The University of Osaka
Institutional Knowledge Archive

Title	文の長さの統計モデル
Author(s)	石井, 正彦
Citation	現代日本語研究. 2017, 9, p. 109-123
Version Type	VoR
URL	https://doi.org/10.18910/61379
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

https://ir.library.osaka-u.ac.jp/

The University of Osaka

文の長さの統計モデル

On Statistical Models of Sentence Lengths

石井 正彦
ISHII Masahiko

キーワード：文長、文構造、単文、複文、確率分布モデル、回帰モデル

要　旨

文の長さの統計モデルには、従来、単一の確率分布が想定され、対数正規分布やガンマ分布などのあてはめが試みられてきた。しかし、同じ確率分布モデルでも、文長において単文は有限、複文・重文は原理的に無限と仮定すれば、前者は左右対称に近い分布形、後者は右に裾を引く正に歪んだ分布形というように２つの分布型に分離できる。さらに、すべての文にはその従属節の数に応じた文長の限界がある（単文は従属節数０の文）と仮定すれば、従属節数を説明変数、文長を被説明変数とする単回帰モデル、あるいは、従属節の種類別に説明変数を設ける（被説明変数は同じく文長）重回帰モデルが成り立つ。ただし、回帰モデルの場合は、説明変数に文構造以外の要因を組み込む可能性があり、そのためには文長を決定する言語学的な機構の検討が必要となる。

1. はじめに

一見、単一の確率分布モデルで説明できそうな現象が、実は異なる複数の確率分布から成っている[1)]、ということがままある。個人所得の分布はその有名な例で、上位１％程度の金持ちの所得はベキ分布、残り 99%ほどを占める勤め人の所得は対数正規分布に従うといわれている（高安 2004)。同じ所得でも、ある程度上限がある給与所得と投資や投機で上限なく増やせる資産所得とではその性質が大きく異なり、したがって、分布型も異なるというわけである。

こうしたことは、言語現象にも認められる。文の長さはその有力な候補だろう。文は、回帰性や並列性という性質によって無限に長くすることができると

いわれる[2)]。回帰性とは「再帰性」「繰り返し性」ともいわれ、(1)(2)のように、ある種の言語的なまとまりの中にそれと同じ種類のまとまりが繰り返して現れ得るという性質である。

(1) ［［［［Ａさんの弟］の友達］の恋人］の姉］の……

(2) ［［［Ａさんが帰った］とＢさんが言った］のをＣさんが聞いた］と……

［［［Ａさんが帰った］のにＢさんが来た］からＣさんは残った］が……

(1)では回帰性によって長い句がつくられ、(2)では長い複文がつくられる。句の回帰は単文･複文を問わず起こり得るが､文の回帰は複文にのみ認められ、単文では生起しない（生起すれば単文ではなくなる）。文の回帰性に注目するなら、単文は有限、複文は原理的に無限ということになる。

並列性にも、(3)のように語や句を並べて並列句をつくる場合と、(4)のように文を並べて重文をつくる場合とがある。

(3) ＡさんとＢさんとＣさんと……

Ａさんの友達とＢさんの先輩とＣさんの後輩と……

(4) Ａさんが来て、Ｂさんが帰り、Ｃさんが現れ、……

語や句の並列は単文・複文を問わず起こり得るが、文の並列は重文にのみ認められ、単文では生起しない（生起すれば単文ではなくなる）。文の並列性に注目するなら、単文は有限、重文は原理的に無限ということになる。

このように、（語や句ではなく）文の回帰性・並列性からみれば、単文の長さは有限、複文・重文（以下「複文」で代表させる）の長さは（原理的には）無限ということになり、したがって、その文長の分布型も異なることが予想される。しかし、これまでの（日本語の）文長分布型の研究では、こうしたことはあまり考慮されなかったようである。

本稿では、まず、単文と複文とを上述のように質的に異なる単位体ととらえることで、確率分布モデルにもとづく単一の（と考えられていた）文長分布型が有限の単文の分布と無限の複文の分布とに分離できることを示す。一方、単文と複文とを（従属節の数という点で）連続する単位体ととらえることで、文長を従属節数によって説明する回帰モデルが想定し得ることをも示す。これらの検討を通して、単文と複文という、文の構造分類としてはごく単純な区別を考慮することが、文の長さの統計モデルを定める上で有効であることを確認す

