言語の長相関に内在する再帰的生成過程
Long-Range Correlation Underlying Language and Recursive Generative Processes

田中久美子
Kumiko Tanaka-Ishii
† 東京大学先端科学技術研究センター
RCAST, University of Tokyo
kumiko@cl.rcast.u-tokyo.ac.jp

Abstract

経済、地震、気象など、自然発生的な時系列に対して知られている物理学上の特性として、長相関がある。長相関は、与えられた時系列の任意の二つの部分が、離れていても類似性を持つという、再帰的な性質を示唆する。本発表では、CHILDES コーパスや音楽、プログラミングといった人間が生成した時系列に長相関が成り立つことを示す。その原因を考えるため、数学的生成過程を複数考え、ある単純な再帰的なふるまいが関わっている可能性を示す。

Keywords — 長相関, CHILDES corpus

1. はじめに

経済、地震、気象など、自然発生的な時系列では、大域的な統計法則が成り立つことが知られる。自然言語にもそのようなものが知られており、冪則の形式をとることが多い。冪則とは、ある時系列に対して、異なる二変量の関係を両対数で観測した時に、線形の関係が現れる性質をいう。Zipf 則 [Zipf(1965)]、Heaps 則 [Heaps(1978)] [Herdan(1964)] は代表的で古くから知られるものである。

言語に関する大域的な統計法則の中で、比較的近年に計測できるようになったものに、長相関がある。長相関とは、与えられた時系列の二つの部分が、それがどんなに離れていても一定の類似性を持つ性質をいう。この定義からも、長相関は再帰的な概念であり、時系列のフラクタル的性質を検証するための一つの方法論でもある。長相関はもともとは Hurst が水量データの解析に用い [Hurst(1951)]、以来、経済時系列、地震、気象など、数値時系列に対して計測方法が考案されてきた。一方で、自然言語は数値時系列ではないため、長相関を算出する適切な方法が、長い間わからなかった。近年この点の研究が進み、自然言語の時系列を単語の間隔で捉えることにより、計算可能となった。本稿では、まず、この方法を用いて、さまざまな種類の自然言語に関するデータの中に、Zipf 則、Heaps 則、長相関が成立することを示す。

その上で、なぜこのような法則が言語において成立するのか、本研究では Simon の生成モデルを元に議論する。Zipf 則の成立に関しては、Mandelbrot が、コミュニケーションの全体最適化が背景にあるとして説明している。しかしながら、発話の時々で、発話全体を見通した上で最適化をしているとは、にわかには考えられない。人が毎回毎回行う発話行為の中に何か隠された「からくり」のようなものがあり、その行為は実は全体最適化につながる行為であり、人はその行為を学ぶことにより言語を使えるようになる、と考える方が自然ではなかろうか。本稿では、その「からくり」として何があるのかを数学的生成過程を通して考える。

なお、本稿の元論文は英文で現在投稿中の内容となる。本稿はこれのポイントをまとめ、また関連事項を付記したものである。

2. 言語に成り立つ３つの大域的な統計法則

本稿では、自然言語について成り立つことが報告されている Zipf 則 [Zipf(1965)]、Heaps 則 [Heaps(1978)]、長相関の三つの大域的な統計法則を考える。Zipf 則とは、与えられた文書に対し、u を単語の頻度順位、その単語の頻度を $F(u)$ とすると、

