ID	KK280702
種別	調査報告
タイトル	文長分布型と係り受け関係に基づいた文構造の解析
Title	Analysis of Japanese Sentence Structure Based on the Sentence Length Distribution and Dependency Relations among Segments
著者	古橋　翔・早川　美徳
Author	FURUHASHI Sho・HAYAKAWA Yoshinori
掲載号	28巻7号
発行日	2012年12月26日
開始ページ	250
終了ページ	260 （英文要旨：p.276）
著作権者	計量国語学会

計量国語学第二十八巻第七号〔Mathematical Linguistics, Vol.28 No.7〕2012 年

調査報告

文長分布型と係り受け関係に基づいた文構造の解析

古橋　翔（東北大学）
早川　美徳（東北大学）

キーワード：文の長さ，係り受け，対数正規分布，負の二項分布

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，文体を特徴づける量として用いられ，文献の作者を特定する際に活用されている．

日本語の文長研究において，文字数を単位とした文長分布は，対数正規分布（安本（1958），佐々木（1976）と新井（2001）），やガンマ分布（佐々木（1976））に従うと報告されている．一方で，形態素数を単位とした場合は hyper-Pascal 分布（Ishida, Ishida（2007））とみるのが適切であるとの報告もある．

文長分布型の起源として，佐々木（1976）は，対数正規分布が得られる場合については，Kapteyn のアナログマシンを例に挙げ，文生成が乗算的（multiplicative）な確率過程と見なせる可能性を指摘している．また，ガンマ分布が得られる理由として，文の構成要素の長さが指数分布に従い，それらの畳み込みとして，文長分布がガンマ分布となる可能性についても言及している．

佐々木の考察で挙げられたモデルはいずれもシンプルで，仮に文生成がこのようなモデルで説明できるならば興味深い．本研究では，コーパスを用いて，佐々木が挙げたモデルを手掛かりとして文構造を解析した．文構造の表現には係り受け関係に基づく依存構造木を使用した．なぜなら，我々は，主語と述語，修飾や非修飾などの係り受け関係に注意しながら文を書いていると考えられ，また，日本語については，文構造を係り受け関係で表す方法が広く用いられているからである．

２．確率分布

この節では，先行研究で挙げられたガンマ分布および対数正規分布と，それらを生じさせる確率過程についてまとめる．

FURUHASHI Sho (Tohoku University), HAYAKAWA Yoshinori (Tohoku University) — Analysis of Japanese sentence structure based on the sentence length distribution and dependency relations among segments —

２．１　ガンマ分布

ガンマ分布は，形状母数kと尺度母数θの二つのパラメータを用い，

$$f_G(x)=\frac{x^{k-1}}{\Gamma(k)\theta^k}\exp\left(-\frac{x}{\theta}\right)(x\geq 0),\quad =0(x<0), \tag{1}$$

で定義される連続型確率分布である．指数分布$f_E(x)=\theta^{-1}\exp\left(-\theta^{-1}x\right)(x\geq 0)$に従う確率変数を$X_1,X_2,\cdots,X_k$とすると，その和$X_1+X_2+\cdots+X_k$はガンマ分布に従う．これは，佐々木が挙げたガンマ分布の生成モデルに対応する．

ガンマ分布の生成モデルで，指数分布を離散型確率分布である幾何分布$f_{\mathrm{GE}}(x)=p^x(1-p)(x\geq 0)$に置き換えると，幾何分布に従う確率変数$X_1,X_2,\cdots,X_n$の和$X_1+X_2+\cdots+X_n$は負の二項分布に従う．すなわち，ガンマ分布に対応する離散型確率分布は，負の二項分布，

$$f_{\mathrm{NB}}(x)=\binom{n+x-1}{x}p^x(1-p)^n\ (x=0,1,2,\cdots), \tag{2}$$

である．負の二項分布は，確率pのベルヌーイ事象が連続して「成功」する回数を調べる作業をn回繰り返したとき，「成功」の総数が従う分布である．本来，文長分布は離散型分布として考えるのが自然であるから，以下の議論では，ガンマ分布ではなく，負の二項分布について検討を行う．