る。ただし、回帰モデルの場合は、文構造（の複雑さ）を従属節数以外の尺度でとらえたり、文構造以外の説明変数を設定したりする余地が残されており、それらを検討するためには文長を決定する言語学的な機構の追究が必要であることをも述べる。

2. 確率分布モデル

文の長さについては、従来、その統計モデルとして単一の確率分布が想定され、対数正規分布やガンマ分布などのあてはめが試みられてきた（安本 1957、佐々木 1976、新井 2001、Ishida and Ishida 2007 など[3)]）。しかし、上述のように、原理的には無限の長さが可能な複文と、長さに限界のある単文とでは、その文長分布型も異なることが予想される。つまり、文長の分布型は単一ではなく、単文の分布と複文のそれとに分離できる可能性がある。

そこで、このことを、文長研究の先駆者である波多野完治が『文壇出世作全集』(1935 年刊）から抽出した（口語小説の地の文）480 文[4)] の文長データ（波多野 1950）を用いて、検証してみる。なお、安本(1960)は、この（文字数による）文長データについて、単文・複文の区別をすることなく、対数正規分布に従うとしている。

単文・複文の判別にあたっては、橋本(2003)を参考に、(5)のように単一の節からなる文を単文、(6)(7)のように複数の節からなる文を複文（重文も含む）とした（用例は、『文壇出世作全集』から)。

(5) 私はその儘楽屋へ上つた。

(6) お仙は羽織を持つて来なかつたので、一郎の妻がその白ぽい小紋の羽織を出してやつた。

(7) 日はきらめく、梢は濃い緑に飾られる、大地の草は時を得て勢ひよくのびて行く。

節は述語（述部）を含む複数の語からなるものとし、(8)(9)の下線部のように単独のものは節＝述語とは認めず（単文)、(10)(11)のように２語以上からなるものは節と認めた（複文)。また、(9)(12)の「見詰めてゐた」「思つたのであつた」などは単一の述語（述部）とした。

(8) 中に<u>古い</u>バツクなどが有つた。

(9) 一郎は驚いて妹を見詰めてゐた。

(10) 妹はしまひには声を立てゝ泣いた。

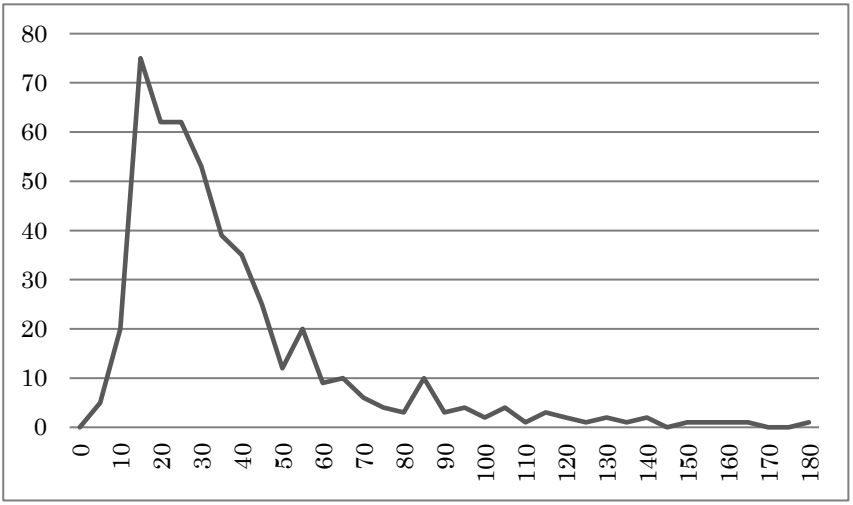
(11) 汽笛ががらんとした構内に響き渡つた。

(12) 爺さんは肚のなかで、然う思つたのであつた。

次ページ以降の図 1a・2a は、波多野のデータ全体の文長を文字数および文節数で測り、それぞれの度数分布を描いたもの、図 1b・2b は、同じデータを単文と複文とに分け、それぞれの度数分布を描いたものである。予想通り、文字数・文節数のいずれで測っても、長さに上限のある単文データは左右対称に近い分布形となり、原理的に上限のない複文データは右に裾を引く正に歪んだ分布形となる[5)]。なお、２標本コルモゴロフ=スミルノフ検定を行うと、文字数・文節数のいずれも $p<0.001$ となり、単文の分布と複文の分布とは有意に異なるという結果になる。以上のことは、文長もまた、個人所得の分布などと同じように単一の分布型ではなく、単文と複文という、異なる複数の確率分布から成っていることを推測させる。

