ことばのゆらぎ，ことばのかたち

Fluctuations and structures in natural language

宮崎修次，高口太朗
Syuji Miyazaki, Taro Takaguchi

〒606-8501 京都市左京区吉田本町　京都大学　情報学研究科
Graduate School of Informatics, Kyoto University, 606-8501, Kyoto

Abstract

Based on large deviation analysis and recurrence time statistics, natural languages are studied. It is shown that fluctuations hidden in the languege reflect a style of writing depending on the author, and quantify the naturalness of human writing or the artificiality of unsolicited blog constructed automatically as senseless arrangements of keywords.

Keywords

Large Deviation Analysis, Language

われわれは主に非線形物理学・非平衡統計力学の観点からカオス力学系やグラフ・ネットワーク上の酔歩に関する大偏差統計解析を行ってきた [1)]．ランダムまたはカオス的に変動する量について，その局所平均の分布の，中心極限定理に従った正規分布への収束の速さを測る量を揺らぎスペクトルと呼ぶ．時系列の揺らぎの特徴は揺らぎスペクトルなどを用いる大偏差統計解析によって捉えられる．われわれの研究テーマはこのような大偏差統計解析の手法を更に深化させるとともに，その手法により自然言語に潜む揺らぎを捉えることである．

１つの方針は，文章を単語の時系列とみなし大偏差統計解析を適用することである．今日の文章の数量解析では，文の長さ・各品詞の使用率など種々の静的な統計量を調べることにより，文学作品の作者の真贋判定などが研究されている [2)]．英語の単語を構成するアルファベット数，言い換えば単語の長さを例にとろう．ある文章の中で単語の長さの頻度分布や平均・最頻値を見るのが従来の研究であるが，われわれは文章を，その初めから終わりまでの単語の出現順を時間とみなした各単語の長さという観測量のランダムな時系列だと捉え，有限の時間幅での局所平均の分布を中心にした大偏差統計解析を考えたい．この方が，文章の前後関係を無視した静的な統計量に比べ文章の流れをある程度含むことができると考えており，文章に潜むこのような揺らぎに作者の個性や文章の自然さが反映されるものと期待している．

もう１つの方針は，単語間の関係から構成されるネットワークに大偏差統計解析を用いることである．文章の普遍的な統計則として，単語の出現頻度が逆冪則となるジップ則が知られている [3)]．言語の習得過程にスケールフリーネットワークの生成原理である優先的選択 [4)] が潜み，単語の繋がりをネットワークとみなしたときにスケールフリー的となっているのかといった複雑ネットワークの視点からの自然言語の解析も試みたい．われわれは既に複雑ネットワークの大偏差統計解析に関する研究を行っているが，その知見が何らかの役に立つものと期待している．また，例えば地震の統計則にはジップ則的なマグニチュードと頻度の関係と，1/f スペクトル的な余震の発生間隔の分布の冪則が対となって現れるが，自然言語については 1/f スペクトル的な解析が少ないように思われる．また，われわれの研究によれば，旧約聖書のある単語の出現間隔は品詞や格によって指数関数分布になったり，冪分布になったりして単純ではない．単語間のネットワークにおいて，単語の出現間隔はあるノードへの再帰時間と解釈できる．その統計特性を中心に自然言語の 1/f スペクトル的な捉え方を探求するとともに，再帰時間の大偏差統計解析も試みたい．

Reference

[1)] S. Miyazaki, *Forma* **22**, 141-155 (2007).
[2)] 村上征勝他, “言語と心理の統計”(岩波書店) (2003).
[3)] M. E. J. Newman, *Contemp. Phys.* **46**, 323-351 (2005).
[4)] A. L. Barabási and R. Albert, *Science* **286**, 509-512 (1999).