２．２　対数正規分布

対数正規分布は，

$$f_{\mathrm{LN}}(x)=\frac{1}{x\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(\ln x-\mu)^2}{2\sigma^2}\right\}(x>0), \tag{3}$$

で定義される連続型確率分布である．対数正規分布に従う確率変数Xに対して，パラメータμは$\ln X$の平均値で，σ^2は$\ln X$の分散である．この分布は，後述するとおり，乗算確率過程の結果として得られることが知られている．

３．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で表現できる．

図１：依存構造木

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ることで構築され，係り受けは一般的に非循環性が仮定されているので，文全体の係り受け関係はツリーとなる．図 1 は「友人の太郎は次郎が持っている本を花子に渡した。」という文の依存構造木である．

4．確率分布の生成モデルと文構造の対応付け

本節では，負の二項分布と対数正規分布の生成モデルに文構造をどのように対応付けることが出来るか説明する．

4．1　負の二項分布の場合

2 節でも述べたように幾何分布に従う確率変数 $X_1, X_2, \cdots, X_n$ に対して，$X_1 + X_2 + \cdots + X_n$ は負の二項分布に従う．この生成モデルを依存構造木と対応づけるために，図 1 の依存構造木を図 2 のように表示し直す．

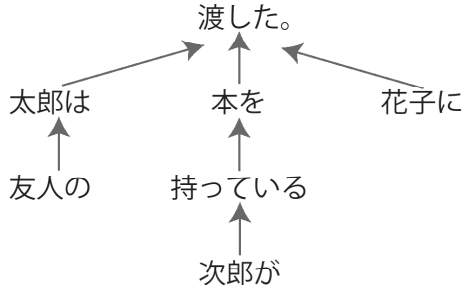

図 2：依存構造木．矢印は係り受けを表す．

図 2 の依存構造木では，ルート「渡した。」に対して，「友人の太郎は」は動作主を表す句，「次郎が持っている本を」は動作の対象を表す句，そして「花子に」は動作の方向を表す句である．従って，リーフからルートの子ノードまでの部分は，ルートに対する主格や対格等に対応する句である．このようなノードの集合は，依存構造木のルートの子ノードをルートとする部分木に対応する．本研究では，このような部分木を単純に「部分木」と呼ぶことにする．各依存構造木において，「部分木」は文中の出現位置に応じてラベル付けできる．図 2 の場合，「友人の太郎は」は 1 番目の「部分木」，「次郎が持っている本を」は 2 番目の「部分木」，「花子に」は 3 番目の「部分木」である．

負の二項分布の生成モデルを依存構造木に対応付ける場合，「部分木」を構成するノード数を長さと定義し，全ての依存構造木は n 個の「部分木」を持つと仮定する．実在する「部分木」の数が n より少ない場合，残りの「部分木」は長さ 0 とする．i 番目の「部分木」の長さを確率変数 X_i とし，もし，$X_1, X_2, \cdots, X_n$ が同じ幾何分布に従うならば，ルートに対応する文節を除いた文長 $X_1 + X_2 + \cdots + X_n$ は負の二項分布に従うことになる．

4．2　対数正規分布の場合

対数正規分布に従う確率変数を生み出すモデルの一つである，乗算確率過程を説明する．

ある確率変数 X_m が，

$$X_m = \alpha_{m-1} X_{m-1} = \prod_{i=0}^{m-1} \alpha_i X_0, \tag{4}$$

のように，ある分布に従う確率変数 α_i を独立に m 回掛け合わせて作られるとする．X_m を作る試行を繰り返すと，

$$\ln X_m = \sum_{i=0}^{m-1} \ln \alpha_i + \ln X_0, \tag{5}$$

より，m が十分大きければ中心極限定理から，$\ln X_m$ は正規分布に従う．よって，X_m は対数正規分布に従う．

係り受け過程を乗算確率過程に対応づけるために，依存構造木を図 3 のように書き換えた．図 3 は，リーフである文節から係り受け関係に従って文節をまとめていき，最終的にルートである文が完成する過程を表している．