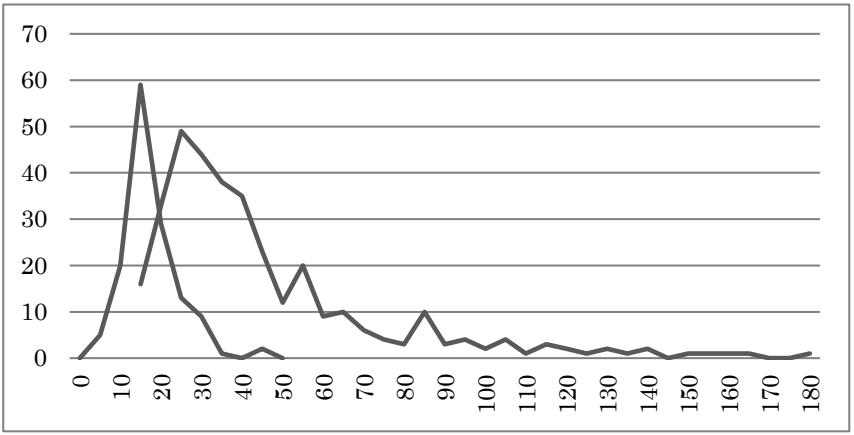
3. 回帰モデル

しかし、単文の文長と複文の文長とは、有限対無限と質的に二分することとは別に、従属節の数という点から連続的にとらえることも可能である（単文は従属節数が０の文、重文は文末の節以外を従属節とみなす）。複文がすべて無限の文長を持ち得ると考えるのではなく、従属節が１つの文にはそれによる文長の限界があり、また、従属節が２つの文にはそれに応じた文長の限界があり、……というように、すべての文にはその従属節の数に応じた文長の限界がある（取り得る従属節の数には限界がない）という考え方である。



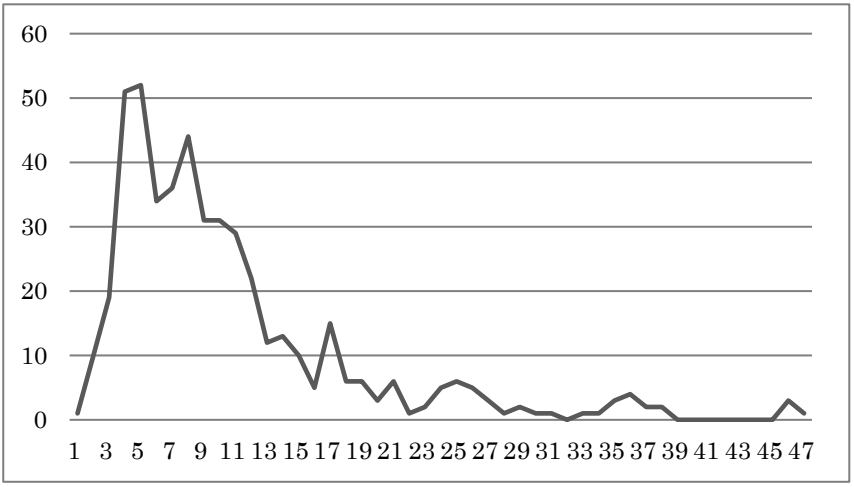
	N	平均	標準偏差	最小値	最大値
全体	480	35.2	27.7	1	179