$$F(u) \propto u^{-\xi}, \quad \xi \approx 1.0 \tag{1}$$

との性質が成り立つことである。一方、Heaps 則は、文書中の語彙量と文書量の関係であり、m を文書量, $V(m)$ を語彙量として、

$$V(m) \propto m^{\zeta}, \quad \zeta < 1.0 \tag{2}$$

が成り立つという性質である。特に、自然言語の場合には、指数 ζ が、1.0 よりもかなり小さい点に特徴がある。

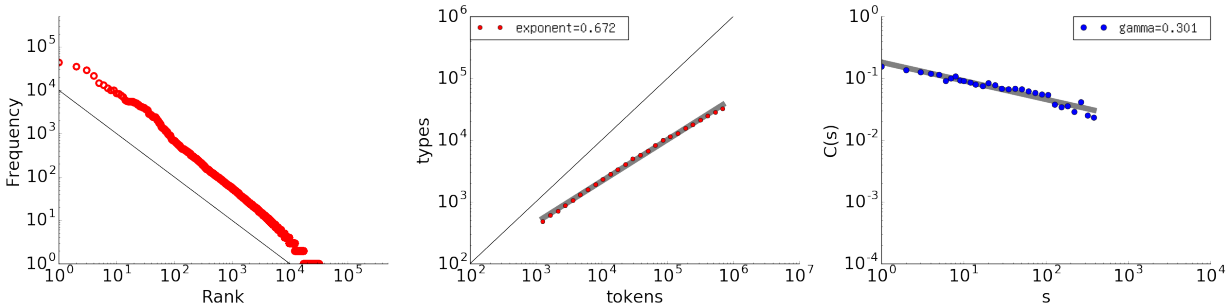

図 1　レ・ミゼラブル全文 (フランス語) で成り立つ 3 つの大域的な統計法則 : 左から Zipf 則、Heaps 則、長相関。

例として、レ・ミゼラブルの場合を図 1 に挙げる。左の図がレ・ミセラブルの順位頻度分布を示す。横軸は順位、縦軸は頻度を表し、赤い点群がデータの実プロットである。下側の細い黒い線が傾き-1 の直線を表している。赤プロットと、黒直線を比較すると、上の定義のとおり、順位頻度分布においては冪指数はほぼ-1.0 にに近いことがわかる。

中央の図は文書量に対する語彙量を表している。横軸は文書量、縦軸は語彙量で、赤点群がレ・ミゼラブルの実プロットである。図中の黒直線は傾き 1.0 であるが、それとの比較では、赤プロットの傾きは小さい。傾きを算定すると、0.672 となり、回帰直線を赤プロットの背後に薄いグレーで示している。

長相関とは、与えられた時系列の中の二つの (長い) 部分列が、ある程度離れていても似ているという再帰的な性質を表す。歴史的に、Hurst 法 [Hurst(1951)] や、Detrended Fluctuation Analysis [Kantelhardt *et al.*(2001)] [Kantelhardt(2002)] などさまざまな検出方法があるが、それぞれの手法の問題なども近年では明らかとなり、昨今では、最も基礎的な自己相関関数で計測するのが妥当であると報告されている [Lennartz and Bunde(2009)]。$R = r_1, r_2, \ldots, r_N$ を時系列として、その平均と分散を μ,σ とすると、自己相関関数は、

$$C(s) = \frac{1}{(N-s)\sigma^2} \sum_{i=1}^{N-s} (r_i - \mu)(r_{i+s} - \mu) \qquad (3)$$

で定義される。つまり、距離 s の開きがある二つの時系列部分の共分散を時系列の分散で正規化しており、すなわち、距離 s の二つの時系列の類似度を計算するものである。これが、

$$C(s) = C(1)s^{-\gamma}, s > 0 \qquad (4)$$

つまり、距離 s に対して、冪となる時、長相関があるという。

この自己相関関数を自然言語にどのように適用するかが長く問題であった [Ebeling and Pöschel(1994)] [Altmann *et al.*(2009)] [Altmann *et al.*(2012)]。近年の報告によると、単語の間隔時系列に対して、極値解析を適用することにより、長相関がある場合には安定してそれを観測することができる [Tanaka-Ishii and Bunde(2016)]。具体的には、文書中の稀な単語集合を考え、集合に含まれる単語群全てを用いて間隔の時系列を構成して R とする。これに対して自己相関関数を計測することで、長相関を観測することができる。 [Tanaka-Ishii and Bunde(2016)] に基づき、以下、本稿で示す長相関は、文書長の 1/16 の単語を稀な順からとり、それをもとに、間隔時系列を構築し、自己相関関数を計測したものである。