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 の係り先は c_j の係り先とする．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．

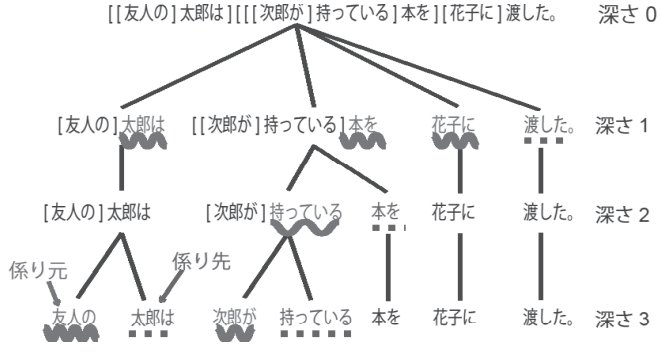

図 3：依存構造木．係り受け関係に基づき文節をまとめていき，文を生成する過程を表している．

この依存構造木の枝分かれによるノード数増加を，乗算確率過程に対応づける．すなわち，式（4），（5）に対し，深さ d のノード数を X_d，深さ $d-1$ から深さ d へのノード数の増加比率を α_{d-1} とする．ただし，X_d は整数，$\alpha_{d-1} = X_d / X_{d-1}$ より，α_{d-1} の取りうる値は $1/X_{d-1}$ の整数倍に制限される．よって，厳密には，α_d は深さごとに異なる分布を持つはずであるが，ここでは各深さで分布関数も変わらないと仮定する．

５．サンプル

本研究では，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[1]を用いた．京都大学テキストコーパスは，毎日新聞の記事に各種言語情報を人手で付与したテキストコーパスである．京都大学テキストコーパスは，そのウェブページ上のパッケージに含まれる形態素・構文・関係の付加情報と毎日新聞 1995 年版 CD-ROM に収納されている新聞記事で全体が構成される．しかしながら，本研究で必要なのは係り受け情報なので，毎日新聞 1995 年版 CD-ROM のテキスト情報は使用せず，パッケージ中の NUM ファイルに記録されている形態素・構文・関係の付加情報のみを用いた．

図 4 は，NUM ファイルの内容である．#で始まる行の S-ID は文 ID である．＊で始まる行は，二列目の数字は文節番号，三列目は，数字が係り先の文節番号，英大文字は係り受けの種類を表している．＊で始まる行以下が，その文節を構成する形態素情報である．

収録されている総文数は 38,400 であるが，本研究で用いた文数は 33,082 である．除外された文は，次のような文である．

```
# S-ID:950106080-002 KNP:97/01/20 MOD:
＊0 2D
＋0 2D
0-2 こんげつ＊名詞 時相名詞＊＊
2-1 ちゅう＊接尾辞 名詞性名詞接尾辞＊＊
3-1 に＊助詞 格助詞＊＊
＊1 2D
＋1 2D
4-1 しん＊接頭辞 名詞接頭辞＊＊
5-2 せいど＊名詞 普通名詞＊＊
7-1 を＊助詞 格助詞＊＊
＊2 21D
＋2 26D
8-2 はっぴょう＊名詞 サ変名詞＊＊
10-2 する＊動詞＊サ変動詞 基本形
12-1 が＊助詞 接続助詞＊＊
13-1 、＊特殊 読点＊＊
```

図 4：京都大学テキストコーパスの NUM ファイル

- 自分自身に係る文節を含んでいる文．
- 格関係，照応・省略関係，共参照タグ付きコーパスに含まれる文 ID（rel.dat ファイルに記載されている）に対応している文．京都大学テキストコーパスは，形態素・構文タグ付きコーパスと格関係，照応・省略関係，共参照タグ付きコーパスで構成されている．この中で，格関係，照応・省略関係，共参照タグ付きコーパスは，形態素・構文タグ付きコーパス中の約 5,000 文を使用して作られており，重複しているため除外．
- 係り受けのマークが I となる係り受け関係を含んでいる文．I でマークされた係り受けは，係り先のない文節に対して定義された係り受けであるため除外．