図 1a：文長（文字数）の度数分布（全体）



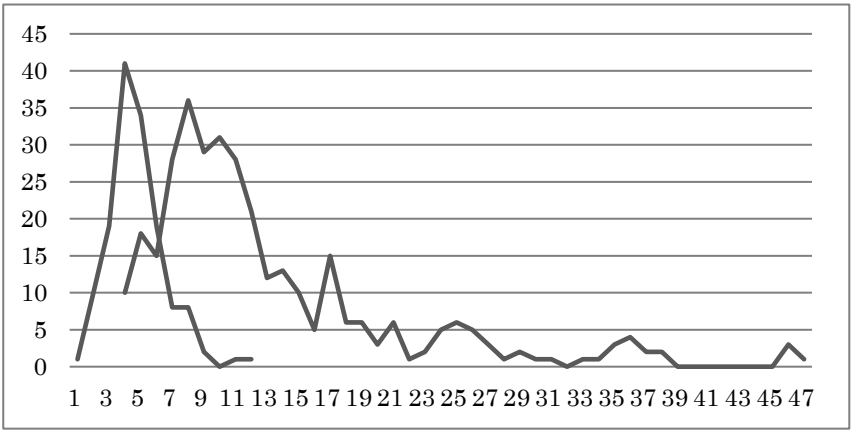
	N	平均	標準偏差	最小値	最大値
単文	144	15.3	6.1	1	42
複文	336	44.0	28.9	9	179

図 1b：文長（文字数）の度数分布（左：単文、右：複文）



	N	平均	標準偏差	最小値	最大値
全体	480	10.6	7.9	1	47

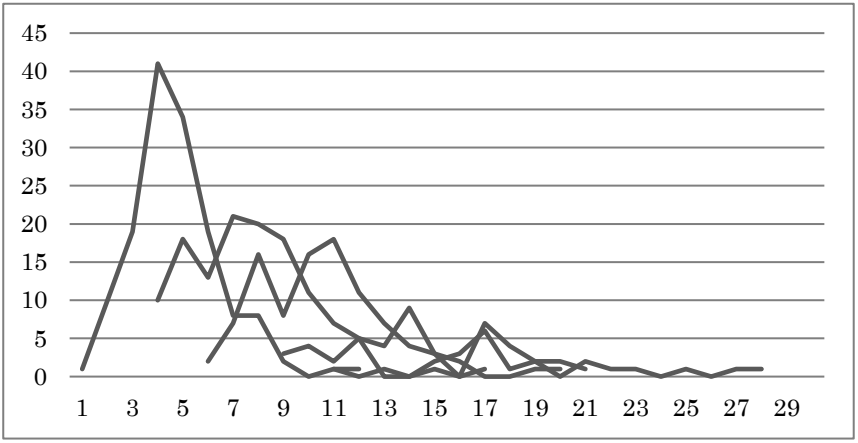
図 2a：文長（文節数）の度数分布（全体）



	N	平均	標準偏差	最小値	最大値
単文	144	4.8	1.8	1	12
複文	336	13.1	8.2	4	47

図 2b：文長（文節数）の度数分布（左：単文、右：複文）

そこで、文長測定の単位としてより一般的な文節数のデータ（図 2a）について、従属節の数ごとに文長の度数分布を描くと図 3a のようになる（従属節数の最大値は 12 であるが、ここでは４までの分布を示す）。また、図 3b は、見やすさを考慮して、従属節数ごとの度数分布を（縦軸のスケールを同じにして）分離したものである。これによれば、従属節数が増えるに従って文長の分布の位置は大きくなり、また、それぞれの分布の形はきれいな左右対称とはいえないが、いずれも同じような範囲（最大値－最小値）に収まっていて、従属節の数によって文長の範囲＝限界が定まることを示す結果となっている。



従属節数	0	1	2	3	4
N	144	125	96	46	23
平均	4.8	7.6	10.6	14.4	18.6
標準偏差	1.8	2.4	2.6	3.2	4.2
最小値	1	4	6	9	11
最大値	12	17	20	21	28

