数理解析研究所講究録
第1658巻 2009年 205-218

チョムスキー文章の分布則とこれに基づくゆらぎ解析

福岡教育大学教育学研究科数学教育専攻　麻生庸介
松尾美幸
福岡教育大学教育学部数学教育講座　濃野聖晴
日本大学文理学部情報システム解析学科　岩澤秀樹
鈴木　理

Abstract

長さを指定して得られるチョムスキー文章をセンテンスに分解し、その長さに関する頻度数の分布則を与える。この分布則を基礎として、クラスター数が1となる文章の分布から$1/f$-ゆらぎ、折れ棒モデルがどのように近似モデルとして得られるかを示す。この分布は指数分布に近く全体としては$1/f$-ゆらぎは記述できないことが分かる。相転移にかかわると思われる「小山現象」はクラスター数の高い文章により表現可能なことが示される。

1. はじめに

本研究の目的はチョムスキーが考察した文脈自由文法の基礎となった整合括弧列の分布を用いてゆらぎ解析を行うことである。$1/f$-ゆらぎは様々な所に現れているにもかかわらず、その数理理論は僅かに知られているにすぎない([17])。また、$1/f$-ゆらぎには十分大きなfについて$1/f$-ゆらぎを破る新しい現象(これを「小山現象」という)も観察されている。小山現象は$1/f$-ゆらぎを生じる現象の相転移現象と考えることが出来るにもかかわらず殆ど注目されていないように思われる。ここではこの現象も視野に入れて研究を行う。チョムスキーは、彼の生成文法理論を構成するにあたり、「言語機能は先天的にヒトに与えられており、進化的立場で考察されるべきである。」と主張している([3])。本研究においては$1/f$-ゆらぎは考察される対象の生成過程或いは進化過程と結びついていると予想し、チョムスキー文章の分布のような単純な分布を考察することで、このゆらぎのみならず小山現象も進化過程として捉えられるのではないかと期待してこの研究をおこなった。

2. $1/f$-ゆらぎ

統計学で取り扱われる現象は正規分布をはじめポアソン分布等の様々な分布を用いて記述される。近年、複雑系にこれらの分布とは異質なべき則ゆらぎと呼ばれる分布が見出されている。その特徴はその分布が自然科学のみならず人文・社会科学分野で見出されており、その分野を特定しないところにある。その例を幾つか述べる。$1/f$-ゆらぎのグラフは、両変数の対数をとって対数座標で表すと、傾きが-1の直線になる。第一の例は$1/f$-ゆらぎの最も典型的なものとして知られる Zipf の法則である(図1)。これは英文で書かれた一冊の書物に現れる単語の頻度数を適当に規格化して表現したものである。次の例は雪崩の大きさに関する頻度数を表す(図 2)。これは直線というより上に凸となるように見える。この形は折れ棒モデルに見られる分布であり、これに基づくモデルづくりが行われている。同様な現象がDNAの塩基配列にも見られる(図8)([15])。コード化されていない所謂ジャンクと呼ばれる偽遺伝子は$1/f$-ゆらぎに近く、コード化されている配列については折れ棒モデル型の分布が見られるのは興味深い。

音楽にみられる$1/f$-ゆらぎは報道番組等でとりあげられよく知られている(図５)。その他、銀河の分布、人口分布にも見られる(図 4,6,7) ([4],[5])。

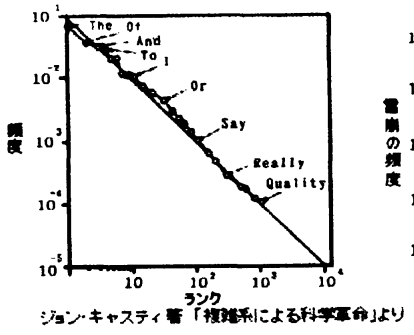


図１単語とランク（Zipf の法則）

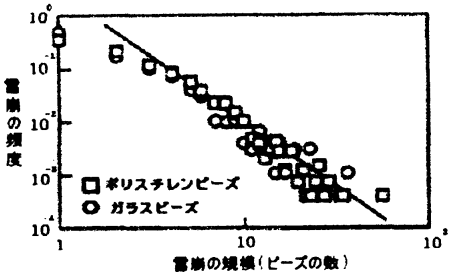


図２雪崩の大きさと頻度の関係

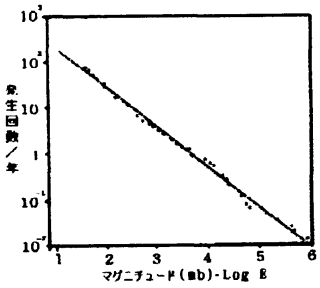


図３地震のマグニチュードと頻度
(グーテンベルグ・リヒター則)

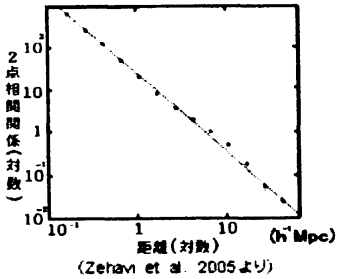
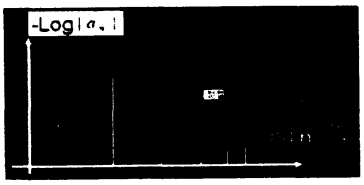


図４銀河における２点相関関係



$f = \sum_{n=1}^{\infty} a_n f_n$

図５音楽に見られる$1/f$-ゆらぎ

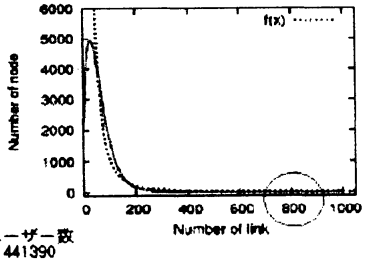


図６mixi ネットワークの冪則ゆらぎ

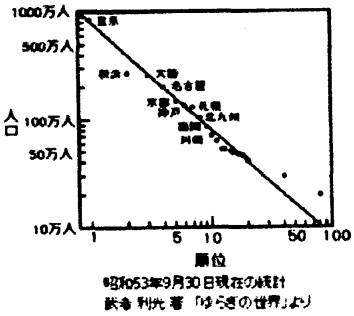


図７日本の都市人口と順位

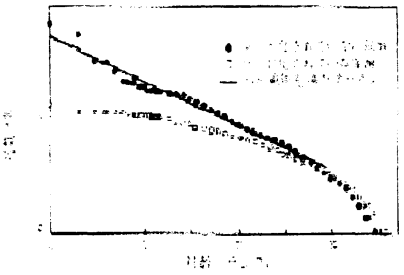


図８イースト菌の DNA

３．小山現象

ある現象に着目しその頻度数の分布を調べると一般にはベキ則等の右下がりの形状を示すのが通例であるが、単調に右下がりにならないことがしばしば見受けられる。例えば、地震については、規模の小さい地震は、図 3 のようなグーテンベルグ・リヒター則として知られるべき則ゆらぎとなるが、規模の大きい地震はこれには従わず、図 9 のように小山(小谷)のような部分が見出される。これを小山(小谷)現象と呼ぶ。また、分子進化学においても同様な現象が存在する。図 10 は DNA の突然変異が生じる範囲の塩基列の長さに関する頻度数分布である([2])。この図を見ると、全般的には右下がりになるが、280 前後のところでは小山現象が現れている。この小山現象は ALU と呼ばれる、所謂動く遺伝子によってもたらされた変異であり、ヒトやチンパンジーにおいて顕著である。この遺伝子は知能に関係するのではないかと考えられている。さらに、図 11 のように、銀河の分布についても同様の小山現象が現れている。これは宇宙誕生 38 万年後のいわゆる宇宙の晴れ上がりと関係するのではないかと考えられている([4])。その他、小山現象は、単語の長さに関する頻度数、ソー