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

先行研究で挙げられたモデルの妥当性を確かめるには，先行研究と同じコーパスを用いるべきであるが，係り受けデータが利用可能ではなかったため，係り受け関係を付与されている既存のコーパスを用いることにした．

６．文長分布

文長分布として負の二項分布と対数正規分布のどちらがより妥当であるか検討する．

本研究では，文長は一文を構成する文節数と定義した．なぜなら，依存構造木の構成単位が文節であり，また，京都大学テキストコーパスの形態素・構文・関係の付加情報を用いた解析になじむからである．

本研究で設定した文長の単位は，文長の単位を文字とする佐々木の設定と異なる．したがって，解析を始める前に，文長の単位の違いによる影響を検討しなければならない．京都大学テキストコーパスでは，テキストがないので検討出来ない．代わりに現代日本語書き言葉均衡コーパス（BCCWJ）を用いた．BCCWJ の C-XML 下の PN フォルダにある XML ファイルから毎日新聞（人名 ID 258,099）の記事を収集する．収集した記事から<sentence>タグに挟まれた文字列を文とし，<sentence>タグの種類が“quasi”ではない，会話文やキャプションなどではなく地の文，等の条件を満たすものを対象とした（約 1,300 文）．日本語構文・格解析システム KNP で係り受け解析して，一文当たりの文節数 l_s と文字数の分布を比較した．図 5 では，一文当たりの文節数 l_s の分布と平均値が一致するようにスケール変換した文字数分布を比較しており，分布はほぼ一致する．したがって，文長の単位によらず本研究の解析結果は成り立つと考えられる．

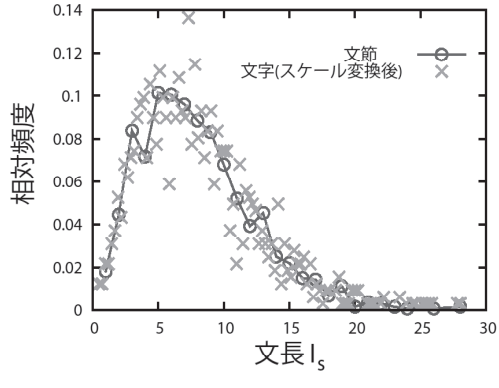

図 5：一文当たりの文節数 l_s 分布と文字数分布の比較（BCCWJ を使用）．
文字数分布は，文節数分布と平均値が一致するようにスケール変換を施した．

図 6 に文長 l_s の分布を示す．対数正規分布と負の二項分布のパラメータには，最尤法による最尤推定値を用いた．式（2）より，負の二項分布は本来パラメータが 2 つであるが，本研究では n をあらかじめ与え p のみを推定した．

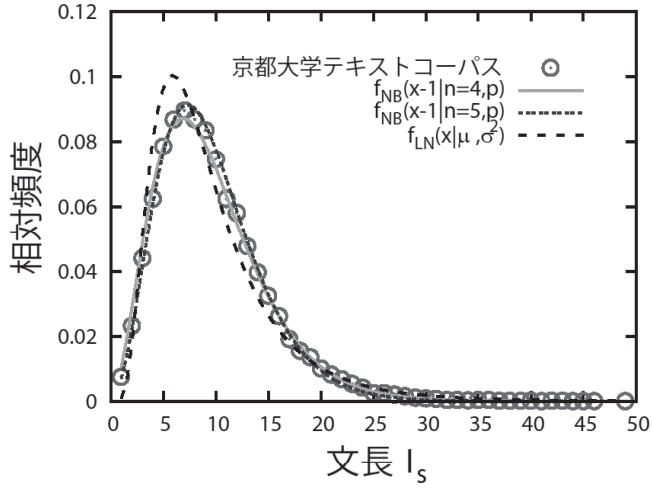

図 6：文長 l_s の分布

n は整数でなければならないため，負の二項分布の二項係数をガンマ関数で近似した上で，Gnuplot の関数フィッティング機能で求めた $n=4.5$ に近い整数 $n=4,5$ を用いた．最尤推定値は，$\mu=2.1,\sigma^2=0.33,p=0.68(n=4),p=0.63(n=5)$ である．

