文の長さ分布に見られる対数正規性

古橋　翔（東北大学大学院理学研究科）†

Lognormality of the Distribution of Sentence Length

Sho Furuhashi (Graduate School of Science, Tohoku University)

１．はじめに

言語学の一分野である計量文献学では、統計的手法により文献の分析が行われている。その際に着目する量として、単語の長さ、品詞の使用率や句読点の打ち方などがあり、文の長さ（文長）もその一つである。文長は英語など多くの言語で調べられており、日本語でも調べられている。その結果、文長分布は、文字数でみた場合、安本(1958)、佐々木(1976)と新井(2001)らにより対数正規分布、また佐々木によってガンマ分布の場合もあると報告されている。一方で、Ishida, Ishida(2007)により形態素数の場合は Hyper Pascal 分布であると報告されている。

文長分布の特徴に対する考察として、安本は、文長分布の対数正規性は Weber-Fechner の法則に依るのではないかと提案し、佐々木は、対数正規分布のモデルの一例として Kaptyen のアナログマシンの例を挙げている。また、ガンマ分布に対しては、佐々木は、文の構成要素の長さが指数分布に従い、この指数分布のたたみこみ分布であるから文長分布はガンマ分布となるのではないかと考察している。しかしながら、このような考察の一方で計量的な研究は行われていない。

対数正規分布は、文長分布以外でも多くの自然現象や社会現象で見られる。例えば、落下によるガラス破片のサイズ分布や論文の発表数の分布がある。近年、物理学の分野において対数正規分布が注目されるようになり、その生成メカニズムの研究が行われている。本研究では文長分布の対数正規性に着目し、この性質を既存のモデルにより説明できないか試みた。

２．サンプル

本研究では、インターネット上にある著作権の切れた作品を収蔵している青空文庫と京都大学大学院情報学研究科黒橋・河原研究室が提供している京都大学テキストコーパス Version 4.0 を利用した。これらの資料から文を収集するのだが、京都大学テキストコーパスはあらかじめ文ごとに区切られているが、青空文庫は独自に文章を文に分割しなければならない。本研究では、青空文庫の作品から次のような処理により文を収集した。

まず、クローラーを使用してインターネット上の青空文庫から出来る限り多くの作品のテキストファイルを収集する。この時、文字コードを Shift-JIS から UTF-8 へと変換した。次にファイルの最初と最後に作品情報が記載されているのでそれを削除する。この処理を行った後、次のルールに従いテキストファイルを選別する。

- ▲や□などの記号を含んでいない。
- 日本語のみで書かれている。
- 括弧の開閉の数が一致している。
- 詩、俳句などの韻文を含んでいない。
- 箇条書きを含んでいない。章立てになっていない。
- 脚本になっていない。

これらの条件を満たしていない作品は除外した。

残ったファイルを更に加工する。まず、青空文庫に収蔵されている作品は、電子化するに

† furuhashi@cmpt.phys.tohoku.ac.jp

当たり原本の情報を残すため注釈が書かれているので、これを取り除く。次に、踊り字をそれが表している文字に置き換えた。この置換は、形態素解析器などを利用するに当たり、踊り字のままだと誤りが多くなると考えたからである。以上のような処理を行った後、文章を句点で区切っていき文にばらしていった。最後に、得られた文を次のルールに従い選別した。

- カタカナと句読点のみの文ではない。
- 日本語の文字と句読点のみで構成されている。
- 文の長さに制限はない。

得られたサンプルは、青空文庫では、作品ファイルは 2213、著者は 150 名以上で、文の数は 116719、京都大学テキストコーパスでは、文の数は 38397 である。

３．一文当たりの文字数の分布

まず始めに、一文当たりの文字数l_cの分布を調べた。図 1 がその分布であり、平均 42.9 標準偏差 32.9 であった。但し、調べたのは青空文庫のみである。京都大学テキストコーパスは、形態素・構文情報のみ公開しており、元となる毎日新聞データは含まれていないからである。