シャルネットワークシステム(SNS)等にも見られる(図 6,12,14)。これより小山現象は$1/f$ゆらぎの表す現象に何らかの異変が生じそこで新しい現象に移行している、所謂「相転移」が起きていると考えられる。

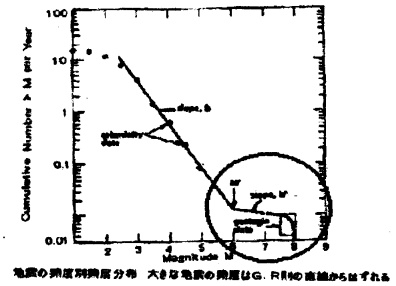


図 9　地震のマグニチュードと頻度

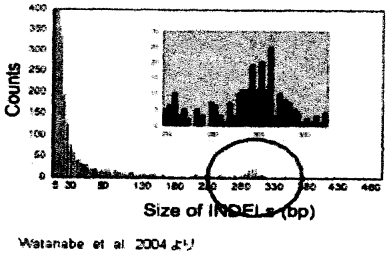


図 10　DNA(人間とチンパンジー)

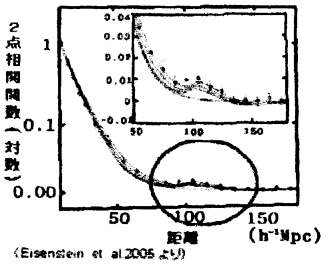


図 11　銀河の 2 点相関関係

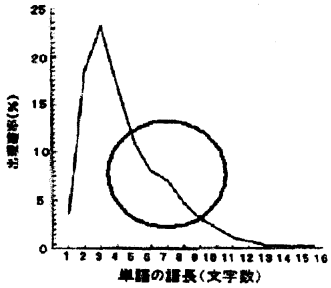


図 12　単語の語長と頻度

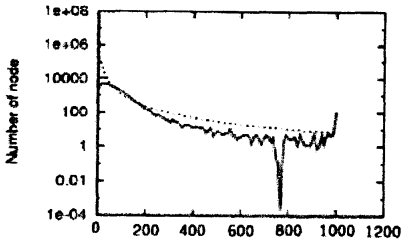


図 13　mixi ネットワーク

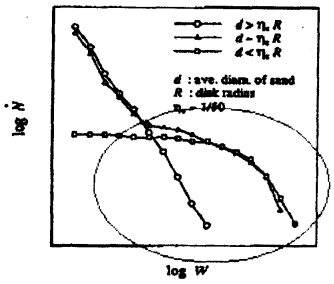


図 14　雪崩の大きさと頻度

4．チョムスキー文章

チョムスキーは文章の背景には図 15-1 のような木構造があることに着目してチョムスキー生成文法理論を構成した([3])。その最も基本的な部分は、この木構造に対して，図 15-2 のように括弧列を考え文脈自由文法(整合括弧列文法)を定義するところにある。英文で書かれた文章の場合にその木構造の一例を示す(図 15-1)。この木構造に括弧の列を対応させることが出来る(図 15-2)。文脈自由文法は書かれた単語をすべて□に置き換え内容を問わずにその括弧の列のみに着目することにより定められる文法である。本論文において考察する文章をチョムスキー文といい、形式言語理論でいう整合括弧列と呼ばないのは、その生成過程を進化論の立場で考え、その文脈依存性にまで入り込むからである(§ 9)。以下文章といえば断らないかぎりチョムスキー文章のこととする。

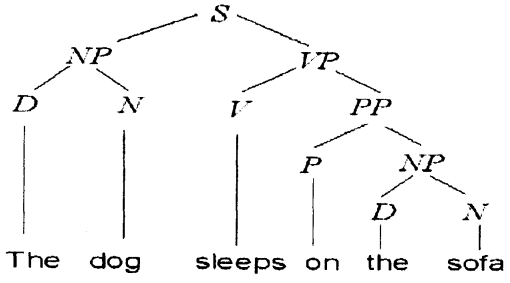


図 15-1　文章の木構造の例

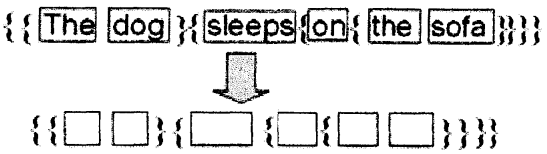


図 15-2　括弧列の例

どのような括弧列が受理言語となるかどうかの判定法は、次の条件により与えられる：

受理条件

与えられた括弧列に対し、左から順に括弧の数を数え、すべての時点において

（開き括弧の総数）≧（閉じ括弧の総数）

となり、最後に（開き括弧の総数）＝（閉じ括弧の総数）となるときに限り受理される。

次に，チョムスキー文章を考察するにあたって必要となる基本的な用語を幾つか述べる：

定義 1 文章Sに対して、開き括弧と閉じ括弧の対の個数をSの長さといい、length Sと表す。

定義 2 長さがnである文章Sに対して、上記の受理条件が成り立ち、等号が成り立つのはnのときに限るとき、Sは既約な文章であるといい、そうでないときSは可約な文章であるという。長さがnとなる既約文章の個数がmとなる文章の全体を$C(n,m)$と書く。

例 {{}},{{{}}{}}(長さ2,4)は既約であるが、{}{{}},{}{}{}(ともに長さ3)は既約ではない。全ての文章を既約な文章の列として表わすことができる。これを文章の既約分解といい次のように表す：

$$\xi \in C(n,m) \quad \Rightarrow \quad \xi = \xi_1 * \xi_2 * \ldots * \xi_m (\xi_i (i = 1,2,..,m) \text{は既約な文章})$$

定義 3 既約な文章Sに対して、開き括弧の直後に閉じ括弧が現れるとき、この間に丸印を描く(図 16-1、図 17-1)。この丸印の総数をSのクラスター数といい、clus Sと表す。可約な文章S'においてはS'を$\{S_1,\cdots,S_m\}$と既約分解することによりクラスター数を次のように定める:

$$\text{clus } S' = \max\{\text{clus } S_1,\cdots,\text{clus } S_m\}$$

長さがn、クラスター数がαとなる文章の全体を$C^{(\alpha)}(n)$とかく。$C(n,m)$に属する文章のうち、クラスター数がαとなる文章の全体を$C^{(\alpha)}(n,m)$と書く。

単純な文章は図 16-2 のように、クラスター数が 1 の文章となる。それに対し、図 17-2 のようにクラスター数の大きい文章は、一般に幾つかの短文が接続詞等で結合された、より高度な文章を記述していると考えられる。従ってクラスター数の大小により文章の複雑さが分類されるといえる(§ 9)。

単純文章(I-型文章)

clus{{{}}} = 1 の文章をI-型文章という。

例

{○},{{○}},{{{○}}},.....