文長の分布型のモデルとして，対数正規分布と負の二項分布のどちらがより適切であるかを，赤池情報量規準（AIC）で評価した．AIC は，

$$\mathrm{AIC}=-2\times(\text{最大対数尤度})+2\times(\text{パラメータ数}), \tag{6}$$

で定義され，AIC の値が小さいほど当てはまりが良いとされている．AIC は，197,048（対数正規分布），195,929（$n=4$ の負の二項分布），196,091（$n=5$ の負の二項分布）であり，殆ど差は見られない．しかしながら，図 6 を見る限り，対数正規分布は文長分布からのずれが顕著にみられる．実際に，ずれの程度を調べるため，

$$R=\sum_{i=1}^{N}\frac{|E(x_i)-f(x_i)|}{NE(x_i)}, \tag{7}$$

を計算し比較した．ここで，$\{x_i\}(1\le i\le N)$ は測定データ，$E(x_i)$ は測定データの相対頻度，$f(x_i)$ はモデル分布である．モデル分布が対数正規分布の場合 $R=0.15$，負の二項分布 $(n=4)$ の場合 $R=0.044$，負の二項分布 $(n=5)$ の場合 $R=0.044$ となり，明らかに対数正規分布での残差が大きい．したがって，本研究で用いたコーパスについては，文長分布は対数正規分布に較べ負の二項分布により近いと言える．

7．乗算確率過程モデルの妥当性

依存構造木の深さ d におけるノード数を S_d とする．もし枝分かれ過程を乗算確率過程と見なすことができれば，4.2 節から，$\alpha_{d-1}=S_d/S_{d-1}$ とし，

$$\begin{aligned}\langle \ln S_d \rangle &= \sum_{i=0}^{d-1} \langle \ln \alpha_i \rangle + \langle \ln S_0 \rangle, \\ &= d\langle \ln \alpha \rangle + \langle \ln S_0 \rangle,\end{aligned} \tag{8}$$

が成り立つ．$\langle\ \rangle$はサンプル文に対する平均である．したがって，$\langle \ln S_d \rangle$はdに比例するはずである．実際に，$\langle \ln S_d \rangle$のdに対する変化をプロットした結果を図 7 に示す．その際，図 3 の依存構造木のリーフの深さd_lが文ごとに異なる点を考慮して，$S_d = l_s (d \geq d_l)$とした．

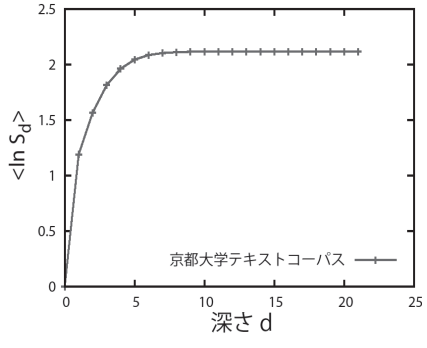

図 7：$\langle \ln S_d \rangle$の深さdに対する変化

図 7 から，明らかに$\langle \ln S_d \rangle$は深さdに比例していないことが判る．

次に，リーフの深さd_lとリーフの深さd_lが同じ文の文長の対数平均$\langle \ln l_s \rangle$の関係を調べた．もし，依存構造木の枝分かれ過程が乗算確率過程で説明できるならば，$\langle \ln l_s \rangle \propto d_l$となるはずである．結果，図 8 より$\langle \ln l_s \rangle$は$d_l$に比例していない．したがって，依存構造木の枝分かれ過程を単純な乗算確率過程と見なすことには無理がある．