図 1 の右図はこのようにして得られた長相関である。横軸は距離 s、縦軸は $C(s)$ である。距離に対して類似度が冪でしか減少しない、つまり、どこまで s を大きくしても、一定の類似性が見られることがわかる。たとえば 0 と 1 がランダムに現れるバイナリ時系列の場合には、類似度 $C(s)$ は、0 付近の正負の小さな値をとることになる。図は両対数であるので、当然正の領域のみを示している。つまり、レ・ミゼラブルの自己相関関数は、負の値は一切現れず大きく正となっており、しかも、どの距離 s をとっても時系列は類似していることを示している。

[Tanaka-Ishii and Bunde(2016)] では、十の長い単著の文学作品においては長相関がよく成り立つことが報告されている。また、筆者の未発表の研究ではあるが、Gutenberg プロジェクトから長いテキストを 1000 以上得て、それに対して長相関を調べたところ、長相関が成り立たないものはむしろ珍しく、おしなべて長相関が見られることがわかっている。

3.　幼児の発話、音楽、プログラム

以上のような統計法則は、実は自然言語に関連する、他の人間のコンテンツでも成り立つ。本節では、

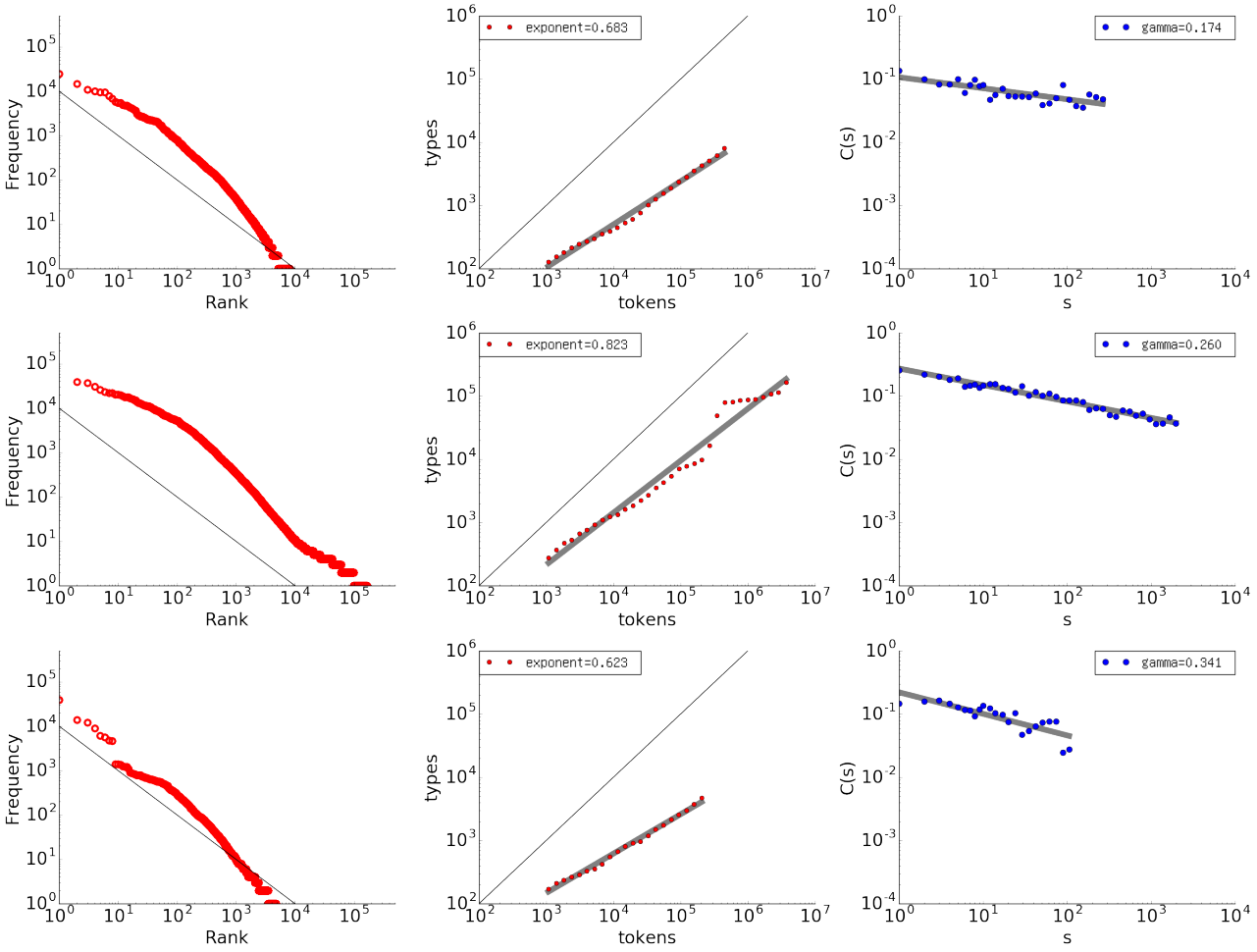

図 2　CHILDES コーパスの Thomas の発話 (上)、Lisp(中央)、バッハのマタイ受難曲 (下) での三つの大域的な統計法則。

三つの異なる例として、CHILDES コーパス最長の発話データである Thomas のデータ、プログラムの Lisp コード、バッハのマタイ受難曲、を挙げる。