図 16-1

{{{}}}

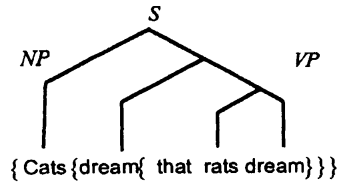


図 16-2

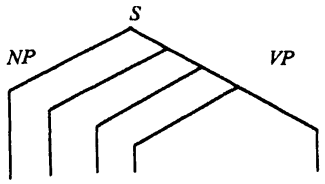


図 16-3 Ⅰ-型文章の木構造

II-型文章

clus{{○}{{○}}} = 2 となる文章。

{{○}{○}},{{{○}}}{{○}}},.....

図 17-1

{{} {}}

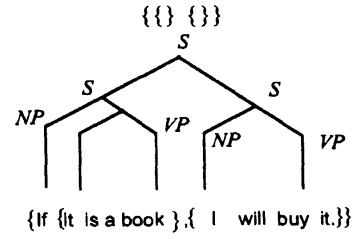


図 17-2

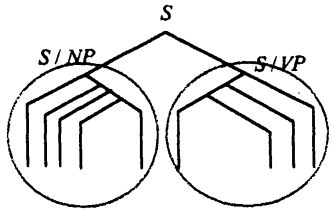


図 17-3 Ⅱ-型文章の木構造

5. チョムスキー文章の生成

整合括弧列文は、次の生成規則によって作られる([1])。次の演算は整合括弧列の生成を帰

納的に記述している。

(1)記号列{}は整合した括弧列文である。
(2)記号列 A が整合した括弧列文ならば、$\{A\}$も整合した括弧列文である。
(3)記号列 A とB が整合した括弧列文ならば、AB も整合した括弧列文である。

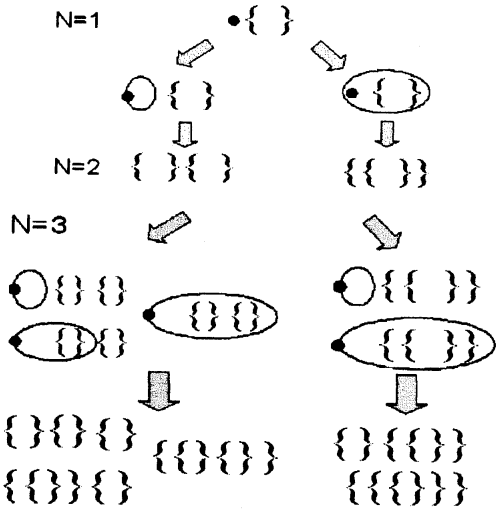


図 18　チョムスキー文章の生成ダイアグラム

次に文章の生成規則を与える。この方法の導入により議論が明快になっていることに注意する：

(基本構成)

長さがnである既約な文章または既約な文$\{S_1,\cdots,S_m\}$に分解されている可約な文章Sを考える。このとき、文章Sから長さが$n+1$となる文章を次のように生成することができる:

1. Sの前に点Pをうつ。
2. 点Pを始点としてSを囲まない円を描き、この円を括弧の対に置き換える。
3. すべての$a(1\le a\le m)$に対して点 P を始点として $S_1,\cdots,S_a$ のみを囲む円を描き、その円を括弧の対に置き換える(図 18)。

定理Ⅰ　基本構成によりすべての文章を生成することができる。

この証明は自明ではない。すべての括弧列の総数はカタラン数に一致していることはよく知られている([13])。定理は基本構成によりこれらがすべて得られることを主張している。証明は§ 8で与える。

次に、クラスター数が 1 となる文章について考える。長さがnとなる文章Sが与えられたとする。Sの既約分解を$\{S_1,\cdots,S_m\}$とする。$\mathrm{length}\,S_a$を$k_a(1\le a\le m)$として、この文章を$\xi(k_1,k_2,...,k_m)$と表すことにする。基本構成の定める演算、σと書く、は次の様になる:

$$\begin{array}{ccc}\sigma: & C^{(1)}(n) & \to & C^{(1)}(n+1)\\ & \cup & & \cup\end{array}$$

$$\xi(k_1,k_2,...,k_m)\mapsto\xi(1,k_1,k_2,...,k_m)+\xi(k_1+1,k_2,...,k_m)$$

ここで二つの文章A,Bに対してその直和を$A+B$と表わした。従って、長さがnでクラスター数が 1 のすべての文章は次のように表されることが分かる。

$$\sigma^{n-1}(\xi(1))=\sum_{m=1}^{n}\sum_{\substack{k_1+\cdots+k_m=n\\ k_j\ge 1\,(1\le j\le m)}}\xi(k_1,k_2,\cdots,k_j,\cdots,k_m)$$

この文章の総数をP_nとすると、$P_n=2^{n-1}$となることに注意する。

この生成規則から次の分布則が得られる：

定理Ⅱ　$C^{(1)}(n)$のすべての文章を既約な文章に分解したとき、長さがkとなる既約な文章の個数を$P_k(n)$とすると、次のことが成り立つ:

(1) $P_k(n)=\begin{cases}(n-k+3)2^{n-k-2} & (n\neq k)\\ 1 & (n=k)\end{cases}$.

(2) $P(n)=\sum_{k=1}^{n}P_k(n)$ とおくと、 $P(n)=(n+1)2^{n-2}$ が成り立つ。

(3) $\mu_k(n)=P_k(n)/P(n)$ とおくとき、

$$\mu_k(n)=\begin{cases}\dfrac{(n-k+3)}{(n+1)}2^{-k} & (n\neq k)\\ \dfrac{1}{(n+1)2^{n-2}} & (n=k)\end{cases}$$

となる。この分布をチョムスキー文章の長さ分布と呼ぶことにする。

[証明]

(1) 定義よりすべてのnに対して、 $P_n(n)=1$ となることがわかる。そこで、

$$\sigma^{n-1}(\xi(1))=\sum_{m=1}^{n}\sum_{\substack{k_1+\cdots+k_m=n\\ k_j\geq 1\,(1\leq j\leq m)}}\xi(k_1,k_2,\cdots,k_j,\cdots,k_m)$$

さらにσを作用すると、

$$\sigma(\sigma^{n-1}(\xi(1)))=\sum_{m=1}^{n}\sum_{\substack{k_1+\cdots+k_m=n\\ k_j\geq 1\,(1\leq j\leq m)}}\xi(1,k_1,k_2,\cdots,k_j,\cdots,k_m)+\sum_{m=1}^{n}\sum_{\substack{k_1+\cdots+k_m=n\\ k_j\geq 1\,(1\leq j\leq m)}}\xi(k_1+1,k_2,\cdots,k_j,\cdots,k_m)$$

