文の長さ分布から見た文生成のメカニズム

古橋　翔（東北大学大学院理学研究科）†

Mechanism of Writing Sentences Based on the Distribution of Sentence Lengths

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、文を構成する文字や単語を数え、最頻値、平均値や分布型などから作者の文体を特徴づけてきた。

日本語における文の長さ（文長）に関する研究に、長さの単位を文字とし、文長分布が対数正規分布であると示した安本(1958)と新井(2001)、対数正規分布とガンマ分布であると報告した佐々木(1976)の先行研究がある。一方で、長さの単位を形態素とし、文長分布が Hyper Pascal 分布と報告した石井と石井(2007)の研究がある。

現象の仕組みを理解する上で、実験や観測により得られたデータの分布型を再現するモデルを考えることは、重要である。上記の先行研究においても文長分布の生成モデルが挙げられている。佐々木は、対数正規性を生む一例として Kaptyen のアナログマシンを例に Multiplicative な確率過程を挙げ、ガンマ分布に対しては、文の構成要素の長さが指数分布に従うから、文の構成要素が合わさった文長はガンマ分布に従うのではないかと考察している。石田らは、G. Altmann (1988)の Hyper Pascal 分布の生成モデルを挙げている。

本研究では、先行研究で挙げられた日本語の文長分布型から、日本語文の生成メカニズムを解明しようと試みた。研究に当たり、佐々木が提示したガンマ分布の生成モデルに注目する。佐々木は考察の中で、

「述べようとする『思想』あるいは『事柄』をこまかく分割したとき、その分布は指数分布に従う。そして、文の長さは、その指数分布に従うものが、いくつかくっついたものと考えることができる。従って、文の長さの分布は、『ガンマ分布』に従う。」

と述べている。また、くっつくべき個数 m を定数と考えることに若干の問題があるとし、m がある分布に従うとするならば、「複合分布」を考えなければならないと考察している。

本研究では、「述べようとする『思想』あるいは『事柄』」が文を構成する句や節に対応すると考え、佐々木の考察により文長の分布型が説明できるかどうか調べた。

２．文長の分布型

先行研究で取り上げられた分布型を説明する。

２.１対数正規分布

対数正規分布は、

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

で定義される連続型確率分布である。パラメータ μ は $\ln x$ の平均値で、σ^2 は $\ln x$ の分散で

† furuhashi@cmpt.phys.tohoku.ac.jp

ある。変数 x の対数 $\ln x$ が正規分布に従うので、$f_{\mathrm{LN}}(x)$ は長い裾をもつ分布である。

２.２ガンマ分布

ガンマ分布は、

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

で定義される連続型確率分布である。パラメータは、形状母数 k と尺度母数 θ の二つである。特徴として、長い裾が挙げられる。確率変数 $X_1, X_2, \cdots, X_\alpha$ が指数分布 $\beta^{-1}\exp(-x/\beta)$ に従うならば、$X_1 + X_2 + \cdots + X_\alpha$ はガンマ分布（$k = \alpha$ 、$\theta = \beta$）に従う。

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。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。依存構造木のルート「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。本研究では、このようなノードの集合を枝と呼ぶとする。