図 3a：従属節数別の文長（文節数）分布（左から、0→1→2→3→4）

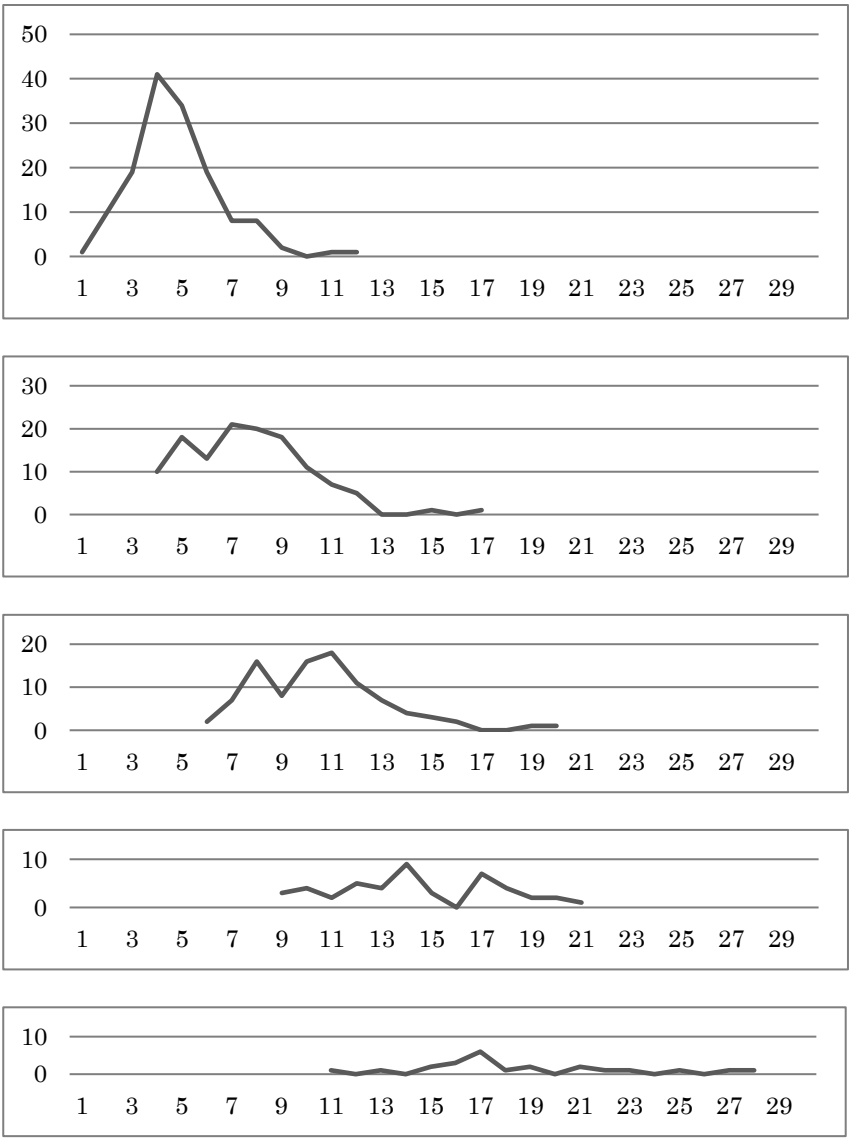


図 3b：従属節数別の文長（文節数）分布（上から、0→1→2→3→4）

試みに、従属節の数を横軸に、文長（文節数）を縦軸にして散布図を描くと（図４）、当然、従属節の数ごとに文長が縦にばらつくが、そのばらつきは同じような形状（範囲）をもってほぼ直線的な右上がりの傾向を示しており、両者の間には、従属節数を説明変数、文長を被説明変数とする単回帰モデルを想定することができる（$y=3.5744x+4.2134$、$R^2=0.8636$、$p<0.001$）。