図 2 は、それぞれのデータの順位頻度分布 (左)、文書量に対する語彙量 (中央)、長相関 (右) を上から順に表している。

CHILDES コーパスは、前処理をほどこし、XXX などとなっている部分を削除している。また、Lisp については、世界中の Lisp のソースを集め、その上で、プログラムコードは、コピー&ペーストで生成されることも多いため、大きな重複のみ削除するなどの処理を施し、括弧を削除したデータを時系列としている。最後に、マタイについては、MIDI データからおこしているが、前処理にかなりの恣意性がある。今回は、複数に重なる楽器ごとの旋律を、一旋律として単純につなげた状態で時系列としている。

結果の図は、前節のレ・ミゼラブルの結果と、どれも若干異なりはする。たとえば、幼児の発話は、順位頻度分布がレ・ミゼラブルのそれよりも上に凸の傾向が強いし、Lisp では、語彙の増大が途中で不連続になってもいる。しかし、おおまかには、これらの人由来のデータにおいて、3 つの大域的な統計法則が成立するといっても過言ではないであろう。事実、順位頻度分布の傾きは-1.0 に近いといえるであろうし、文書量に対して語彙量が増大する傾きは自然言語同様に 1.0 よりは小さい。そして、長相関はどれにも現れており、自己相関関数に負値は現れない。

同様の傾向を CHILDES コーパス最長の 10 人のデータ、10 の長い交響曲、他のプログラムコード (C++や python など) で調査したところ、それぞれに個性はあるが、やはりおおまかな傾向は一致する。

言語、音楽、プログラムなどはどれも人の記号活動を基礎とする。また、幼児の発話は、文法や語彙の使用もたどたどしい。にもかかわらず、これらにおいてこのように 3 つの統計法則が成立する理由は何であるのか。