これより $P_k(n+1)=P_k(n)+P_k(n-1)+\cdots+P_k(k)+2^{n-k}$ となることが示される。

また、 $P_k(n)=P_k(n-1)+P_k(n-2)+\cdots+P_k(k)+2^{n-1-k}$ より、

$$P_k(n+1)=2P_k(n)+2^{n-1-k}$$

$$\frac{P_k(n)}{2^{n-2-k}}=n-k-1+2P_k(k+1)$$

ここで右辺第二項に寄与する文章は$\xi(1,k)$と$\xi(1,k)$のみであるから$P_k(k+1)=2$となり

$$P_k(n)=(n-k+3)2^{n-k-2}$$

(2)
$$\begin{aligned}P(n)=\sum_{k=1}^{n}P_k(n)&=\sum_{k=1}^{n-1}(n-k+3)2^{n-k-2}+1\\&=(n+3)2^{n-2}\sum_{k=1}^{n-1}2^{-k}-2^{n-2}\sum_{k=1}^{n-1}k\cdot 2^{-k}+1\\&=(n+3)2^{n-2}(1-2^{-n+1})-2^{n-2}\left(2-(n+1)2^{-n+1}\right)+1\\&=(n+1)2^{n-2}\end{aligned}$$

(3) (1)(2)より従う。

6.クラスター数が1のチョムスキー文章の分布によるゆらぎ解析

ここではクラスター数が1の文章の作る分布と従来考えられている折れ棒モデル、および$1/f$-ゆらぎの分布とを比較する。

(1) 折れ棒モデルとの比較

生態生物学者マッカーサーは、一定の領域に幾つかの種類の生物が共存している生態系のモデルとして折れ棒モデルを提案している([9],[10])。ここでは折れ棒というとき、棒の長さは自然数であり、折り目はすべて自然数座標に限るものとする、すなわち自然数の分割を考える。長さnの棒をm個$(0\leq m\leq n-1)$の折り目を入れて$m+1$個の部分に分割し、それぞ

れの分割片の長さを $p_k(k=1,2,..,m+1)$ とするとき $p_a \le p_{a+1}(a=1,2,\cdots,m)$ を満たすものとする。この分割を次のように表示することにする：

$$n = p_1 * p_2 * \ldots * p_{m+1}$$

長さ 4 の折れ棒についてそのすべての分割を図 19-1 に示す。数字を用いて表すと図 19-2 のようになる。長さ n の棒を分割したとき p_k の個数を $Q_k(n)$ と書くことにする。また、長さ n の折れ棒の種類の総数を Q_n とする(図 19-3)。

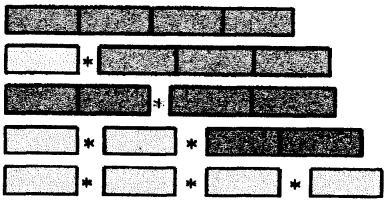


図 19-1

4
1 * 3
2 * 2
1 * 1 * 2
1 * 1 * 1 * 1

図 19-2

$Q_4(4) = 1$
$Q_4(3) = 1$
$Q_4(2) = 3$
$Q_4(1) = 7$
$Q_4 = 5$

図 19-3

P_n と Q_n の比較

次にクラスター数が 1 の文章の分布と折れ棒モデルの分布を比較する(図 20-1,2)。十分大きな n に関する P_n と Q_n の漸近挙動を比較する(P_n については § 5 を参照せよ):

$$P_n = \exp((n-1)\log 2),\ Q_n \approx \frac{1}{4n\sqrt{3}}\exp(\pi\sqrt{\frac{2n}{3}})$$

Q_n の漸近挙動は Hardy-Ramanujan の公式として知られている([12])。長さ n の折れ棒モデルについて、$p_a \le p_{a+1}$ の条件を除くと長さ n の文章となることが分かる。これをチョムスキー文を対称化すると云うことにする。

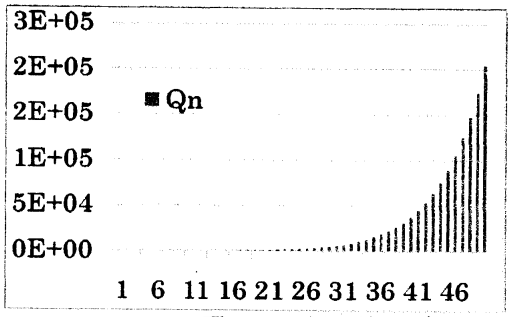


図 20-1　折れ棒モデル

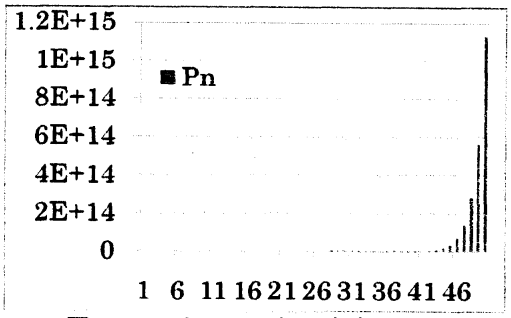


図 20-2　チョムスキー文章
（クラスター 1 ）

(2)　$P_k(n)$，$Q_k(n)$ と $1/f$ -ゆらぎの比較

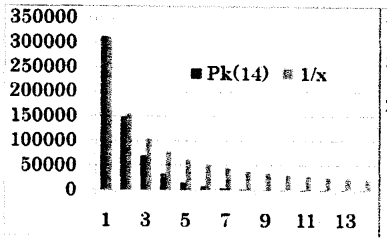


図 21-1　$P_k(14)$ と $1/f$ の比較

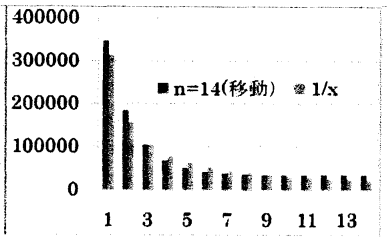


図 21-2　$P_k(14)$ の平行移動と $1/f$ の比較

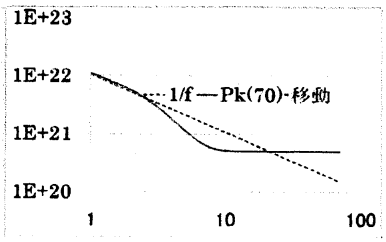


図 21-3　$P_k(70)$ の平行移動と $1/f$ の比較(対数座標)