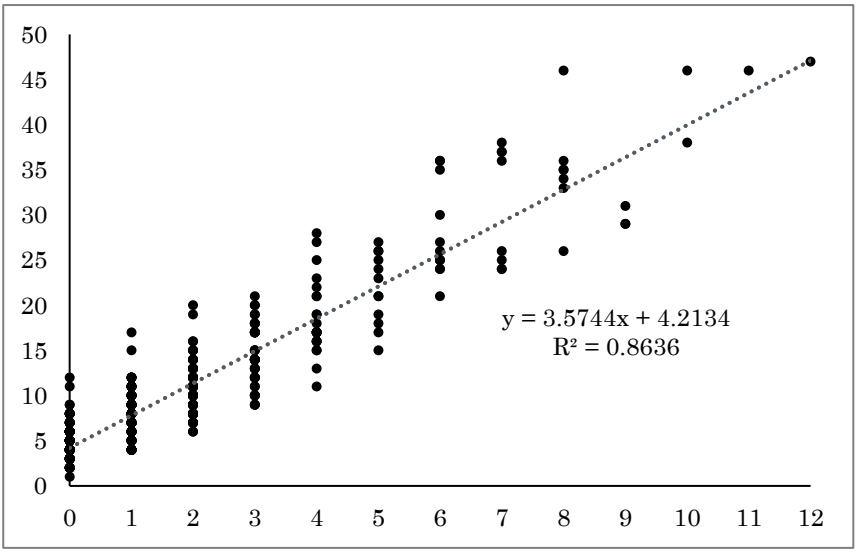


図４：従属節数（横軸）と文長（文節数）（縦軸）

ただし、こうした単回帰モデルでは、従属節の長さがほぼ一定であることが仮定される。しかし、南不二男の文の階層性の研究（南 1974・1993）などが示すように、従属節には様々なタイプがあり、その長さは一定でない可能性が高い[6)]。とすれば、種類の異なる従属節ごとにその数を説明変数とするような重回帰モデルを想定する必要があるかもしれない[7)]。

4. まとめと今後の課題

文の長さについて確率分布モデルを想定するときには、単文の分布と複文の分布とを区別し、それぞれの分布型を検討していくことが必要である。そこで

は、単文の長さは有限、複文の長さは無限という、両者の質的な違いを重視することになる。一方、回帰モデルを想定するときには、単文と複文とは従属節の数という点から連続的なものととらえられ、両者の違いというよりも直接的には従属節の種類とその長さが問題となる。その際には、一つの文がとり得る従属節の数は原理的には無限であっても、現実の文における従属節の数や種類はどのようにして決まるのかという問題もかかわってくるだろう。いずれにしても、以上の検討を通して、単文と複文という、文の構造分類としてはごく単純な区別を考慮することが、文の長さの統計モデルを定める上で有効であることが確かめられたといえる。

今後の課題としては、単文と複文とを質的に異なるものとする確率分布モデルの場合は、十分なデータを用意して分布のあてはめを行い、単文・複文それぞれの文長分布型を追究していくことが考えられる。

単文と複文とを連続的にとらえる回帰モデルの場合は、まず、文の構造を文長の説明変数とする際、その最適な測り方（尺度）を検討するという課題が考えられる。前節では従属節数を尺度としたが、他の方法も十分に考えられるからである。飯豊(1956)は、「文の長さと文の構造とを関係づけて考察すること」を明示した数少ない論考[8)]であるが、そこでは文構造の複雑さの尺度として述語に対するいわゆる「かかりの次数」が採用され、文長との関係が論じられている。かかりの次数とは、述語に直接かかる文の成分を第一次の成分、述語にではなくそれら一次成分にかかる文の成分を第二次の成分、という具合に、述語以外の文の成分の述語に対するかかり方の次数を求め、その最大値をその文の次数とするものである。この方法の特徴は、従属節数より細かい分類が可能になるということであり、たとえば、(13)(14)はともに従属節数０（の単文）、(15)は従属節数１（の複文）であるが、文の次数で分類すれば、それぞれ、１次・２次・３次の文として区別される。