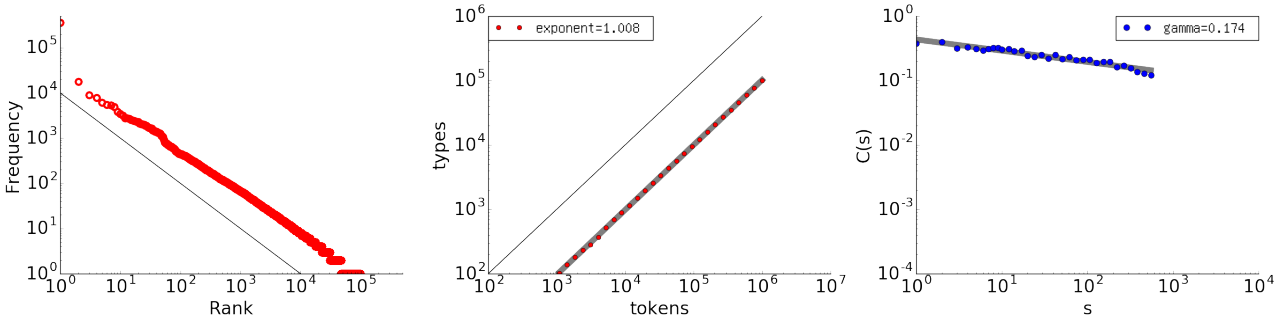

図 3　$\alpha = 0.10$ の場合に Simon モデルを用いて 100 万要素生成した時系列における順位頻度分布、文書量に対する語彙量、長相関。

4. Simon モデルとその変種

Zipf 則の成立原因については、Mandelbrot の [Mandelbrot(1952)] の論が有名である。Mandelbrot はコミュニケーションの効率を最良にすることにより、Zipf 則を導くことができることを数学的に証明した。言語の背景に、効率化の概念があることは、言語学上同時期に他の論も知られる [Martinet(1960)]。とはいえ、Mandelbrot の論は Zipf 則のみを説明し、特に法則の中でも長相関との関係は明らかではない。

しかしながら、果たしてコミュニケーションを全体最適化するようなことを我々は行って発話をしているだろうか。むろん、発話に意図がある以上、その意図に沿って、なるべく話が伝わるように工夫して話をするだろう。しかし、毎回の発話において、発話の最後の部分まで見通して、全体最適化を行っているかは疑問である。特に、幼児の発話においても冪則は成立するのである。すると、むしろ、発話一回一回の行為に何かがあり、それは、結果として全体として最適化をする行為となっており、そのような行為の仕方を子供は学ぶのである、と考える方が自然であろう。すると、大域的な統計法則の成立原因を考える上では、一回一回の発話行為を考察することが必要となる。その基礎的なモデルとしての数学的生成過程の統計的ふるまいを考察することが一つの手がかるとなるだろう。

基礎的な数学的生成過程には、マルコフ過程やポアソン過程、再生過程など含まれるが、以上はいずれも語彙数が有限のモデルである。自然言語は発話に応じて新規語彙が常に導入される系であり、だからこそ、Heaps 則が成り立つ。

語彙数が無限に増える生成過程の代表には、複雑系科学の先がけとして提案された、Simon モデル [Simon(1955)] がある。以下、紙面の幅の関係から、簡略的な記法をとる。K_t を時刻 t における (単語や文字など) 要素数とし、時刻 t までに要素 i が現れた数を $S_{t,i}$ とする。時刻 $t = 0$ では、

$$K_0 = 1, \quad S_{0,1} = 1, \quad S_{0,i} = 0, \quad i = 1, 2, 3 \ldots$$

として生成を始める。Simon 過程は、時刻 $t > 0$ においては、α を与えられたパラメータとして、

$$\begin{aligned} P(\text{新規単語を生成}) &= \alpha, \\ P(\text{要素 } i \text{ を生成}) &= (1-\alpha)\frac{S_{t,i}}{t}, i = 1, \ \ldots, K_t. \end{aligned}$$

と生成する。上の一行目は、事前に決めた α の確率で、新規単語を生成することを表す。二行目は、それ以外の $1-\alpha$ の確率で、それまでに生成した時系列から無作為に一つ要素をサンプリングすることに相当する。式上は、各要素を、過去に現れた頻度に比例する割合で選択する、という定義であるが、それは無作為選択と同じことである。

Simon 過程を $\alpha = 0.1$ として、100 万要素生成し、図 3 に、左に順位頻度分布、中央に文書量に対する語彙量、右に長相関を示している。Zipf 則がだいたい成り立つこと、また語彙量が増大する傾きが 1.0 であることは、数学的に証明されており、理論どおりといえる。長相関は、その処理工程が複雑であるため、証明は困難であるが、実験的に成り立っていることが図からわかる。この三つのグラフを見ると、Simon モデルは自然言語の特性をふまえているかにみえる。

しかしながら、自然言語と決定的な差があり、それは文書量に対して語彙量の増大速度が早すぎるという点である。通常、自然言語の Heaps 則の傾きは 0.55-0.75 くらいであり、1.0 ということは筆者の知る限り観測されたことはない。とすると、語彙量の増大速度を抑えたモデルを考える必要がある。

この点を改良したモデルは、Pitman-Yor 過程として知られている [Pitman(2006)]。Pitman-Yor 過程は、2000 年代に Bayes 推定 [Teh(2006)] [Goldwater *et al.*(2009)] の枠組みとして自然言語処

