修士論文 要旨

# 自然言語の Taylor 則に関する研究

数理情報学専攻 48176208　小林 達
指導教員　田中 久美子 教授

## 1　はじめに

自然言語とは我々が普段から使う日本語や英語といった言語のことを指す．自然言語には Zipf 則 [7] や Heaps 則 [4] といった経験則が存在することが知られている．本研究の対象である Taylor 則 [5, 6] は，時間あたりの事象の発生回数や空間あたりの個体の分布などについて，その分散が平均のべき乗に比例するという法則である．

本研究では単語の出現を事象の発生とみなして自然言語に Taylor 則を適用する手法を提案し，その手法に基づいて 1400 以上の自然言語や関連したテキストにおける Taylor 則を解析する．さらに，Taylor 則の指数が自然言語に近い値をとり，他の自然言語の経験則も満たす言語モデルとしてランダムネットワーク上のランダムウォークを考える．

## 2　関連研究

### 2.1　自然言語の経験則

自然言語にはいくつかの経験則が存在し，言語や文章の種類にかかわらず成立する．以下ではそのうちのいくつかについて述べる。

Zipf 則 [7] は，文章中に出現する単語を頻度が大きい順に並べたときに，それぞれの単語の頻度が順位の $-1$ 乗に比例するという法則である．

Heaps 則 [4] は，文章の冒頭 $N$ 単語中に $V(N)$ 種類のことなる単語が含まれているとすると，$V(N)$ が $N$ のべき乗に比例するという法則である．

長相関 [1] は文書中の単語の出現について，文をまたいだスケールでの相関が観測される現象で，べき則の形で相関が減衰することによる定式化が多く用いられている．

### 2.2　Taylor 則

Taylor 則 [5, 6] は，事象の発生回数や空間的な分布について，標準偏差 $\sigma$(分散 $\sigma^2$) が平均 $\mu$ のべき乗に比例するという法則である．

$$\sigma \propto \mu^{\alpha}, \tag{1}$$

Taylor 則は [2] は生態学，物理学，経済学から人間工学，ネットワークなど幅広い分野で報告されている [2] が，自然言語に関する例 [3] は少なく、自然言語において Taylor 則が成立する意味を深く考察していない。

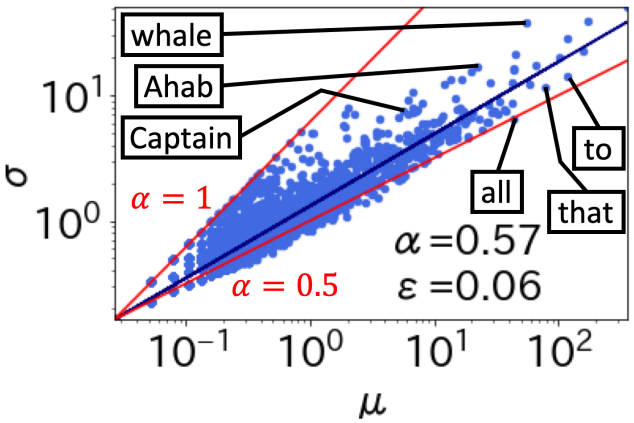


図 1:　*Moby Dick* における Taylor 則の結果．$\Delta t = 5620$ としており、横軸、縦軸はそれぞれの単語の $\Delta t$ 単語あたりの頻度の平均 $\mu$ と標準偏差 $\sigma$ である。直線は $\sigma = \hat{c}\mu^{\hat{\alpha}}$ であり、$\hat{c}, \hat{\alpha}$ は (3) によって求めた。直線の傾きが Taylor 指数 $\alpha$ に対応し、赤色で示した直線はそれぞれ傾き 1 と 0.5 である。

## 3　自然言語の Taylor 則

解析対象の単語列を $X = X_1, X_2, \ldots, X_N$ とし，単語列中に含まれる単語の集合を $W$ とおく．本研究では，単語列 $X$ を $\Delta t (\in \mathbb{N})$ 単語ごとに区切り，各単語 $w_k \in W$ について $\Delta t$ 単語ごとに何回現れるかを計測し，出現回数の平均 $\mu_k$ と標準偏差 $\sigma_k$ を求める．Taylor 指数 $\alpha$ と比例定数 $c$ の推定値 $\hat{\alpha}, \hat{c}$ は

$$\hat{c}, \hat{\alpha} = \arg\min_{c,\alpha} \varepsilon(c, \alpha), \tag{2}$$

$$\varepsilon(c, \alpha) = \sqrt{\frac{1}{|W|}\sum_{k=1}^{|W|}(\log \sigma_k - \log c\mu_k^{\alpha})^2}, \tag{3}$$

によって求めた．

Taylor 指数 $\alpha$ は単語列に対して経験的に $0.5 \leq \alpha \leq 1.0$ を満たす。各単語が独立同一分布に従って生成された単語列では、$\alpha = 0.5$ であり、各単語がそれぞれ一つの $\Delta t$ の区間の中にかたまって現れる場合は $\alpha = 1.0$ である。本研究で扱った実データにおいては $0.5 < \alpha < 1.0$ であった。