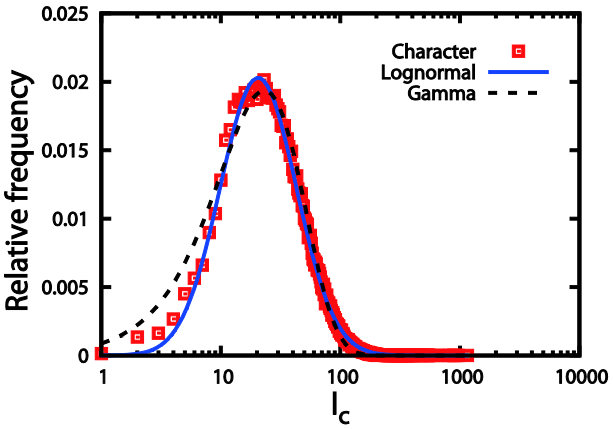

図 1　一文当たりの文字数l_cの分布（青空文庫）

図 1 の分布型が先行研究で報告されていた対数正規分布

$$f_{LN}(x)=\frac{1}{x\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(\ln x-\mu)^2}{2\sigma^2}\right\}, \tag{1}$$

とガンマ分布

$$f_G(x)=\frac{x^{k-1}}{\Gamma(k)\theta^k}\exp(-\frac{x}{\theta}), \tag{2}$$

のどちらに近いかを調べた。まず Gnuplot のフィッテイング機能を利用し、パラメータの値を推定した（表 1）。

表 1 パラメータ値（文字数）

	μ	σ^2	k	θ
青空文庫	3.557 ± 0.003	0.536 ± 0.003	2.40 ± 0.02	16.3 ± 0.2

次に、実データとフィッティングで推定したパラメータによる分布関数との差の累積を計算した。

$$R = \sum_{x=x_0}^{x_{\max}} |O(x) - E(x)|, \tag{3}$$

$O(x)$ は実データの長さ x である文の相対頻度、$E(x)$ は当てはめた分布関数である。x が大きい領域ではデータ点はまばらであり揺らぎが大きい。その影響を除くために、x の範囲はデータ点が多い領域に制限した。$x_0 = 1$、$x_{\max} = 242$ として、$E(x) = f_{LN}(x)$ の場合、$R = 0.061$ 、$E(x) = f_G(x)$ では $R = 0.096$ であった。よって、青空文庫から収集した文は、一文当たりの文字数 l_c の分布は対数正規分布に近い。

４．文の構造

日本語の文構造は、文節間の係り受け関係を表した依存構造木（図 2）で表現できる。依存構造木は、文節をノードとして係り元から係り先へ矢印を張り表現される。（矢印の向きを反対に書く場合もある）

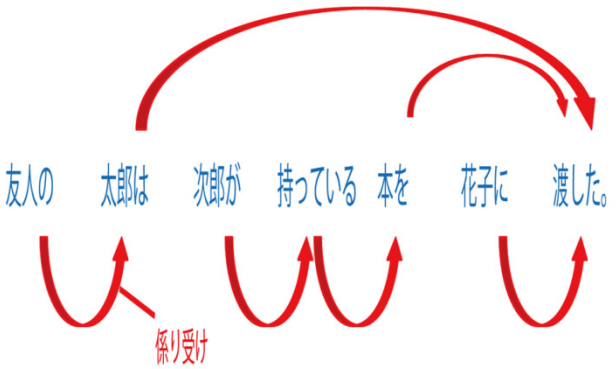

図 2　依存構造木

本研究では、依存構造木に着目して文長分布の対数正規性を生むメカニズムを調べた。京都大学テキストコーパスは既に依存構造木の情報が与えられているが、青空文庫には与えられていない。そのため、青空文庫の文に対して、依存構造木の情報を得るために、形態素解析に MeCab 0.98（辞書は MeCab-Ipadic）を用いた日本語係り受け解析器 CaboCha 0.60 pre4（TinySVM と YamCha なし）を使用した。