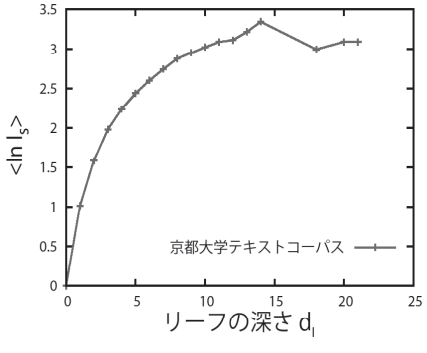

図 8：$\langle \ln l_s \rangle$とリーフの深さd_lの関係

８．依存構造木の生成モデルとしての負の二項分布の生成モデルの妥当性

依存構造木について，負の二項分布の生成モデルによる予想と実際の構造を比較する．まず，「部分木」の数に着目する．4.1 節の負の二項分布の生成モデルにおいて，実在する「部分木」の数は 0 より大きい値をとる確率変数の個数に対応し，その分布は二項分布，

$$f_B(x) = \binom{n}{x} p^x (1-p)^{n-x} (x = 0, 1, \cdots, n), \tag{9}$$

となる．

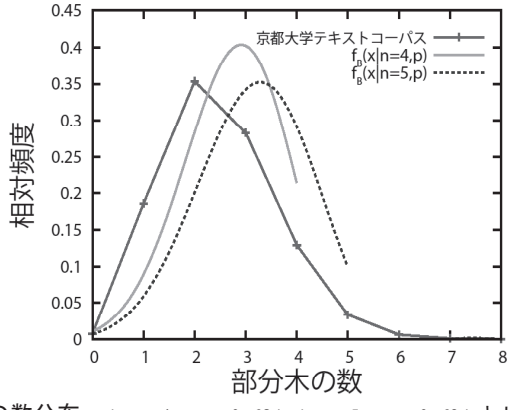

図 9：「部分木」の数分布． $(n = 4, p = 0.68), (n = 5, p = 0.63)$ **とした二項分布も表示．**

図 9 に，一文を構成する「部分木」の数分布を，文長分布で得た最尤推定値にパラメータを設定した二項分布とともに示す．結果，明らかに，実データは負の二項分布の生成モデルから大きくずれる．

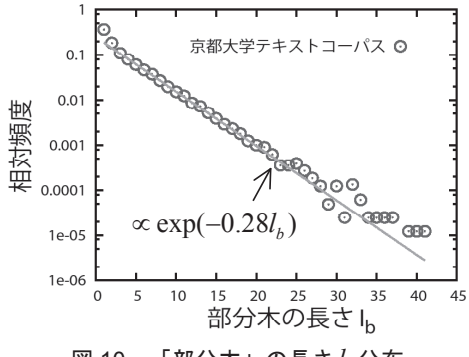

図 10：「部分木」の長さ l_b 分布

次に，実在する「部分木」の長さ l_b 分布を確かめる．サンプル全体の「部分木」の長さ l_b 分布を図 10 に示す．図 10 より，「部分木」の長さは大まかに指数分布に従うもの