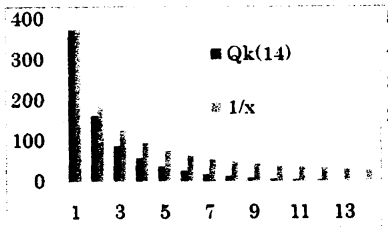


図 22-1 $Q_k(14)$ と $1/f$ の比較

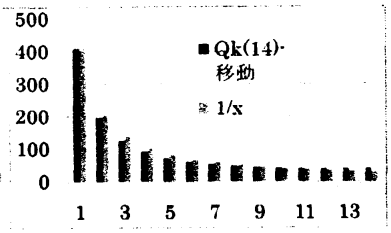


図 22-2 $Q_k(14)$ の平行移動と $1/f$ の比較

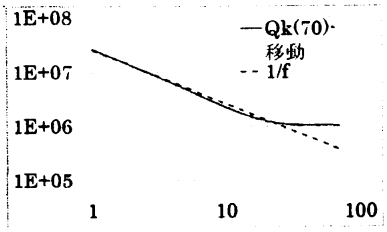
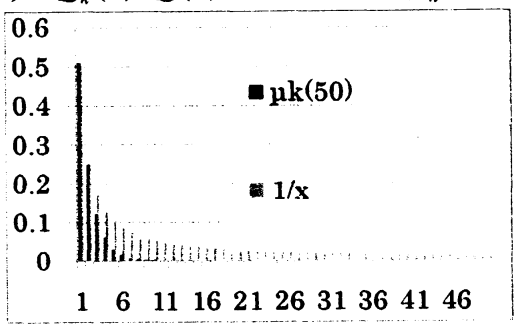


図 22-3 $Q_k(70)$ の平行移動と $1/f$ の比較(対数座標)

(3) $\mu_k(n)$, $\rho_k(n)$ と $1/f$-ゆらぎの比較

長さ n の折れ棒をとりその長さが k となる折れ棒片の個数を $Q_k(n)$ として、この分布を $\rho_k(n) = Q_k(n)/Q(n)$ とおく。以下 $\mu_k(n)$, $\rho_k(n)$ および $1/f$ ゆらぎの分布を比較する:



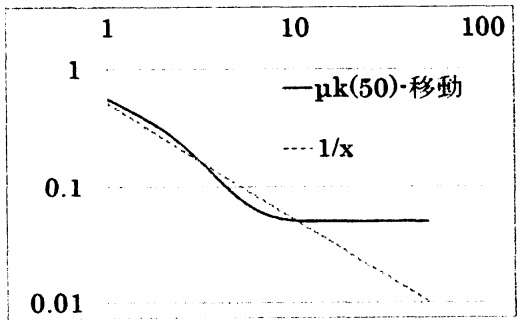
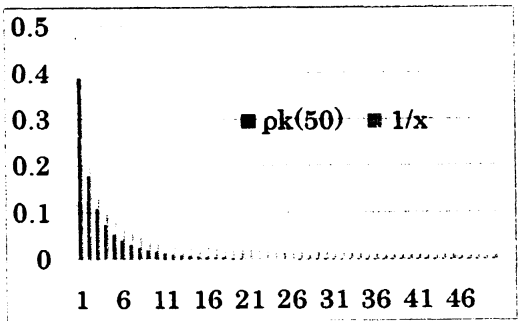


図 23-1 $\mu_k(50)$ の長さ分布 (右：対数座標表示)



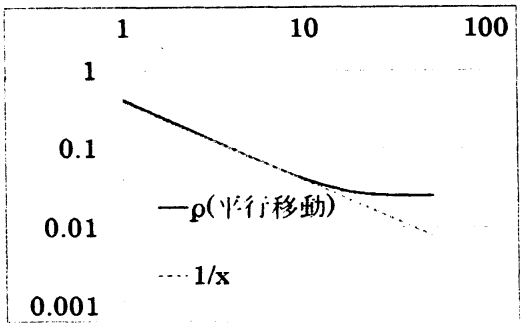


図 23-2 $\rho_k(50)$ の分布(右：対数座標表示)

以上の考察から次の事柄がわかる：

(1)長さ n が十分小さいとき、クラスター数が 1 の文章と折れ棒モデルはともに $1/f$-ゆらぎをよく近似するが、折れ棒モデルはより広い範囲でよく近似している。

(2)十分大きな n について、クラスター数が 1 の文章および折れ棒モデルの分布はともに指数ベキの振る舞いとなる。

(3)十分大きな n について、チョムスキー文章は $1/f$-ゆらぎよりも早く減衰し、$1/f$-ゆらぎを記述しない。

7．クラスター数の大きいチョムスキー文章の分布によるゆらぎ解析(小山現象)

ここではまずクラスター数の大きい文章がクラスター数 1 となる文章からどのようにして得られるかを述べる。次にこれに基づき既約成分とクラスター数を与えて文章の集合を考え、この分布が小山現象を記述する可能性のあることを示す。

文章をクラスター数が１となる既約文章の積を用いて表す。例を述べる:

$$\{\{\}\{\{\}\}\}=\{\xi_1*\xi_2\} \quad \text{ただし、} \xi_1=\{\},\xi_2=\{\{\}\}$$

すべての文章はひとまずクラスター数が１となる既約な文章の積を考え、次にこれらの幾つかを整合的な括弧でまとめることにより表すことができる。例えば文章$\xi(k_1,k_2,...,k_m)$の既約成分の第１成分と第２成分、第k成分と第$k+1$成分をまとめたとき、これを

$$\xi((k_1,k_2),...,((k_k,k_{k+1})),...\xi_m)=\{\xi_1*\xi_2\}*...*\{\{\xi_k*\xi_{k+1}\}\}*....\xi_m$$

と表すことにする。次の変換をクラスター変換という：

$$\begin{array}{ccc} Q:C^{(\alpha)}(n,m) & \rightarrow & C^{(\alpha')}(n+1,m-k_1-k_2+1) \\ \cup & & \cup \\ \xi(k_1,k_2,...,k_m) & \rightarrow & \xi((k_1,k_2),...,k_l) \end{array}$$

ここでは第一成分と第二成分のまとめ上げを例示してある。一般に$\alpha\le\alpha'$であり$\alpha'=\alpha+1$とは限らないことに注意する。次の事柄が成り立つ：

クラスター変換に基づく文章の生成

（１）すべてのクラスター数が２以上の文章はクラスター数が１となる文章にクラスター変換を何回か作用することにより得られる。

（２）特に$\alpha=1$のとき、引き続く既約成分(k_l,k_{l+1}とする)に関するクラスター変換を行うと、$\alpha'=2$となる。クラスター数がαとなる文章のつくる分布を$P_k^{(\alpha)}(n)$と書く。変換前の文章の分布を$P_k^{(1)}(n)$とすると$P'^{(2)}_k(n+1)$は次の様になる：

$$P'^{(2)}_k(n+1)=\begin{cases} P^{(1)}_{k_l+k_{l+1}+1}(n)+1 \\ P^{(1)}_{k_l}(n)-1 \\ P^{(1)}_{k_{l+1}}(n)-1 \\ P^{(1)}_k(n)(k\neq k_l,k_{l+1},k_l+k_{l+1}+1) \end{cases}$$

このことからクラスター数が２の文章の分布はクラスター数が１の文章の分布からその座標がk_{l_1}とk_{l_2}上にある分布をひとつ減じて座標$k_{l_1}+k_{l_2}+1$上にひとつ加えることにより得られることが分かる。これより大きなクラスター数をもつ文章は小山現象を記述する可能性のあることが分かる。