５．一文当たりの文節数の分布

依存構造木の構成単位が文節なので、一文当たりの文節数 l_s の分布を調べた（図 3, 4）。

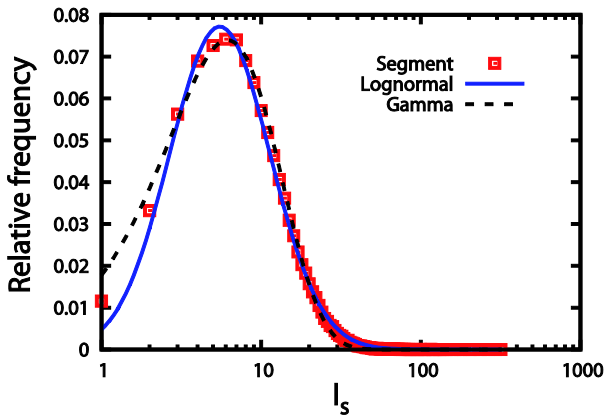

図 3　一文当たりの文節数 l_s の分布（青空文庫）

青空文庫では、平均が 11.1、標準偏差は 8.49、京都大学テキストコーパスの平均が 9.69、標準偏差は 5.27 であった。長さの単位を文字から文節へ切り替えたことで、文長分布型が変化するかどうか確かめるために、文字数の場合と同様に、対数正規分布 f_{LN} とガンマ分布 f_G のどちらがより当てはまるか R の値で比較した。フィッテイングにより得られたパラメータ値を表 2、R の値を表 3 に示す。

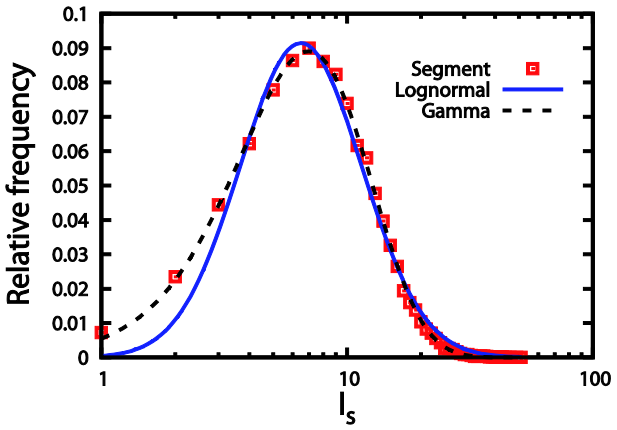

図 4　一文当たりの文節数 l_s の分布（京都大学テキストコーパス）

表 3 より、青空文庫では文字数の場合と異なりガンマ分布と対数正規分布では差が無く、京都大学テキストコーパスではガンマ分布の方が当てはまった。したがって、長さの単位を文節にすると、分布型はガンマ分布に近くなり、対数正規性は小さくなる。

表 2　パラメータ値（文節）

	μ	σ^2	k	θ
青空文庫	2.227±0.005	0.522±0.006	2.45±0.02	4.22±0.05
京都大学テキストコーパス	2.19±0.01	0.32±0.01	3.57±0.03	2.70±0.02

表 3　R の値（文節）

	青空文庫 （$x_0=1$ ，$x_{\max}=88$）	京都大学テキストコーパス （$x_0=1$ ，$x_{\max}=46$）
対数正規分布	0.071	0.11
ガンマ分布	0.070	0.024

６．乗算過程

対数正規分布を生み出すモデルの一つに乗算過程

$$X_n = \alpha_{n-1} X_{n-1} = \prod_{i=0}^{n-1} \alpha_i X_0, \qquad (4)$$

がある。変数 X_n は、ある分布に従う確率変数 α_i を独立に n 回掛け合わせて作られる。X_n を作る試行を多く繰り返すと、n が十分大きければ中心極限定理より X_n は対数正規分布に従う。

７． 依存構造木の枝分かれ過程

本研究では、乗算過程が文中の係り受け過程に表れるのではないかと考えた。係り受け過程を乗算過程と比較するために、依存構造木を図 5 のように書き換えた。図 5 は、葉である文節から係り受け関係に従い文節をまとめていき、最後に根である文になる過程を表している。

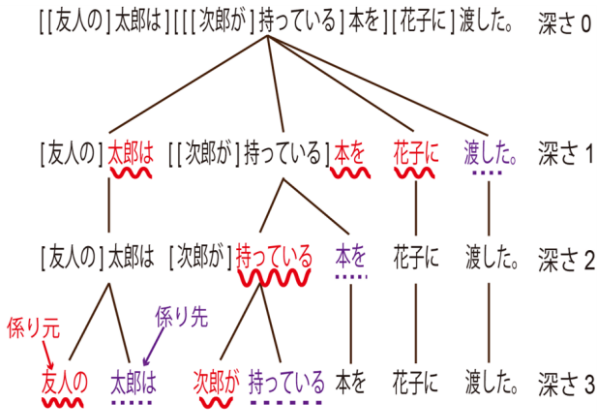

図 5　依存構造木

この依存構造木の作成方法は以下の通りである。

i. 他の文節から係られない文節の集合U を作る。
ii. 文節$b_j \in U$ に対して、その係り先c_j に係る文節が全てU の要素であるか確認する。
iii. ii. が確認された場合、c_j とそれに係る全ての文節をまとめてb_k とする。b_k はc_j の係り受け情報を引き継ぐ。U からc_j とそれに係る全ての文節を削除しb_k を追加する。
iv. U の要素数が一つになるまで ii と iii を繰り返す。

