ビット表現を用いた日本語テキストの正規数性の評価

田窪 洋介 (新居浜工業高等専門学校、高エネルギー加速器研究機構) *
浅原 正幸 (国立国語研究所) †
山崎 誠 (国立国語研究所) ‡

Studying Borel normality of Japanese texts in bianry expression

Yosuke Takubo (Niihama College, High Energy Accelerator Research Organization)
Masayuki Asahara (National Institute for Japanese Language and Linguistics)
Makoto Yamazaki (National Institute for Japanese Language and Linguistics)

要旨

0 と 1 をランダムに出力する乱数生成器は、暗号化通信で多用されている。近年の通信技術の発展によって、通信のセキュリティを向上させるためにより高い乱数性をもつ乱数生成器の必要性が増している。そのため、乱数生成器の開発において、ビット列の乱数性を定量的に評価することが重要となる。我々は、乱数性の評価に使用されている手法をテキストの統計的解析に応用することを考えた。最初の試みとして、日本語テキストの正規数性 (Borel normality) を評価した。日本語テキストを UTF-8、SJIS、EUC の文字エンコーディングを用いて 0 と 1 のビット表現に変換し、正規数性の指標を計算した。本稿では日本語テキストの正規数性の特徴について議論する。

1. はじめに

統計的解析手法は、言語や（新聞、雑誌、書籍などの）レジスタの特徴を定量的に分析するために非常に有効である。このような解析では、一般的には文字を最小単位として取り扱う。一方で、文字は UTF-8、SJIS、EUC などの文字エンコーディングによって、ビット表現としても記述できる。文章は文字で構成されているため、文字エンコーディングで表現された文章も言語やレジスタの情報を含有していると予想される。

乱数生成器は暗号化通信に欠くことのできない技術であるが、その性能を評価するためにはビット列の乱数性を定量的に分析することが必要不可欠である。正規数性 (Abbott et al. 2019) はそのような例の 1 つで、ビット列内に出てくる n ビット数の出現確率を評価する手法である。本研究は、ビット列の乱数性の評価手法を日本語テキストの解析に応用することを目標としている。最初の試みとして、日本語テキストの各レジスタについて正規数性の指標を計算し、その分布の違いを分析した。文字エンコーディングでは、ASCII 制御文字のように各文

* Y.Takubo@niihama.kosen-ac.jp
† masayu-a@ninjal.ac.jp
‡ yamazaki@ninjal.ac.jp

字共通の固定ビット列を含む。従って、日本語テキストのビット表現は完全な乱数にはならない。しかしながら、乱数性の水準は言語やレジスタの情報を含むはずなので、正規数性を用いて文章を分類できるかも知れない。

本稿は以下のように構成される。第 2 章では正規数性の理論的定義を行う。第 3 章で本研究で用いた解析手法を説明し、第 4 章でその結果について議論する。最後に、第 5 章で本研究の結論と今後の展望について述べる。

2. 正規数性

有限長または無限長のビット列について乱数性を評価するため、様々な方法が開発されてきた。乱数性を定式化する最初の試みとして、Émile Borel は無限長のビット列に対する正規数性の指標を定義した。無限長のビット列が m ビット数によって構成されていて、各 m ビット数 (0, 1, 2, ..., $2^m - 1$) が等しい確率 (2^{-m}) で現れる場合に正規数 (Borel normal) と呼ぶ（図 1）。