(13) ［男が］［山道を］歩いている　……　１次の文

(14) ［男が］［［薄暗い］山道を］歩いている　……　２次の文

(14) ［男が］［［［日の］［差さない］］山道を］歩いている　……　３次の文

飯豊(1956)は、この文の（かかりの）次数と文長との関係を、志賀直哉の「雨蛙」と谷崎純一郎の「金と銀」で調査し、その結果を詳しい表[9)]にまとめてい

る。このデータを、文の次数を横軸、文節数で測った平均文長を縦軸として散布図に表すと、以下の図 5a・5b を得る。

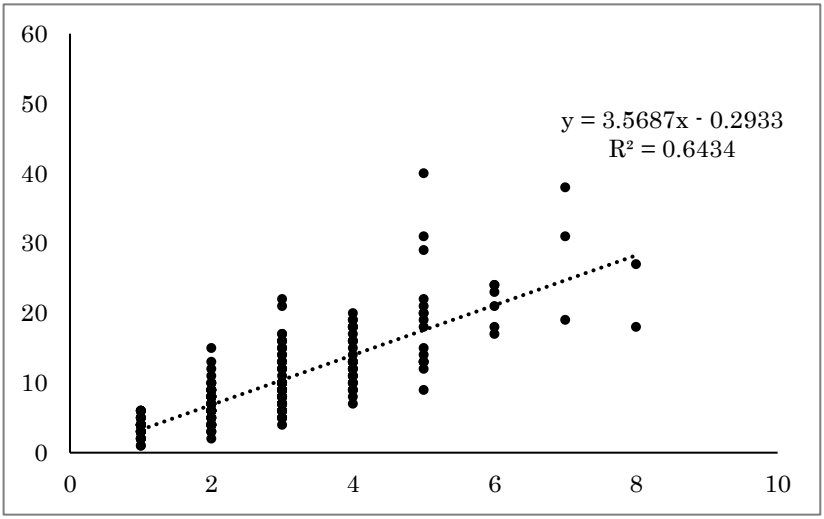


図 5a :「雨蛙」（横軸：文の次数、縦軸：文長平均値（文節数））

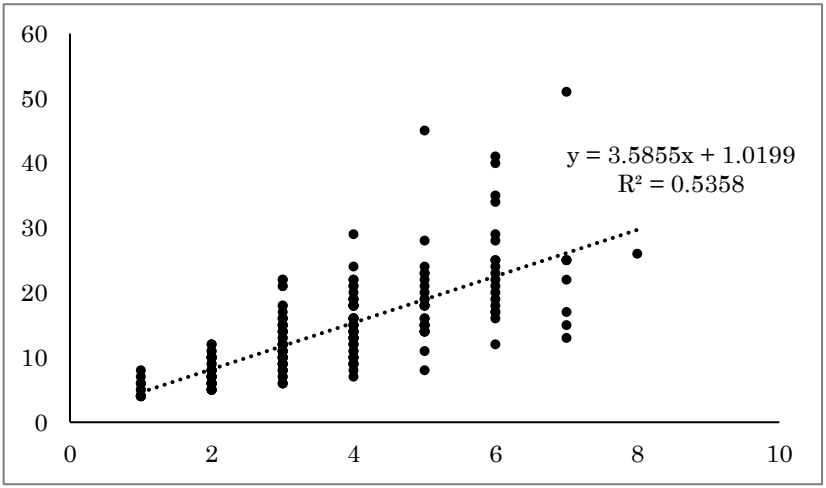


図 5b :「金と銀」（横軸：文の次数、縦軸：文長平均値（文節数））

いずれも、飯豊の指摘するように「次（文の次数）が高くなるに従って文の長さも増す」傾向、すなわち、従属節数と同様、文の次数によっても文長の範囲＝限界が定まることが示唆される結果となっている。

ただし、図 5a・5b に単回帰直線をあてはめても、図４に比べて決定係数の値も小さく、なにより文の次数が大きくなるに従って文長のばらつきも大きくなることから、従属節数の場合とは分布型が異なることも推測される。文構造の測り方にはこのほかにもいろいろな方法があり得るから、それらを一つ一つ実際のデータに適用し、文長の説明変数としての適否を検討していく必要があるだろう。

回帰モデルについては、このほかにも、文の長さがさまざまな要因によって決まるものだとすれば、文構造以外の諸要因も説明変数に組み込んだ精緻なモデルを構築するという、より大きな課題が考えられる。そして、そのためには、かつて樺島忠夫が行ったような（樺島 1953）、そもそも文長というものがどのような仕組みによって決まるのかについての言語学的な追究・知見[10]を前提にし、かつ、そうした追究をさらに深めていくことが求められるだろう。文の長さの統計モデルを定めるためには、なお検討の余地が多く残されている。

注

１）こうした分布を「混合分布」あるいは「分布の混合」ということがある。ある分布を混合分布と仮定してそれを複数の確率分布に分割（クラスタリング）する「混合分布モデル」という考え方もある（新納 2007 など参照）。

２）「文は回帰性という特性によって無限に長くすることができる」という考え方については、中島(2016)にわかりやすい解説がある。

３）安本(1957)、新井(2001)は、文長を文字数で測った場合、その分布は対数正規分布になるとし、佐々木(1976)は対数正規分布のほかにガンマ分布の場合もあるとしている。また、Ishida and Ishida(2007)は、文長を形態素数で測った場合、Hyper Pascal 分布になるとしている。

４）波多野(1950)の表（p.213）には「総合計 500（文）」とあるが、字数 20～24 の小計に誤りがあり、490（文）が正しい。ただし、本稿の調査では 480 文であり、これとも一致しない。

5）もちろん、原理的に上限がないといっても、現実の複文の文長には限界があり、度数分布は必ず収束する。複文の限界文長がどれほどかという問題は、また別の研究課題となるだろう。