の，「部分木」の長さ 1 は明らかに指数分布から有意にずれており，極端に多い．

最後に，実在する「部分木」の数で依存構造木を区別し，それぞれで「部分木」の長さ分布を調べた結果を図 11 に示す．

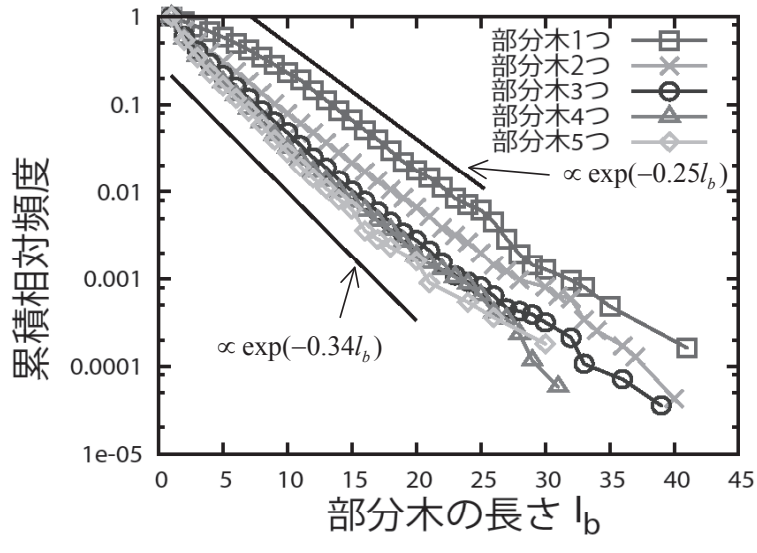

図 11：「部分木」の数ごとの「部分木」の長さ l_b 分布．グラフは直線（$\propto \exp(-0.34l_b)$）から直線（$\propto \exp(-0.25l_b)$）の間である．「部分木」が 6 つ以上は，データ数が少ないで省略した．

負の二項分布のモデルでは，これらの「部分木」の長さ分布が，着目した「部分木」の数に依らず同じ分布に従うと仮定されているが，現実には，その分布型は顕著に異なる．これらの結果から，負の二項分布の生成モデルで仮定されている条件を，実データは満たしていないと考えられる．

以上のように，パラメータ n を定数とした負の二項分布の生成モデルでは，文構造をよく説明できなかった．しかしながら，図 11 のように，「部分木」の長さは全体として指数分布に従っていることから，n を実在する「部分木」の数とした，負の二項分布の複合分布として説明できる可能性がある．実際，佐々木（1976）もこの描像について言及している．

９．結論

係り受け関係に基づいた文構造は，佐々木が指摘した単純な乗算確率過程や負の二項分布の生成モデルでは説明できなかった．

今後の課題は，まずは 8 節の最後に示した描像が成り立つかどうかを確かめることである．仮にその描像が成り立つのであれば，文生成の基本はベルヌーイ過程とみなせることになる．つまり，一般に複雑と考えていた文生成が，統計的には，シンプルな過程の組み合わせとして表現できる．次に，新聞以外の小説等の異なるカテゴリーの文でも，同様

な解析を行うことも今後の課題である．既に，青空文庫に対して，乗算確率過程に注目した同様の解析を既に行い，本研究と同様の結論を得ている（Furuhashi・Hayakawa, 2012）．しかしながら，カテゴリーや作者を考慮せず文を収集しているので，得られた結論がこれらの要因を考慮した場合でも成り立つかどうか不明である．したがって，今後，作品のカテゴリーや作者ごとに解析を行う予定である．

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, 20-24.

佐々木和枝（1976）文の長さの分布型『計量国語学』, 78 号, 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, 205-223.

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

S. Furuhashi and Y. Hayakawa (2012) Lognormality of the Distribution of Japanese Sentence Lengths, *Journal of the Physical Society of Japan*, 81, 034004-1-034004-5.

謝辞　本研究に際して，川勝年洋教授には，間違いを指摘頂くなど有益なコメントを頂き，感謝いたします．また，遠藤大樹氏には，有益な議論と助言を頂きました．最後に，京都大学テキストコーパスを公開して下さっている京都大学大学院情報学研究科黒橋・河原研究室に感謝いたします．

（2012 年 10 月 14 日受付）

Report

Analysis of Japanese Sentence Structure Based on the Sentence Length Distribution and Dependency Relations among Segments

FURUHASHI Sho (Tohoku University)
HAYAKAWA Yoshinori (Tohoku University)

Keywords: sentence length, dependency relations, log-normal distribution, negative binomial distribution

Abstract:
In previous studies about Japanese sentence length, two different types of the model for sentence generation have been proposed by Sasaki: one is a multiplicative stochastic model resulting in log-normal sentence length distribution and the other is an additive stochastic model resulting in the negative binomial sentence length distribution.

In the present study, motivated by Sasaki's suggestion, we examined the structure of dependency trees and checked whether those models could explain the obtained structure of dependency trees. To do that, we used Kyoto University Text Corpus (33,082 sentences) which includes the information of dependency relations among segments.

As a result, we found that the structure of the dependency trees did not accord with the expectation of multiplicative nor additive stochastic process.