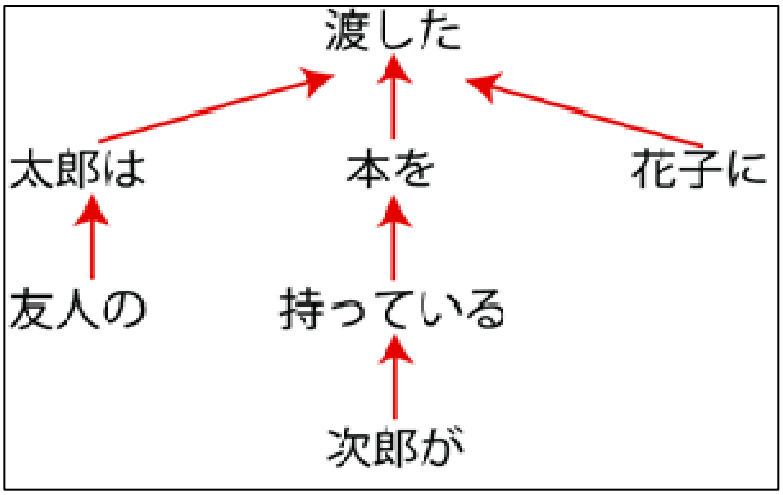

図 1 依存構造木。リンクの向きは、係り受けの向きを表す。

４調査方法とサンプル

本研究の調査方法と用いたサンプルを説明する。

４.１調査方法

本研究は、枝に着目して依存構造木を統計的に解析する。依存構造木の構成単位は文節なので、長さの単位は文節とする。まず始めに、サンプルの文長の分布型を調べる。分布型は、先行研究で挙げられていたガンマ分布と対数正規分布に限定し、赤池情報量規準（AIC）で判断する。次に、一文当たりの枝の数の分布をみる。枝の数が一定であればガンマ分布の生成モデルに当てはまるが、実際は佐々木が指摘したように一定ではないと思わ

れる。さらに、枝の長さ分布を調べ、指数分布となっているか確かめる。そして、一文中の枝の長さに相関があるかどうか調べ、各枝の独立性を確認する。

４.２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 と、現代日本語書き言葉均衡コーパス（BCCWJ）を用いた。

京都大学テキストコーパスは、Web ページからダウンロードできるパッケージを用いた。パッケージに含まれているのは、形態素・構文・関係の付加情報のみであり、テキストは含まれていない。京都コーパスを本来の形に変換するためには、毎日新聞 1995 年版 CD-ROM が必要である。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 は使用しなかった。総文数は 38397 である。

```
<?xml version="1.0" encoding="UTF-16"?>
<sample sampleID="OW1X_00000" version="20070814" type="variableLength">
<article articleID="OW1X_00000_V001" isWholeArticle="false">
<titleBlock>
<title><sentence type="quasi"> 第２節　内外均衡の背景 </sentence><br type="automatic_original"/></title>
</titleBlock>
<paragraph>
<sentence>　５３年度中にみられた内外均衡回復に向けての動きは、それぞれがバラバラに生じてきたわ
けではない。　</sentence><sentence> 以下では、それらの動きの重要な背景として、・・・
</paragraph>
```

図 2　BCCWJ の XML

BCCWJ は、C－XML フォルダ下にあるサブコーパス LB（図書館サブコーパス「書籍」）、OB（特定目的サブコーパス「ベストセラー」）、PB（出版サブコーパス「書籍」）と PN（出版サブコーパス「新聞」）に収蔵されている XML ファイルを用いた。

XML ファイルから文を取り出す方法を示す。まず、XML ファイル中の記事（article タグで挟まれた部分）を、BCCWJ に添付されている記事情報データに基づき実著者の人名 ID ごとに分ける。記事の中には別の ID を持つ記事が含まれているものがあるが、このような記事は除外する。次に、実著者ごとに振り分けた記事から、文として sentence タグで挟まれた部分を、その sentence タグの階層情報とともに取り出す。sentence タグの階層情報とは、article タグから sentence タグに至るまでに通るタグの集合である。例えば、図 2 では、「第二節　内外均衡の背景」のタグの階層情報は “article, titleBlock, title, sentence”、「５３年度中にみられた内外均衡回復に向けての動きは、それぞれがバラバラに生じてきたわけではない。」のタグの階層情報は “article, paragraph, sentence” となる。一部の文は中に別の sentence タグを含んでいる。その場合は、一番外側の sentence タグに従う。また、speech タグが閉じた直後には「と、言った。」が多いが、本研究ではこれを文とは認めない。よって、speech タグ直後の sentence タグは除外した。

更に、取り出した文を以下の条件に従い選別する。

(ア) 階層情報が article, paragraph と sentence タグのみで構成されている。
(イ) sentence タグで挟まれた部分に、ruby と sampling 以外のタグが含まれていない。
(ウ) sentence タグの種類が quasi ではない。
(エ)　▼◇〒@※◆■…＝などの記号が含まれない。
(オ) ”)”、”＜” が含まれない。

(ア) は、会話文、箇条書きやキャプション等を除いた地の文を対象とするため、(イ) は選出基準を簡潔にするため、(ウ) は、文に準ずるものではなく、きちんとした文を対象とす