6）図 3b で、従属節数１～３の分布が双峰性を示しているように見えるのは、従属節の種類によってその長さ（の分布型）が異なることを示唆するものかもしれない。伊藤雅光氏の教示による。

7）いま仮に、従属節を、南(1993)の「従属句」の分類に従って以下のように３分すれば、それぞれの長さ（文節数）は、その構成要素の範囲に応じて異なり、Ａ類＜Ｂ類＜Ｃ類の順に長くなるはずである。

Ａ類：～ナガラ〈非逆接－「カネガアリナガラ出ソウトシナイ」などの逆接の意味でないもの〉、～ツツなど。構成要素の範囲がもっともかぎられている。

Ｂ類：～カラ、～ト、～ナラ、～ノデ、～ノニ、～バなど。構成要素の範囲がＡ類より広くなるが、つぎのＣ類よりはせまい。つまり、ＡとＣの中間。

Ｃ類：～ガ、～カラ、～ケレド（モ）、～シなど。構成要素の範囲がもっとも広い。

したがって、１文の長さ（文節数）を目的変数 Y、１文におけるＡ類・Ｂ類・Ｃ類の数を説明変数 X_A・X_B・X_C とし、その偏回帰係数を β_A・β_B・β_C、定数項を α とすると、以下のような重回帰モデルを考えることができる。

$$Y = \alpha + \beta_A X_A + \beta_B X_B + \beta_C X_C$$

8）ただし、飯豊(1956)では、文の長さの統計モデルを定めるという問題設定はされていない。また、続く（二）以降の論考は発表されていないようである。

9）飯豊(1956)の表（p. 29）では、「金と銀」の「四次文」の総数が 54 文となっているが、これは、その前のページの表で示されているように、正しくは 52 文である。ただ、四次文のうちのどの文長の文数に誤りがあるか不明であるため、今回はこの数値をそのまま利用した。

10）樺島(1953)は、「文の長さに作用すると思われる条件」を、⑴推敲可能性、⑵場面、文脈への依存度、⑶年齢、⑷文章に働く制限、⑸内容、文体意識又

は表現価値、の５項目に分類し、いくつかの計量的な調査・実験により、次のことが確かめられたとする（p. 30）。

○文の長さは表現理解上の制約を比較的受けないものより受けるもの、即ち推敲可能性の高いものより低いもの、場面文脈よりの独立度が強いものより弱いもの、書きことば的なものより話しことば的なものが文長は短くなる。

○年齢が増加するに従って文長は長くなる。ただし、高校生以上においてその差は少ない。

○一定数の文をもつ文章に、その大きさが小さくなるような制限を、重要な語句を省略することなく加えた時、文章中の文の数は減少し、文の平均文長は長くなる。また構造は複雑になる。

○文の表現性は文の長さに関係する。

引用文献

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』125-3:205-223

飯豊毅一(1956)「文の長さと構造（一)」『解釈』02-06:26-31

樺島忠夫(1953)「文の長さについて―条件との相関の分析―」『国語学』15:21-31

佐々木和枝(1976)「文の長さの分布型」『計量国語学』78:13-22

新納浩幸(2007)『Ｒで学ぶクラスタ分析』オーム社

高安秀樹(2004)『経済物理学（エコノフィジックス）の発見』光文社

中島平三(2016)「世界一長い文は？」中島平三編『ことばのおもしろ事典』143-150、朝倉書店

橋本　修(2003)「日本語の複文」北原保雄監修・編集『朝倉日本語講座５　文法Ｉ』181-199、朝倉書店

波多野完治(1950)『現代文章心理学』新潮社

南不二男(1974)『現代日本語の構造』大修館書店

南不二男(1993)『現代日本語文法の輪郭』大修館書店

安本美典(1957)「文の長さの分布型」『計量国語学』1:27-31

安本美典(1960)『文章心理学の新領域』創元社
Ishida, M. and Ishida, K. (2007). On distributions of sentence lengths in Japanese writing. *Glottometrics*, 15, 28-44.

付記

本稿は、計量国語学会第 60 回大会（2016 年 10 月 8 日、日本大学）での口頭発表「文の長さの統計モデル」にもとづくものである。発表の際、有益な教示・助言をいただいた。記して感謝申し上げる。

（文学研究科教授）