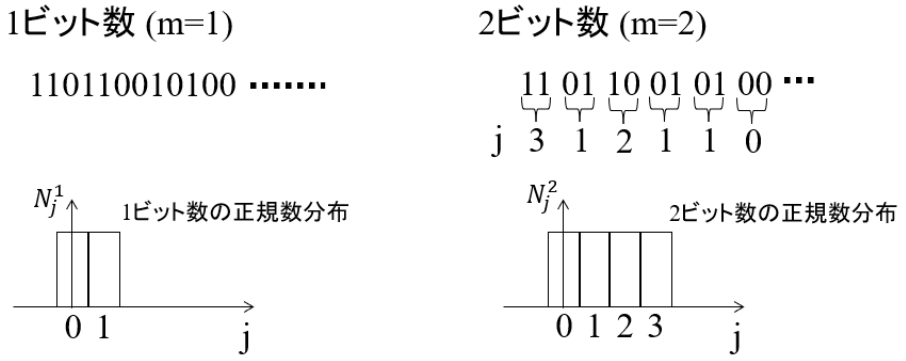

図 1　1 ビット数（左）と 2 ビット数（右）についての正規数分布の概念図。

正規数の定義は、不等式を用いて無限長のビット列から有限長のビット列に拡張できる (Claude 2002)。$B_m = \{0,1\}^m$ を m ビット数の集合とし、$N_j^m(x)$ を m ビット数の中の j 番目の数がビット列 x 中に出てくる回数とする。$|x|_m$ をビット列 x に含まれる m ビット数の数とし、$|x| = |x|_1$ とする。そして、以下の条件を満たすとき、ビット列 x は正規数であると定義する。

$$\psi_j^m(x) = \left| \frac{N_j^m(x)}{|x|_m} - 2^{-m} \right| log_2|x| \leq 1 \tag{1}$$

この不等式の適応範囲は、$l \leq m \leq log_2 log_2|x|$、$1 \leq j \leq (2^m - 1)$ である。本研究の場合、第 3 章で説明する通り、使用するサンプルの大きさが 96k ビット ($log_2 log_2|96000| = 4.05$) であることから、この不等式の適応範囲は $1 \leq m \leq 4$（すなわち、$0 \leq j \leq 15$）となる。全ての (m, j) のペアの出現確率が 2^{-m} である場合、ビット列 x は完全な正規数で $\psi_j^m(x)$ はゼロとなる。

正規数性は乱数についての最も直観的な特徴であるが、抜け穴も存在する。例えば、「0 1 00

01 10 11 000 001 011 100...」の Champernowne 列 (Champernowne 1933) は、m ビット数の全ての要素を同じ出現確率 (2^{-m}) で含んでいるが、明らかに乱数ではない。正規数性だけで乱数性を評価することはできないが、正規数性は乱数が満たすべき特徴の 1 つである。テキストの乱数性を評価する最初の試みとして、本研究では日本語テキストの正規数性について調査した。

3. 解析手法

本研究では、国語研究所の『現代日本語書き言葉均衡コーパス (BCCWJ)』のコアデータに収蔵されている日本語テキストを用いた。テキストは、書籍 (PB)、雑誌 (PM)、新聞 (PN)、白書 (OW)、Yahoo!知恵袋 (OC)、Yahoo!ブログ (OY) の 6 つのレジスタに分類した。元のテキストとの比較のため、各レジスタについて単語をランダムにシャッフルしたテキストも作成した。$\psi_j^m(x)$ の中心値と標準偏差を評価するため、各レジスタのテキストは 96k ビットを持つ 40 個から 150 個のサンプルに分割した。

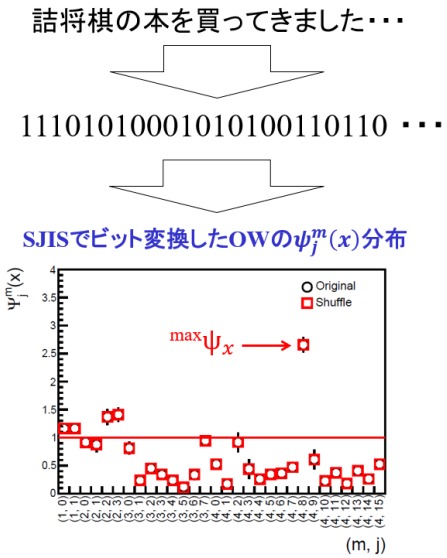

図 2　日本語テキストについて、正規数性の指標 $(^{\max}\psi_x)$ を計算する手順の概念図。下のプロットは、OW の元のテキストと文字をシャッフルしたテキストの $\psi_j^m(x)$ の分布。

日本語テキストについて $\psi_j^m(x)$ を計算する手順を図 2 に示す。UTF-8、SJIS、EUC の文字エンコーディングを用いて日本語テキストをビット列に変換し、$\psi_j^m(x)$ を計算する。α や é などの特殊記号が 16 ビットで表現されることを除いて、UTF-8 は日本語の 1 文字を 24 ビットで表現している。特殊記号の出現確率は 10^{-5} から 10^{-4} と非常に小さいので、テキストの一般的な乱数性を反映していない。そのため、本研究では特殊記号はサンプルから除去した。SJIS と EUC は全ての日本語文字を 16 ビットで表現している。それゆえ、サンプルを用意する上で特別な処理はしていない。3 種類の文字エンコーディングについて、本研究で使用した文字数とビット数を表 1 にまとめた。

