 年ほど前の言語であるが、この法則は時代を越えて適用できるものであるかどうか、話し合いなさい。

リスト 4: 「坊っちゃん」に見られる文字の頻度順位の計算方法

```
% head -513 bochan.txt | texcount -freq - | tail -1892 | nl | head -1891 | awk '{print␣$1,$3}' | \
        graph -TX
```

3　ジップの法則

次の文を読んで、後の問いについて考えよ。

ジップの法則

ジップの法則（Zipf's law）とは、単語の出現頻度を集計し、その順位が、第 k 位だった場合、その単語の頻度が第 1 位の単語の出現頻度のほぼ $1/k$ であるという経験則である。たとえば、第 1 位の単語の出現頻度が 100 であった場合、第 2 位の単語の出現頻度は 50、第 3 位の単語の出現頻度は 33、第 4 位の単語の出現頻度は 25、... である。順位と頻度の間には次の法則があることがわかっている。

ジップの法則は次の式で表される。

$$f_r \simeq \frac{c}{r} \quad (r = 1, 2, 3, ...n)$$

ただし、頻度、順位、定数を f, r, c である。

この理論的説明はまだ成功していない。この法則は言語に限らず、さまざまな頻度と順位との関係にこの法則が当てはまることがわかっている。この法則に従う確率分布（離散分布）をジップ分布という。当初、ジップ自身は Principle of Least Effort という用語を用いて説明している (Zipf 1935, 1949)。

問 6　「経験則」とはどういう意味か。英語では何と言われているか。

問 7　ある分布が「この法則に従う確率分布」であることを調べるにはどんな統計検定をすれば良いか。

問 8　「離散分布」の「離散」とはどういう意味か、具体例を用いて説明せよ。

問 9　“Principle of Least Effort” は「最小努力の法則」と訳されるが、なぜ語の使用順位 (rank) と使用頻度 (frequency) の関係を「最小努力の法則」と呼んだのか。

問 10　「言語に限らず、さまざまな頻度と順位との関係にジップの法則が当てはまることが明らかにされている」と言われているが、言語以外にどのようなものに当てはまるかを議論せよ。

問 11　実は $f_r \simeq \frac{c}{r}$ $(r = 1, 2, 3, ...n)$ は前の「坊っちゃん」のデータには当てはまらない。ならば、どのよ

うな代替案があるかを調べてみよ。

4 ジップの法則が現れる時

次のような現象にジップの法則が成り立つことが確認されている (Wikipedia.org)

1. 単語の使用頻度
2. ウェブページへのアクセス頻度
3. 都市の人口
4. 上位 3% の人々の収入
5. 音符の使用頻度
6. 細胞内での遺伝子の発現量
7. 地震の規模
8. 固体が割れたときの破片の大きさ

問 12 上のそれぞれについて、なぜジップの法則にしたがうのかを考えよ。

問 13 なぜジップの法則が存在するのかを話し合いなさい。

5 関連する概念

ジップの法則は冪乗則（Power law）の一種である。また、ジップ分布は変数変換によりパレート分布（連続分布）と同じ形になる。パレート分布の離散型である。パレートの法則はパレート分布の特別な場合に当たり、また 80-20 の法則とも関係がある。順位規模の法則とも呼ばれる。(Zipf 1935, 中野 1998, 宮島 1990)

6 テキストの統計量

文字や語の数、単語の文字列や一文の長さなど、テキストの要素の特徴を定義する方法には、いろいろなものがある。以下の統計量について調べてみよ。

1. 頻度スペクトル
2. タイプトークン比
3. ユールのK特性値

7 もっと知るには

1. 同志社大学のこちらのページには、テキストの統計量の話がまとめてある
2. “Information on Zipf’s law” には、毎年の Zipf’s law に関する論文の一覧が更新され続けている
3. 「インターネット・マガジン」1999 年 6 月号、310 ページの「後藤滋樹の『新・社会楽』第 53 回『不思議な法則』」には、ジップの法則が簡単に紹介されている
4. 安本 (1960) は、色彩語の分布はポワソン分布よりポリヤ・エゲンベルガー分布に近い分布型にしたがっていることを調査した (安本 1960: 58)。その他、「文の長さ」「比喩の数」「マルの数」「名詞の数」の場合の分布について詳しくわかりやすい文章でまとめてある。
5. 語彙の分布問題に関する日本語の文献としては (水谷 1983: 86–120) が詳しい。最近出版された分布にまつわる解説書としては岩沢 (2016) がある。
6. ベンフォードの法則というものもある。自然界に存在する数値の最初の桁が 1 である確率はほぼ 30% であるというもの。これも簡単に計算できる。
7. ’texcount’ はオンラインでも利用できる。TeXcount web service

参考文献

岩沢宏和 (2016)『分布からはじめる確率・統計入門：実用のための直感的アプローチ』, 分布からはじめる確率統計入門：実用のための直感的アプローチ, 東京図書, 東京, Japan.

宮島達夫 (1990)「単語の使用度数と長さ・古さ」,『計量国語学』, 第 17 巻, 第 6 号, 287–300.
水谷静夫 (1975)「短い作品の語彙の量的構造 昭和初期流行歌の調査から 1 」,『計量国語学』, 第 72 巻, 1–12.
——— (1983)『語彙』, 第 2 巻, 朝倉日本語新講座, 朝倉書店, 第 1 版.
中野洋 (1998)「言語の統計」,『言語情報処理』, 岩波書店, 第 4 章, 149–199.
安本美典 (1960)『文章心理学の新領域 (文芸作品の科学的理解はいかになされるか)』, 創元社, 東京.
Zipf, George Kingsley (1935) *The Psycho-Biology of Language*, Boston-Cambridge Mass.: Houghton Mifflin.
——— (1949) *Human Behavior and The Principle of Least Effort, An Introduction to Human Ecology*: Addison-Wesley Press Inc.