るため、(エ)と(オ)は、係り受け解析の精度を高めるためである。

係り受け解析には、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．結果

京都大学テキストコーパスから得られた結果を示す。

５.１文長分布

文長分布を図３に示す。文長の平均は 9.7 で標準偏差は 5.3 である。最尤法により推定したパラメータ値は $k = 3.45$，$\theta = 2.81$，$\mu = 2.12$，$\sigma^2 = 0.34$ である。また、AIC は、対数正規分布の場合 229744、ガンマ分布の場合 227865 であり、ほぼ同じ値であり、どちらか一方と断言できない。

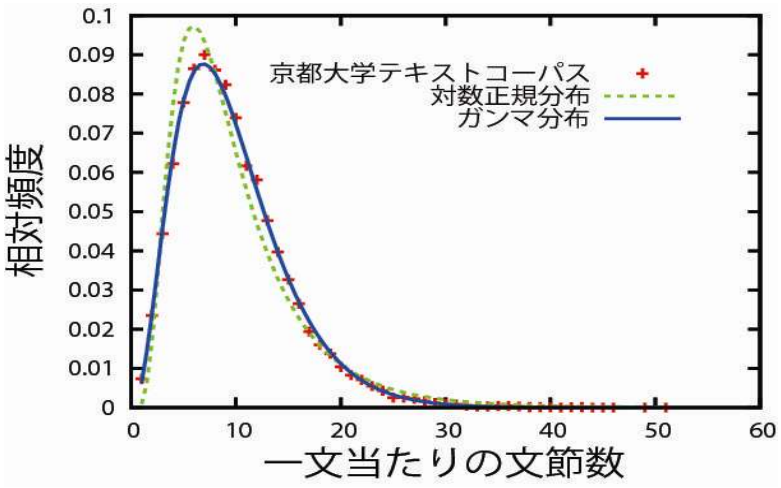

図 3　文長分布

５.２枝の数分布

図４は枝の数分布である。

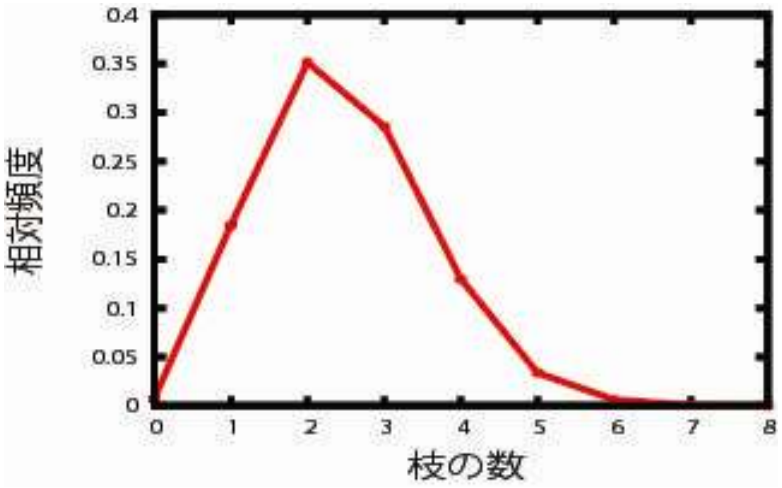

図 4　枝の数分布

平均 2.48、分散 1.22、最頻値は 2 である。ガンマ分布の生成モデルのような定数に枝の数はならないが、狭い範囲に集中している。

５.３枝の長さ分布

枝の長さの累積分布を図５に示す。図４より、依存構造木の枝の数は一定ではないので、枝の数ごとに分布をとった。

分布をみると、枝の数が２の場合全体的に指数分布に従うが、それ以外は枝が短い領域で指数分布からずれている。枝の数が１の分布は、枝の数が２の分布より上、枝の数が３以上の分布は、枝の数が２の分布より下にあり、累積分布では長さ１で相対頻度の累積が必ず１になる点を考慮すると、枝の数が多くなるにつれて、短い枝の割合が大きくなることがわかる。このような変形はあるものの、指数分布のような分布型がみられる。