この依存構造木の枝分かれによるノード数増加の過程が乗算過程になっているのではないかと考えた。依存構造木の深さd におけるノード数をS_d として、もし枝分かれ過程が乗積過程であれば、式 4 より$\langle \ln X_n \rangle \propto n$ なので、$\langle \ln S_d \rangle$ はd に比例するはずである。よって、$\langle \ln S_d \rangle$ のd に対する変化を調べた。その際、依存構造木の葉の深さd_l が文ごとに異なる点に注意して、$S_d = l_s (d \geq d_l)$ とした。

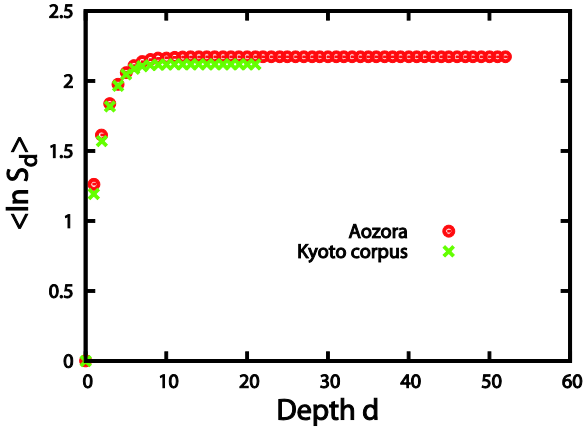

図 6　$\langle \ln S_d \rangle$ の深さd に対する変化

図 6 より青空文庫と京都大学テキストコーパスともに、$\langle \ln S_d \rangle$ はd に比例していなかっ

た。また、文節数l_sとd_lの関係を調べたところ、図６同様に比例関係になってはいなかった（図７）。よって、依存構造木からみた文構造に乗算過程は見られなかった。

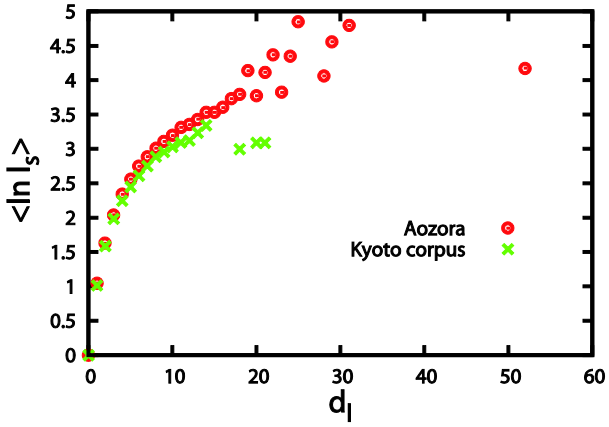

図 7　一文当たりの文節数の自然対数$\ln l_s$の平均と
依存構造木の葉の深さd_lの関係

８．まとめ

文長分布で報告されてきた対数正規性は、文構造に乗算過程が潜在しているのが原因ではなかった。今後は、一文当たりの文節数分布がガンマ分布に近いという結果から、佐々木の考察により文構造を説明できるか確かめるとともに、新たな視点から文長分布の対数正規性を研究していく。

文　献

安本美典(1958)「文の長さの分布型について」計量国語学, 4 号, pp.20-24.

佐々木和枝(1976)「文の長さの分布型」計量国語学, 78 号, pp.13-22.

新井　皓士(2001)「文長分布の対数正規分布性に関する一考察 ： 芥川と太宰を事例として」一橋論叢, 125 号 3 巻, pp.205-223. (http://hermes-ir.lib.hit-u.ac.jp/rs/handle/10086/10418 よりダウンロード可能)

Motohiro Ishida and Kazue Ishida (2007) On distributions of sentence lengths in Japanese writing, *Glottometrics,* 15, pp. 28-44.

関連 URL

Cabocha/南瓜　http://code.google.com/p/cabocha/

MeCab　http://mecab.sourceforge.net/

京都大学テキストコーパス Version 4.0　http://nlp.ist.i.kyoto-u.ac.jp/index.php?京都大学テキストコーパス