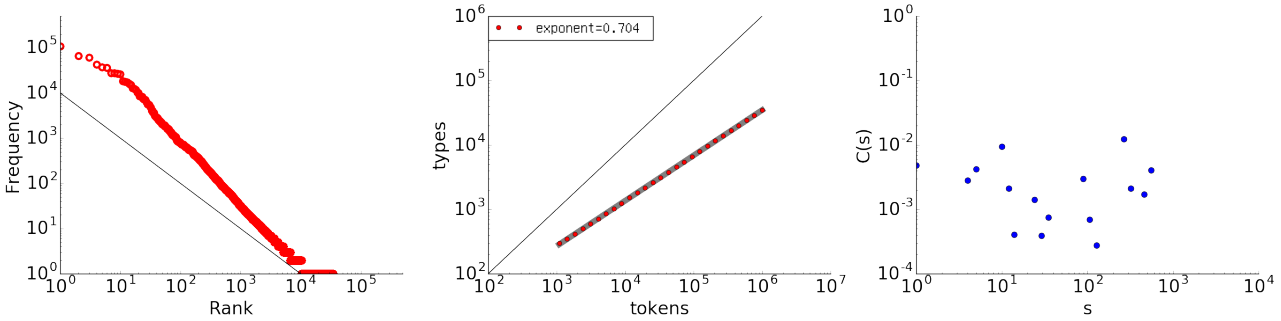

図 4　$a = 0.68, b = 0.80$ として Pitman-Yor モデルを用いて 100 万要素を生成した時系列の、順位頻度分布、文書量に対する語彙量、長相関。

理で用いられた。Pitman-Yor 過程は Simon モデル同様に $t = 0$ から始まる。$t > 0$ においては、事前に与えられたパラメータ a,b に対して、

$$P(\text{新規単語を生成}) = \frac{aK_t + b}{t + b},$$
$$P(\text{要素 } i \text{ を生成}) = \frac{S_{t,i} - a}{t + b}, i = 1, \ldots, K_t$$

として生成される。Pitman-Yor 過程では、各単語が現れた単語の回数を a だけ小さく捉え、その分を新規単語の生成確率として割り振る。また、順位頻度分布は、実データではどれも右上凸の構造をしているが、Pitman-Yor はこの点をモデル化するため、パラメータ b を導入し、b が大きいほど、大きく凸になるように工夫されている。Pitman-Yor 過程は、Simon のような無作為サンプリングに基づく過程ではないため、各単語の現れ方を毎回の生成ごとに統計として保存せねばならず、この意味で、計算機上でも Simon モデルよりも生成速度が遅いのが特徴である。

Pitman-Yor 過程を $a = 0.68, b = 0.80$ の場合として、100 万要素生成した。$a = 0.68$ としたのは、a とだいたい同じ傾きで語彙が増大することを、数学的に示すことができるからである。図 4 に左から、順位頻度分布、文書量に対する語彙量、長相関が示されている。中央の図から、語彙量は傾き 0.704 で増大しており、$a = 0.68$ に近い値となっている。また、左図の順位頻度分布の右上凸の性質は、$b = 0.8$ の場合でも少し現れていることがわかる（さらに b を大きくすると、Thomas の例のように大きく湾曲させることができる)。一方で、長相関については消えてしまっている。a,b のパラメータを網羅的に変化させ、時系列を生成して試したが、Pitman-Yor 過程では長相関がまったく成り立っていない。

Simon で語彙量が早く増大し過ぎ、一方で、Pitman-Yor で長相関が成立しない、との状況から、二つのモデルを折衷した生成過程を考える。最も単純なものとしては、$t > 0$ において、a,b をパラメータとして、

$$P(\text{新規単語を生成}) = \eta, \quad where \ \eta = \frac{aK_t + b}{t + b}$$
$$P(\text{要素 } i \text{ を生成}) = (1 - \eta)\frac{S_{t,i}}{t}, i = 1, \ldots, K_t$$