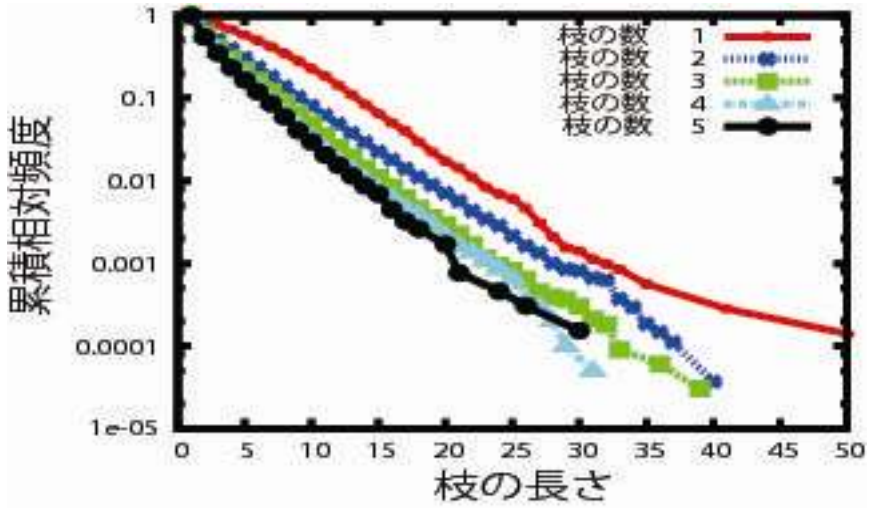

図 5　依存構造木を構成する枝の長さの累積分布。
依存構造木に含まれる枝の数ごとにプロットした。

５.４枝の長さ相関

ガンマ分布の生成モデルでは、確率変数は互いに独立である。文を構成する枝の長さが独立（無相関）かどうかピアソンの相関係数により調べた。

ピアソンの相関係数は、次のようにして求める。まず各文の枝をラベル付けする。具体的には、枝に含まれるルートの子ノードに対して文頭からの出現順位付けを行った。例えば図１の依存構造木の場合、ルート「渡した」の子ノード「太郎は」、「本を」、「花子に」はこの順番に文中に現れるので、一番目の枝は「友人の太郎は」、二番目の枝は「次郎が持っている本を」、三番目の枝は「花子に」である。文 k の枝の数を b_k 、 j 番目の枝の長さを s_{kj} とし、枝の数 x の文に対する i 番目と j 番目の枝の長さの相関を表すピアソンの相関係数 $r_{ij}(x)$ は、

$$r_{ij}(x)=\frac{\sum_k (s_{ki}-\overline{s_i(x)})(s_{kj}-\overline{s_j(x)})\delta_{b_k,x}}{\sqrt{\left(\sum_k (s_{ki}-\overline{s_i(x)})^2\delta_{b_k,x}\right)\left(\sum_k (s_{kj}-\overline{s_j(x)})^2\delta_{b_k,x}\right)}},$$

となる。但し、

$$\overline{s_i(x)} = \frac{\sum_k s_{ki}\delta_{b_k,x}}{\sum_k \delta_{b_k,x}},$$

である。相関係数は枝の数で区別して計算した。理由は枝の長さが枝の数に依存する可能性を排除するためである。表 1 にピアソンの相関係数を示す。相関係数の絶対値は 0.1 程度なので、長さに相関は見られない

表 1　ピアソンの相関係数。左から枝の数が 2（文の数 13487)，3（文の数 10972)、4（文の数 4982）である。

	i=1	i=2
j=1	1.0	-0.15
j=2		1.0

	i=1	i=2	i=3
j=1	1.0	-0.12	-0.099
j=2		1.0	-0.049
j=3			1.0

	i=1	i=2	i=3	i=4
j=1	1.0	-0.11	-0.10	-0.074
j=2		1.0	-0.046	-0.044
j=3			1.0	0.11
j=4				1.0

６．作者別に見た場合

京都大学テキストコーパスで得られた結果が他のコーパスで得られるか確かめるために、BCCWJ を用いて同様の解析をした。PN から、毎日新聞社(人名 ID: 258099、総文数: 1323)、朝日新聞社(人名 ID: 256908、総文数: 1798)、読売新聞東京本社(人名 ID: 259670、総文数: 1871)と中日新聞社(人名 ID: 263664、総文数: 1234)、LB、OB と PB から、赤川　次郎(人名 ID: 873、総文数: 3410)、 森村　誠一(人名 ID: 47302、総文数: 3150)、西村　京太郎(人名 ID: 55252、総文数: 1468)と司馬　遼太郎(人名 ID: 70104、総文数: 4532)を対象とした。
以下に、文長分布、枝の数分布、枝の長さの累積分布を示す。枝のサイズ相関はまだ調べていないので結果を示せない。