次に小山現象の解析を行う。$\mu_k^{(\alpha)}(n)=P_k^{(\alpha)}(n)/P^{(\alpha)}(n)$とおく、ここで$P^{(\alpha)}(n)$は$P_k^{(\alpha)}(n)$の$k$についての総和である。同様に$\mu_k^{(\alpha)}(n)$の対称化$\rho_k^{(\alpha)}(n)$を考える。

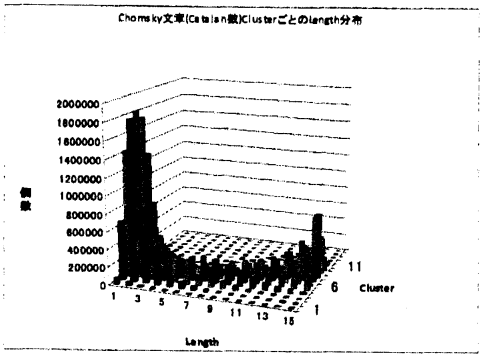


図 24-1 $\mu_k(15)$のクラスター数の大きい分布

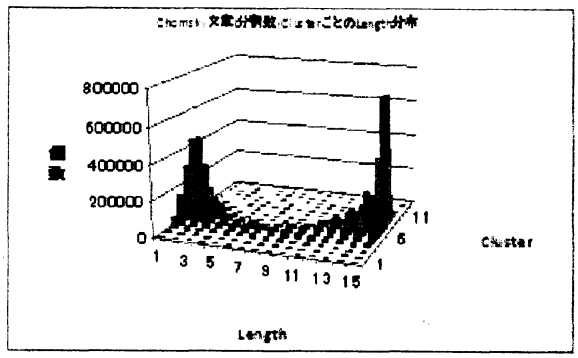


図 24-2 $\rho_k(15)$のクラスター数の大きい分布

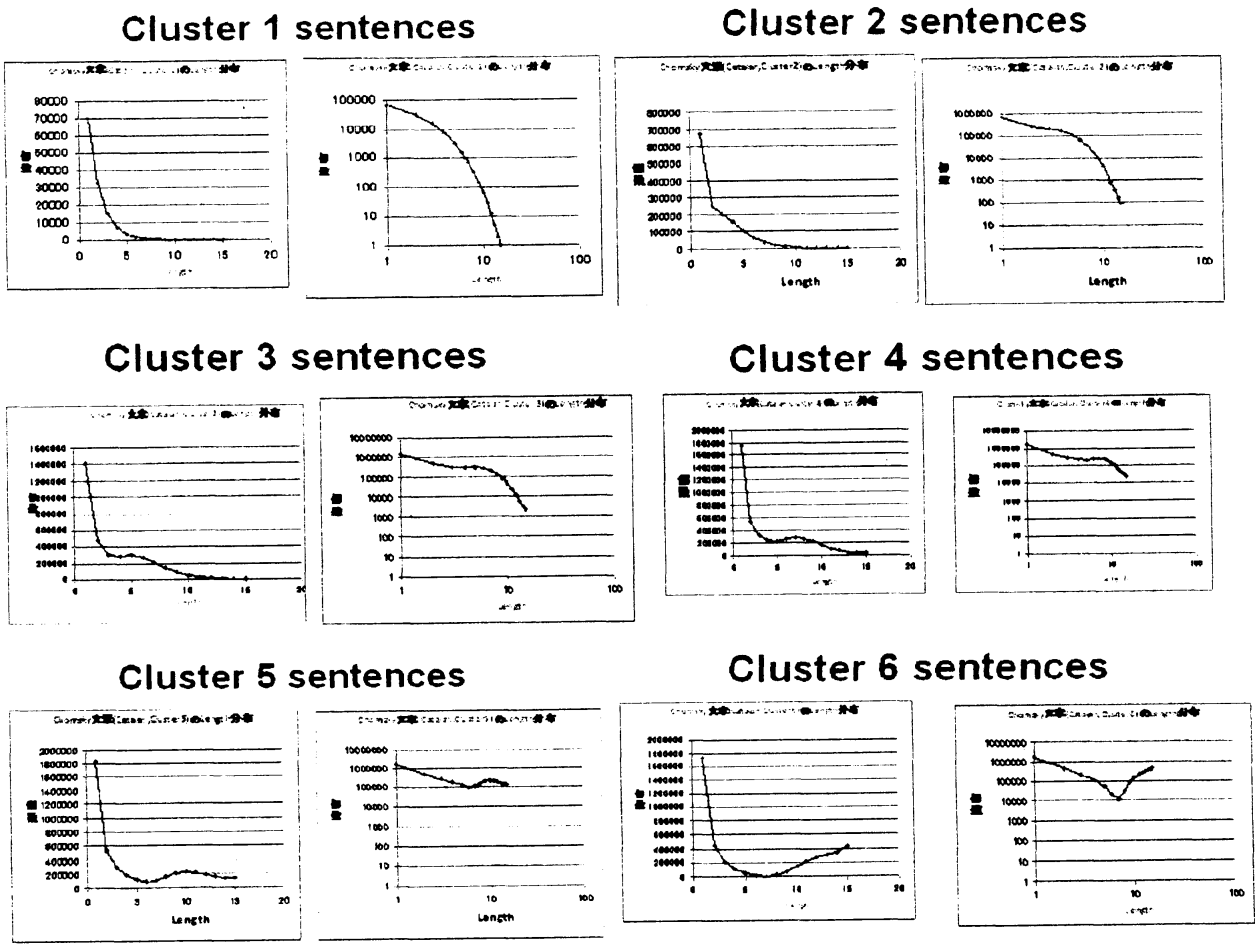


図25　$\mu_k^{(\alpha)}(15)$ $(\alpha=1,2,\cdots,6)$ の分布

すなわち $\rho_k^{(\alpha)}(n)=Q_k^{(\alpha)}(n)/Q^{(\alpha)}(n)$ を考える。この両者の分布を示す(図 24-1,2)。高いクラスター数の文章の分布に小山現象が見て取れる。これより大きいクラスター数をもつ分布は小山あるいは小谷現象を記述する可能性を持っていることが分かる。§ 3の例についてフィッティングを行い、クラスター変換により再構成することは興味ある主題と思われる。