が考えられる。一行目の新規単語生成確率は、Pitman-Yor どおりで要素の語彙数 K_t にしたがって減衰させ、二行目の、すでに現れている要素 i の生成確率は Simon どおりに過去に生成した時系列からの無作為サンプリングにより行う。折衷モデルの生成では、既存単語の生成は Simon 同様であるため、生成は高速に行うことができる。

折衷モデルを用いて、Pitman-Yor の場合同様、$a = 0.68$ and $b = 0.80$ として、100 万要素を生成した。この時系列の順位頻度分布、文書量に対する語彙量、長相関を図 5 に示す。中央の語彙量の増大速度が、自然言語同様に 1.0 よりも小さく実現されている。一方で、右図の長相関も成立している。長相関の傾きがやや小さいなど若干の差はあるが、自然言語の大域的な統計法則を質的に満たす一つの数学的生成モデルとして、折衷モデルがあることになる。

数学的生成過程は、自然言語の生成過程とは根本的に異なるものではある。とはいえ、自然言語で成り立つ大域的な統計法則が成り立つものに関しては、自然言語の生成過程を考えるヒントが隠されていると考えてもよいだろう。以上の論からは、長相関を生成する一つの生成過程は、「新規単語を導入しながら、過去から無作為にサンプリングする過程」であることが示唆される。実際、Simon でも、折衷モデルでも無作為サンプリングにより、既存単語を生成している。一方の、Pitman-Yor は無作為サンプリングでは生成できない。過去の時系列からの無作為サンプリングは、前述のように、計算上も各単語の統計をメモリ上に格納する必要がないため、効率がよい。無作為サンプリ

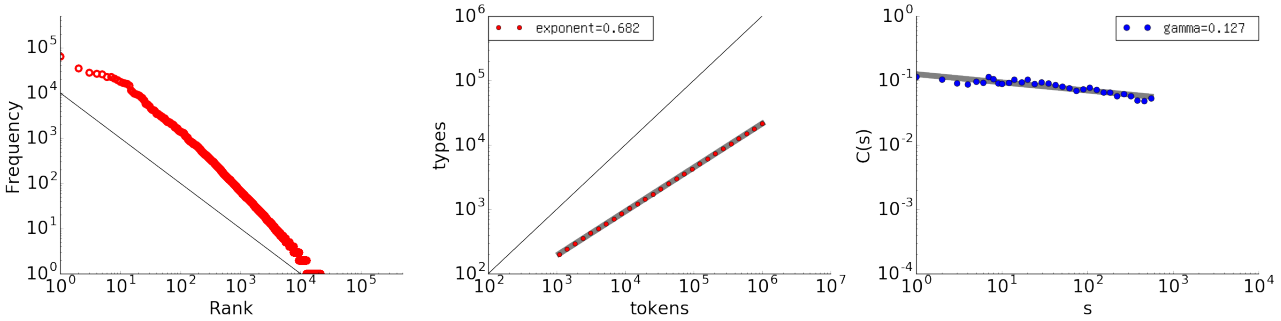

図 5　Pitman-Yor モデルの新規単語導入率に、Simon モデルの既存の単語生成確率を合わせた、折衷生成過程における、順位頻度分布、文書量に対する語彙量、長相関。$a = 0.68$ and $b = 0.80$ の場合、100 万要素。

ングの効率の良さは、人においても同様である可能性はあるだろう。そして、無作為サンプリングの本質とは、自己の過去の発話を再利用する再帰的な過程である。実際の人間の発話は、むろんのこと、無作為サンプリングからはかけはなれてはいようが、無作為サンプリングが人間の言語の長相関に何らかの意味で関わる可能性を否定するものではないだろう。つまり、本稿の最初に述べた、一つ一つの発話行為の背景にある「からくり」として、過去の発話からの再帰的なサンプリングがあるのかもしれない。

5.　結論

本稿では、Zipf 則、Heaps 則、長相関という自然言語の三つの大域的な統計法則を確認し、特に近年計測が可能となった長相関に視点を当て、大域的な統計法則を満たすような数学的生成過程を考えることにより、自然言語の性質を考えた。長相関とは、時系列中の部分列の類似性、つまり自己相似性を表す一つの性質である。