６．１文長分布

図６に、新聞社と作家それぞれの文長分布を示す。ともに、京都大学テキストコーパス（図３）同様に裾が伸びた分布をしている。これらの分布が、対数正規分布とガンマ分布どちらに当てはまるかを AIC により評価する。表２は、最尤法で推定した対数正規分布とガンマ分布のパラメータ値であり、表３は AIC の値である。

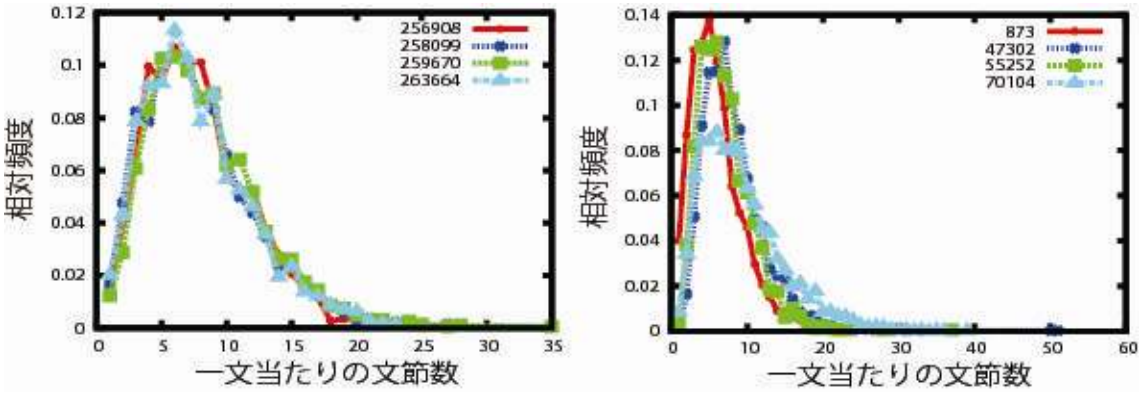

図 6　文長分布。左のグラフが新聞社。右のグラフが作家。

表 2　最尤法によるパラメータ値

	毎日新聞社	朝日新聞社	読売新聞東京本社	中日新聞社	赤川次郎	森村誠一	西村京太郎	内田康夫
k	3.29	3.62	3.60	3.29	3.02	4.41	4.30	4.90
θ	2.38	2.17	2.32	2.36	1.99	1.85	1.65	1.72
μ	1.90	1.92	1.98	1.89	1.62	1.98	1.84	2.02
σ^2	0.361	0.327	0.324	0.364	0.386	0.241	0.253	0.220

表 3　AIC の値

	毎日新聞社	朝日新聞社	読売新聞東京本社	中日新聞社	赤川次郎	森村誠一	西村京太郎	内田康夫
ガンマ分布	7342	9844	10484	6828	17334	17000	7544	7146
対数正規分布	7438	9985	10602	7438	17483	16956	7556	7158

最尤法で推定したパラメータ値は、京都大学テキストコーパスの値と大きな違いは見られない。また、AIC の値をみると、ガンマ分布と対数正規分布間に大きな違いは見られない。したがって、本研究で用いる新聞社と作家の文長分布は、京都大学テキストコーパスの文長分布と似た特徴を有している。