図 2 に載せたプロットは、OW について各 (m, j) ペアの $\psi_j^m(x)$ の分布を表している。ここ

表 1　本研究で用いた、UTF-8、SJIS、EUC の文字エンコーディングにおける各レジスタの文字数 (N_{c}) とビット数 (N_{bit})。

	N_{c}	N_{bit} (UTF-8)	N_{bit} (SJIS, EUC)
PB	235k	8.86M	5.91M
PM	241k	9.35M	6.23M
PN	363k	13.5M	9.02M
OW	230k	8.42M	5.61M
OC	111k	4.26M	2.84M
OY	118k	4.54M	3.03M

で、m は 1 から 4(j は 1 から 24) の値を取る。同様のプロットを全てのレジスタについて作成した。本研究では、全ての (m, j) のペアの中で $\psi_j^m(x)$ が最大となる値 ($^{\max}\psi_x$) を異なるレジスタで比較した。

4.　解析結果

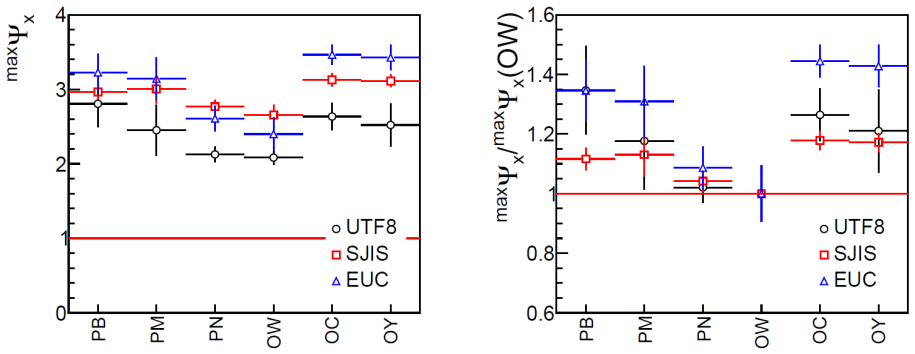

図 3　UTF-8、SJIS、EUC について、各レジスタの正規数性の指標の最大値 ($^{\max}\psi_x$) の分布（左）と、それを OW の値で規格化した分布（右）。

図 3(左) は UTF-8、SJIS、EUC でエンコーディングされた日本語テキストについて、全てのレジスタの $^{\max}\psi_x$ を表している。第 3 章で説明した通り、各レジスタのデータは 96k ビットのサンプルに分割している。各サンプルについて $^{\max}\psi_x$ を計算し、その結果から $^{\max}\psi_x$ の標準偏差を求め、それをプロットの誤差バーとして付与している。全てのレジスタの $^{\max}\psi_x$ は 1 よりも大きいため、日本語テキストは正規数ではないことを意味する。そして、それは N_j^m は一様に分布していないということでもある。テキストは文法や前後関係に依存して記述されるので、この結果は予想通りと言える。

図 3(右) は、レジスタ間の $^{\max}\psi_x$ の相対的な違いを見やすくするため、OW の値で $^{\max}\psi_x$ を規格化したものである。文字エンコーディングに依らず、各レジスタの $^{\max}\psi_x$ に系統的な傾向があることが分かる。OW はどの文字エンコーディングでも最も小さい $^{\max}\psi_x$ の値をもち、それは正規数性が最も高いことに対応する。一方、OC と OY は最も高い $^{\max}\psi_x$ の値をも

つため、正規数性が最も低いということになる。このようなレジスタによる正規性の違いは、文字の出現頻度の一様性の程度に起因すると考えられる。また、この結果から正規数性の程度によってレジスタを分類できる可能性があることが分かった。