上述の三つの大域的な統計法則は、自然言語の単著文書だけでなく、言語的な人間の他の記号過程の例として、幼児の発話、プログラムソース、音楽などでも成立する。なぜこのような大域程な統計法則が成立するのかに関しては、発話全体の最適化の観点からの論が知られている。しかし、発話とは逐次に行うもので、未来の発話を含む全体の最適化を前提とする論には疑問が残る。すると、逐次発話の中に大域的な統計法則を成立させる何らかの要因があるだろう。

逐次の発話のモデルとして、数学的生成過程を考えた。最も基礎となる Simon モデルでは、大域的な統計法則は特に文書量に対する語彙量の増大速度が早すぎる一方で、美しい長相関が観測された。語彙増大速度を抑えた Pitman-Yor モデルでは、今度は長相関が成立しなかった。そこで、Simon モデルと Pitman-Yor モデルを折衷した単純な生成モデルを考えたところ、自然言語の大域的な統計法則を大まかに満たす時系列を生成することができた。状況を総合すると、新規単語を導入しながらも、既存の単語を過去の発話から再帰的に無作為サンプリングを行う過程が、長相関を成立させる要因 となっていることが考えられる。自然言語の発話過程と、過去の発話からの無作為サンプリングとの関わりを考察することが、今後の認知科学上の課題となりえる。

参考文献

[Altmann *et al.*(2009)] Altmann, E.G. , Pierrehumbert, J.B. , and Motter, E.A. (2009). Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words.

[Altmann *et al.*(2012)] Altmann, E.G. , Cristadoro, G. , and Esposti, M.D. (2012). On the origin of long-range correlations in texts. *Proceedings of the National Academy of Sciences*, **109**, 11582–11587.

[Ebeling and Pöschel(1994)] Ebeling, W. and Pöschel, T. (1994). Entropy and long-range correlations in literary english. *Europhys. Letters*, **26**, 241–246.

[Goldwater *et al.*(2009)] Goldwater, S. , Griffiths, Thomas L. , and Johnson, M. (2009). A baysian framework for word segmentation: Exploring the effects of context. *Cognition*, pages 21–54.

[Heaps(1978)] Heaps, H. S. (1978). Information retrieval: Computational and theoretical aspects, academic press. page 206208.

[Herdan(1964)] Herdan, G. (1964). *Quantitative Linguistics*. Butterworths.

[Hurst(1951)] Hurst, H. E. (1951). Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers*, **116**, 770–808.

[Kantelhardt *et al.*(2001)] Kantelhardt, J. W. , Koscielny-Bunde, E. , Rego, H. H. A. , Havlin, S. , and Bunde, A. (2001). Detecting long-range correlations with detrended fluctuation analysis. *Physica A*, **295**, 441–454.

[Kantelhardt(2002)] Kantelhardt, J. W. et al. (2002). Multifractal detrended fluctuation analysis of nonstationary time series. *Physica A*, **316**, 87.

[Lennartz and Bunde(2009)] Lennartz, S. and Bunde, A. (2009). Eliminating finite-size effects and detecting the

amount of white noise in short records with long-term memory. *Physical Review E*, **79**(066101).

[Mandelbrot(1952)] Mandelbrot, B. (1952). An informational theory of the statistical structure of language. *Proceedings of Symposium of Applications of Communication theory*, pages 486–500.

[Martinet(1960)] Martinet, André (1960). *Elements de linguistique générale*. Colin.

[Pitman(2006)] Pitman, J. (2006). *Combinatorial Stochastic Processes*. Springer.

[Simon(1955)] Simon, H.A. (1955). On a class of skew distribution functions. *Biometrika*, **42**(3/4), 425–440.

[Tanaka-Ishii and Bunde(2016)] Tanaka-Ishii, K. and Bunde, A. (2016). Long-range memory in literary texts: On the universal clustering of the rare words. *PLOS One*. `http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0164658`.

[Teh(2006)] Teh, Y.W. (2006). A hierarchical bayesian language model based on pitman-yor processes. In *Annual Conference on Computational Linguistics*, pages 985–992.

[Zipf(1965)] Zipf, G.K. (1965). *Human behavior and the principle of least effort: An introduction to human ecology*. Hafner.