６．２枝の数分布

図７は枝の数分布である。図４の京都大学テキストコーパスと比較すると、全体的に同じ分布型をしている。

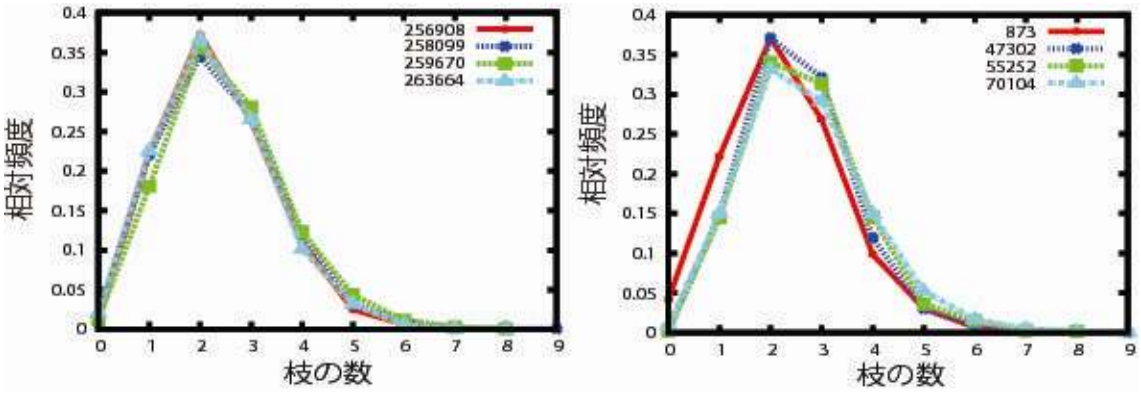

図 7　枝の数分布。左のグラフが新聞社。右のグラフが作家。

６．３ 枝の長さ分布

新聞社から読売新聞東京本社（人名 ID259670）、作家から森村誠一（人名 ID 47302）を選んで、枝の長さ分布をとった結果を図８に示す。京都大学テキストコーパス（図５）と同じ特徴を持った分布型をしている。

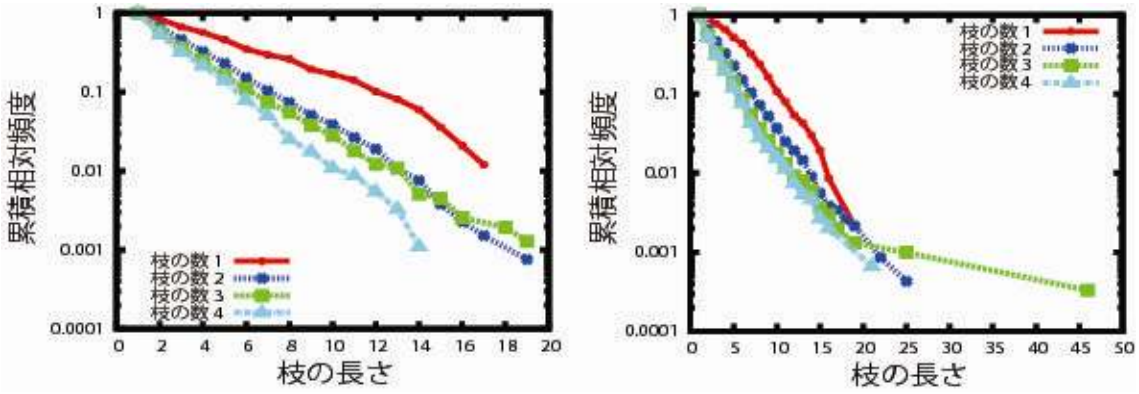

図 8　文を構成する枝の長さ分布。各文の依存構造木に含まれる枝の数ごとにプロット。左のグラフが読売新聞東京本社（人名 ID259670）、右のグラフが森村誠一（人名 ID47302）。

以上から、小説家や他の新聞社でも、京都大学テキストコーパスで得られた結果と非常に類似している。よって、文生成に対する佐々木の考察は一般的に当てはまると推測できる。

7．まとめ

本研究では、係り受け関係と文長に注目して、日本語文の生成メカニズムを調べた。依存構造木のルートに該当する文節に係る句や節の長さが、互いに独立で指数分布に従うことが分かった。ただし、句や節の数は一定ではないので、単純なガンマ分布ではなく複合分布となる。この結果は、佐々木の考察と一致するものであった。

今後の方針として、日本語以外で依存文法に従い同様の解析を行い、文生成における共通点と相違点を明らかにすることが挙げられる。依存文法による解析情報をもつコーパスは多くの言語で公開されている。しかしながら、依存文法に基づいた文構造がツリーにならない、サンプルの量が少ないなど問題がある。このような問題をいかに克服するかが、まず取り組むべき課題である。

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, 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.

Gabriel Altmann (1988) Verteilungen der Satzlängen. *Glottometrika*, 9, pp. 147-170.

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

国立国語研究所の言語コーパス整備計画 KOTONOHA http://www.ninjal.ac.jp/kotonoha/