8. 一般化されたカタラン数によるチョムスキー文章の解析

ここではカタラン数あるいは一般化されたカタラン数を用いて文章の分布を解析する([11])。カタラン数とはベルギー人数学者 E.C.Catalan によって導入された様々な組み合わせの数え上げに現れる数 C_n であり、例えば次の問題を考察する際に現れる([13]):「n 個の量 $a_1,\cdots,a_n$ の積 $a_1\cdots a_n$ の計算を 2 つの積の計算の繰り返しで行うとして何通りの方法が可能か？」。この答えを表す数をカタラン数といい、次のように表わすことが出来る:

$$C_n=\frac{(2n)!}{(n+1)!n!}$$

このことよりカタラン数は長さ n のすべての文章の総数を表すことが分かる。ここでは、次の一般化されたカタラン数も考える([11] [14])。

$$C_{n,m}=\frac{m(2n-m-1)!}{(n-m)!n!}$$

小さいn,mについては次の表を得る。

n \ m	1	2	3	4
1	1			
2	1	1		
3	2	2	1	
4	5	5	3	1

表1 $C_{n,m}(n,m=1,2,3,4)$

次に文章を既約成分の個数に着目して考える。$C(n\,,\,m)$の文章の個数を$P(n\,,\,m)$と表すことにする(定義2, p.4)。小さな個数について具体的な例を表2に挙げておく。

n \ m	1	2	3	4
1	{}			
2	{{}}	{}{}		
3	{{{}}} {{}{}}	{}{{}} {{}}{}	{}{}{}	
4	{{{{}}}} {{{}{}}} {{}{{}}} {{{}}{}} {{}{}{}}	{{}}{{}} {{}{}}{} {}{{}}{} {}{}{{}} {{{}}}{}	{}{}{{}} {}{{}}{} {{}}{}{}	{}{}{}{}

表2 $P(n,m)\ (n,m=1,2,3,4)$

両者の比較を行うと次の定理が予想され、実際に証明することが出来る：

定理Ⅲ 次の等式が成り立つ:

$$P(n,m)=C_{n,m}=\frac{m(2n-m-1)!}{(n-m)!n!}$$

［**証明**］以下$P(n,0)=0$と定める。次の補題を示すことにより示される。

補題 次の等式が成り立つ。

(1) $P(n,n)=1$

(2) $P(n,m)=P(n,m+1)+P(n-1,m-1)$

[証明]

(1)は明らかである。

(2)の証明

基本構成により$C(n,m)$からは下記の集合に属する文章がそれぞれ唯ひとつ生成される。

$$C(n+1,m+1)\,,\,C(n+1,m)\,,\,C(n+1,m-1)\,,\cdots,\,C(n+1\,,\,1)$$

したがって、

$$P(n+1,m)=P(n,m-1)+P(n,m)+\cdots+P(n\,,\,n)\quad(m\geq 2)$$

となり

$$P(n+1,m+1)=P(n,m)+P(n,m+1)+\cdots+P(n\,,\,n)$$

となり、従って(2)が得られる。

[定理Ⅲの証明]　補題(2)と次の生成スキーム(図 26)に従って帰納法により示す。

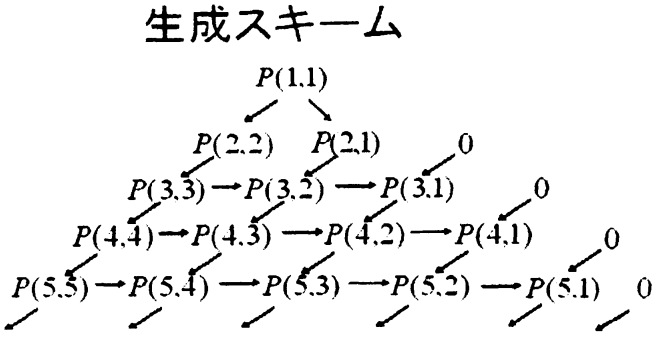


図 26　$P(n,m)$ $(n,m=1,2,3,4,5)$の 生成

$P(1,1)=1,$

$P(2,2)=1, P(2,1)=1,$

$P(3,3)=1, P(3,2)=P(3,3)+P(2,1)=2$

$P(3,1)=P(3,2)+P(3,0)=2$,

.....

.....

$P(n,m)=C_{n,m}(1\le m\le n)(P(n,n)=1)$

と仮定する。

次に$n+1$の時にも成り立つことを示す。

$P(n+1,n+1)=1,$

$P(n+1,n)=P(n+1,n+1)+P(n,n-1)=C_{n+1,n+1}+C_{n,n-1}=C_{n+1,n},$

$P(n+1,k)=C_{n+1,m}$ $(m\le k\le n+1)$と仮定すると

$P(n+1,m-1)=P(n+1,m)+P(n,m-2)=C_{n+1,m}+C_{n,m-1}=C_{n+1,m-1}$

したがって示された。

[定理Ⅰの証明]

定理 III の補題(2)より次の等式を得る：

$$P(n+1,1)=\sum_{k=1}^{n}P(n,k)$$

$P(n+1,1)=C_{n+1,1}=C_n$. 従って全ての文章は基本構成から得られる。

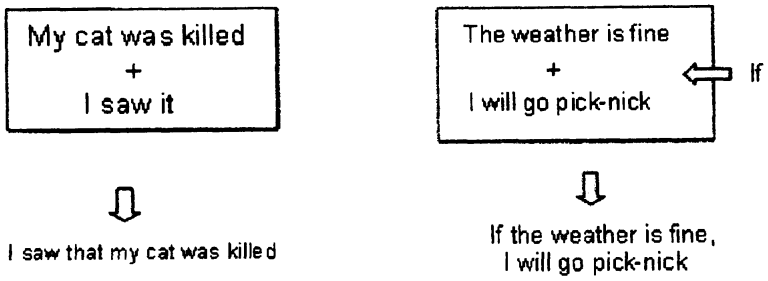
9.チョムスキー文章における Top-down と Bottom-up 構造

最後に文章の作成について考える。まず文章の複雑さはクラスター数により記述されることを例を挙げて説明する。

例文　(1)I　recognize the color red.（Ⅰ-型)

(2)I saw that my cat was killed. (Ⅰ-型)