図 1 は *Moby Dick*(英語) に、$\Delta t = 5620$ として

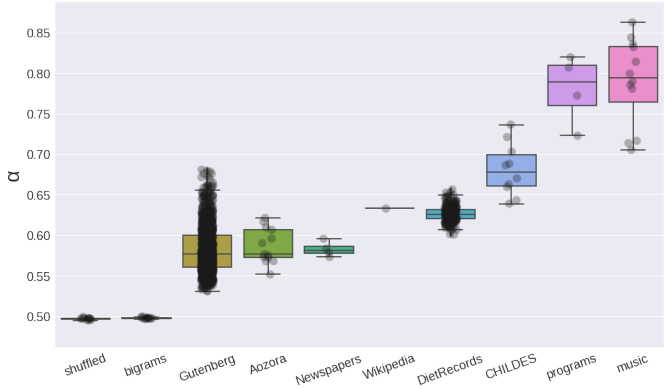


図 2: $\Delta t = 5620$ の場合における Taylor 則の指数 $\alpha$ の文章の種類ごとの分布.

Taylor 則を適用した結果である．横軸を $\mu$, 縦軸を $\sigma$ とした両対数軸上に，各単語を各点で，回帰直線 $\sigma = \hat{c}\mu^{\hat{\alpha}}$ とともに示した．$\alpha$ はこの直線の傾きに相当し，*Moby Dick* においては 0.57 であった．この値は文章の長さにほとんど依存しないため，長さの異なる文章に対しても同様の手法を適用することが可能である．

図 1 において、$\mu$ が大きく $\sigma > \hat{c}\mu^{\hat{\alpha}}$ となる単語の中には文章のキーワードである単語が含まれ、$\sigma < \hat{c}\mu^{\hat{\alpha}}$ となる単語の中には意味的には重要ではないが文法的な役割を果たす単語が含まれる。一方、$\mu$ が小さい単語の中で、$\sigma$ の最大値と最小値に注目すると、それぞれほぼ傾き 1、傾き 0.5 の直線に沿っている。

図 2 に示すように，この $\alpha$ の値は文章の種類によって異なり，自然言語を単語によってシャッフルした文章では 0.50，自然言語の書き言葉では 0.58 付近，自然言語の話し言葉では 0.63 付近，音楽データやプログラミング言語では 0.80 付近であった．

## 4 ランダムネットワーク上でのランダムウォーク

人間が自然言語を生み出すメカニズムを再現するために様々な自然言語の数理モデルが考案されてきた．さまざまな数理モデルでの Taylor 則が自然言語と似た性質を示すのか否かは，Taylor 則そのものの非自明性を判断する基準になるとともに，言語モデルの自然言語らしさも評価しうるため，非常に興味深い．

しかし、既存の言語モデルの多くにおける Taylor 則の指数 $\alpha$ はほぼ 0.50 であった．ニューラルネットワークの中には自然言語に近い Taylor 則の指数を示すモデルもあったが，ニューラルネットワークから言語生成のメカニズムを考えるのは困難であるため、別に自然言語における Taylor 則の指数が 0.50 よりも大きくなる理由を説明しうるモデルを考えたい．本研究では，そのような言語モデルの一つとして，ランダムネットワーク上でのランダムウォークを考える．

本研究では様々なランダムネットワーク上で、様々な条件でのランダムウォークを試みたが、その中で Taylor 則や他の言語の経験則を最もよく満たすものは Barabási-Albert ネットワーク上での次数優先 (Preference) ランダムウォークであった。グラフ上での Preference ランダムウォークとは，各ステップで次に遷移するノードを現在のノードに隣接するノードの次数に比例した確率で選ぶというものである．このモデルによって生成された単語列では、各単語の $\mu, \sigma$ の分布，Taylor 則の指数 $\alpha$ ともに自然言語の結果と近くなった．

## 5 まとめ

本研究では，複雑系における経験則として知られる Taylor 則を自然言語に適用する手法を提案した．文章中の各単語の出現頻度の標準偏差は平均のべき乗に比例していて，その指数は音楽データやプログラミング言語，自然言語，自然言語を単語によってシャッフルした単語列の順に大きい値を示した．この値は文章中における単語出現のゆらぎに対応していると考えられる．

また，数ある言語モデルの中で，ランダムネットワーク上でのランダムウォークにおける Taylor 則の指数は自然言語のそれと近い値を示し，自然言語において単語出現のゆらぎが生じるメカニズムの一つを再現していると考えられる．

## 参考文献

[1] Werner Ebeling and Alexander Neiman. Long-range correlations between letters and sentences in texts. *Physica A*, 215:233–241, 1995.

[2] Zoltán Eisler, Imre Bartos, and János Kertész. Fluctuation scaling in complex systems: Taylor's law and beyond. *Advances in Physics*, pages 89–142, 2007.

[3] Martin Gerlach and Eduardo G. Altmann. Scaling laws and fluctuations in the statistics of word frequencies. *New Journal of Physics*, 16(11):113010, 2014.

[4] Harold S. Heaps. *Information Retrieval: Computational and Theoretical Aspects*. Academic Press, Inc., Orlando, FL, USA, 1978.

[5] H. Fairfield Smith. An empirical law describing hetero-geneity in the yields of agricultural crops. *Journal of Agriculture Science*, 28(1), 1938.

[6] L. Roy Taylor. Aggregation, variance and the mean. *Nature*, 732:189–190, 1961.

[7] George K. Zipf. *Human behavior and the principle of least effort: An introduction to human ecology*. Hafner, 1965.