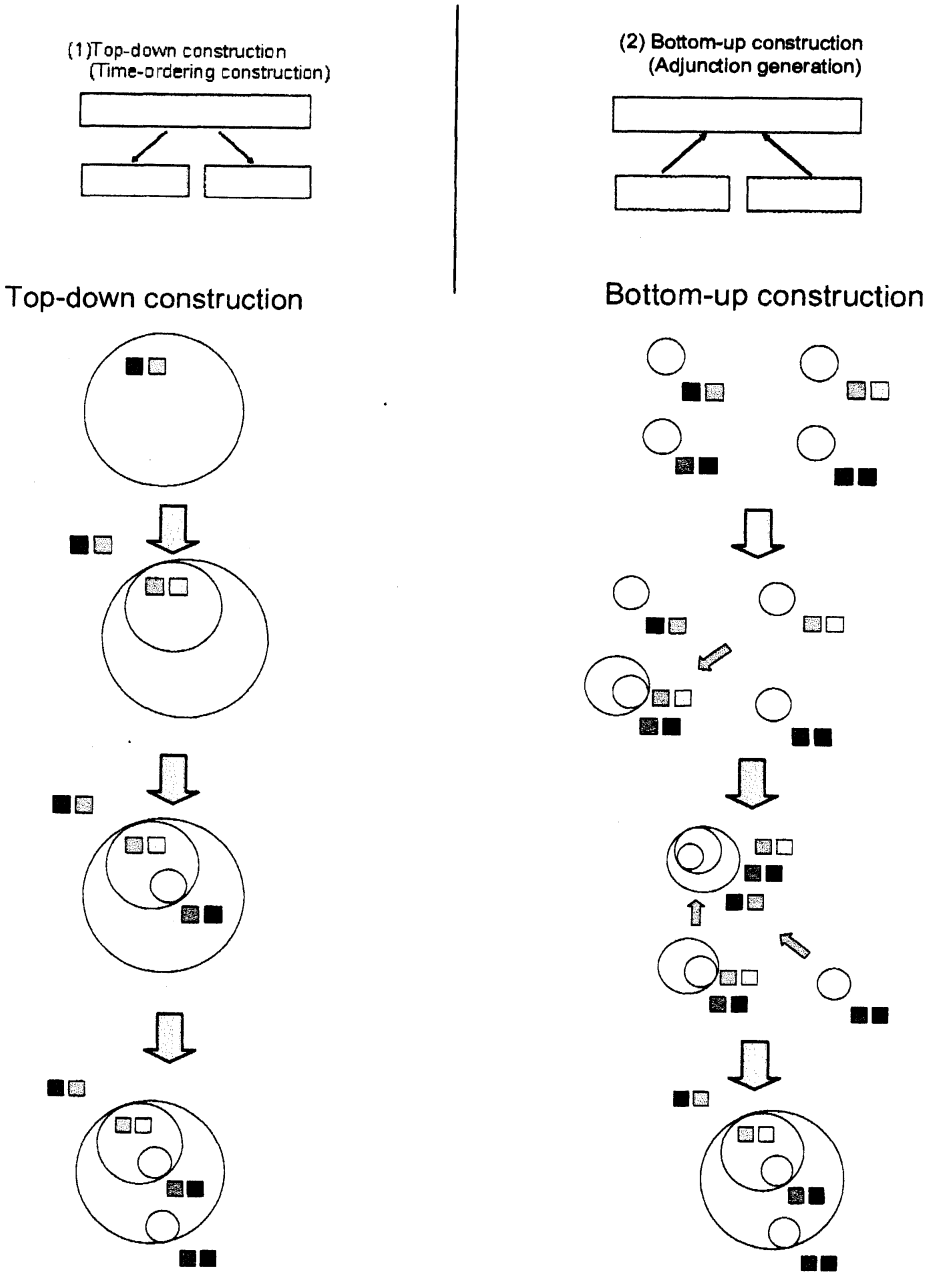
(3)If the weather is fine, I will go pick-nick. (Ⅱ-型)



例文(1)の文章は、論理的順序に配列されて書かれていると考えてよいであろう。これは文章としては基本的なものであると言える(I 型)。

例文(2)の文章は、まず驚きをもって「my cat was killed」があり、次に文章成立条件とし

て「I saw that...」を用意したとも考えられる(I 型)。
例文(3) の文章は、「The weather is fine」および「I go pick-nick」があり、次に文章成立条件として「if」を用意したと考えられる(II 型)。このようにクラスター数の増加するに従って高度な内容を伝える文章になっていることがわかる。
次に「文章はどのように書かれるのであろうか？」という質問について考える。全体の構想があらかじめ定められていて書き下すだけなのか(Top-down method)、それとも全体の構想は持たずに書き始め最後に主張したい内容を実現するのであろうか(Bottom-up method)、これについてチョムスキー文の構成という立場から考えてみる。



上記の流れ図は次の様に理解するものとする:長さが１となる既約な文章を円周で表し、これに色のついた２つの□を与え開き括弧と閉じた括弧をあらわす。このような円周を幾つか用意してどのように配置するかにより文章生成の典型的な構成を２通り考えている。左側の構成では文全体の主張が最初に与えられており、これを詳細に説明するという構成になって

いる。右側の構成では幾つかの文章を用意しておいてこれを組み合わせて構成している。最終的には同一の文章となっていることに注意する。

10. 結論と議論

チョムスキー文章の作る分布が具体的に計算され、既約な文章の個数の数え上げが求められた。これに基づき折れ棒モデル、$1/f$-ゆらぎ及び小山現象が議論された。次ぎの事柄が示された:（1）$1/f$-ゆらぎは変数が十分小さいときにのみクラスター数が1となる文章の分布により記述できることが示された。折れ棒モデルのほうが近似はよいと思われる。（2）折れ棒モデルと我々のモデルとの関係が示された。この両者はともに上に凸となる性質をもつ。その有効性は現実の問題に対するフィッティングにより比較判定されることになる。（3）十分大きなfについて$1/f$-ゆらぎの示す現象の相転移と考えられる小山現象が現れることがある。このことはベキ則だけに固執することは正しくないことを示している。チョムスキー文章のクラスター数による解析は、小山現象を記述する可能性があることが期待させることを示した。以上より我々のモデルは$1/f$-ゆらぎ、折れ棒型のモデルさらに小山現象を同時に記述可能なモデルであると云うことができる。

次の事柄に注意する。$1/f$ゆらぎを完全に記述するにはよりランダムなBottom-upな構造をもつ系を取り上げる必要があるものと思われる。チョムスキー文章のつくる系は木構造のゆえにTop-down的な性質が強すぎるものと思われる。今後はバイオインフォマティックスにおけるDNAの配列に関する不均一性(Inhomogenity)に関する分布を取り上げ$1/f$ゆらぎを記述するかどうかを考察することを考えている([16])。

参考文献

[1]米田正明　広瀬貞樹(他２名)，オートマトン・言語理論の基礎，近代科学社，2003

[2]斉藤成也(他６名)，遺伝子とゲノムの進化(シリーズ進化学２)，岩波書店，2005

[3] N.Chomsky，Syntactic Structures, Mouton ，1957

[4]佐藤勝彦　二間瀬敏久編：宇宙論〈1〉宇宙のはじまり (シリーズ現代の天文学)，2008

[5]井庭崇　福原義久，複雑系入門，NTT出版，1998

[6]武者利光　井上昌二ほか，ゆらぎの科学１，森北出版，1991

[7]吉岡直人，砂崩しの実験　地震ジャーナル第35号，2003/6

[8]長尾真，自然言語処理，岩波書店，2006

[9]R. Macarthur and E. Wilson: The theory of Island biogeography, Princton University press, p.199, 1967,

[10]森　主一　三浦泰三他，集団と生態，朝倉書店 p.306，1977

[11]岩澤秀樹，Catalan 数を用いた整合括弧列の分布の解析とその折れ棒モデル,$1/f$-ゆらぎへの応用，日本大学大学院総合基礎科学研究科修士論文(2009)

[12]G. H. Hardy-S. Ramanujan, Asymptotic formulae in combinatory analysis, Proc. London Math. Soc. (2), 17(1918)75-115.

[13]E. Catalan, "Note sur un probleme de combinaisons," J. Math. Pures Appl. ,(1)3, pp.111-112, 1838

[14] The (Combinatorial) Object Server http://theory.cs.uvic.ca/inf/tree/BinaryTrees.html

[15]John L.Casti.『複雑系による科学革命』. 講談社, 1997 .

[16]高橋秀夫:分子遺伝学概論、コロナ社,p229, 1997

[17] H.A.Simmon: On a class of skew distributions,Biometrica, Vol. 42(1955)425-4400