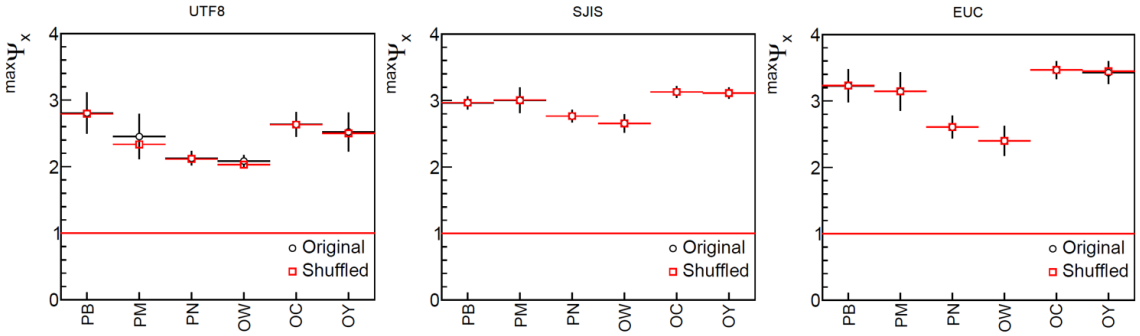

図 4　UTF-8（左）、SJIS（中）、EUC（左）について、各レジスタの ${}^{\max}\psi_x$ を元のテキストとシャッフル・テキストで比較した分布。

図 4 は元のテキストと単語をシャッフルしたテキストについて、${}^{\max}\psi_x$ の値を比較したものである。${}^{\max}\psi_x$ の分布は全てのレジスタについて、元のテキストとシャッフル・テキストで誤差の範囲で一致していることが分かる。今回の研究では $m = 4$ までの $\psi_j^m(x)$ を計算している。1 文字は UTF-8 で 24 ビット、SJIS と EUC で 16 ビットで表現されるため、1 文字内の正規数性を評価していることになる。従って、この手法では隣り合う文字間の相関を評価することはできず、1 文字に含まれるビット間の相関しか評価できない。以上の理由から、元のテキストとシャッフル・テキストで ${}^{\max}\psi_x$ の違いが現れない。前述のとおり、正規数性を評価するための不等式は $1 \leq m \leq log_2 log_2 |x|$ の範囲で適応可能である。UTF-8 で $2^{16777216}(m > 24)$ ビット以上、SJIS と EUC で $2^{65536}(m > 16)$ ビット以上のサンプルを用意できれば、隣り合う文字間の相関を評価できるようになり、元のテキストとシャッフル・テキストの違いが現れると予想される。しかし、このような大きなサイズのサンプルを作成することは現実的ではない。

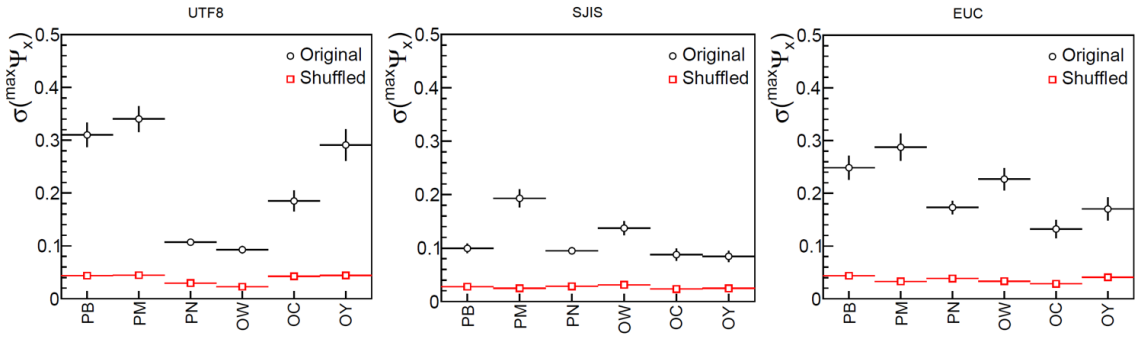

図 5　UTF-8（左）、SJIS（中）、EUC（左）について、各レジスタの正 ${}^{\max}\psi_x$ の標準偏差 $(\sigma({}^{\max}\psi_x))$ を元のテキストとシャッフル・テキストで比較した分布。

図 5 は ${}^{\max}\psi_x$ の標準偏差を元のテキストとシャッフル・テキストで比較したものである。元

のテキストでは、文字エンコーディングに依存しないレジスタの系統的な傾向は見られなかった。一方、シャッフル・テキストの標準偏差は元のテキストに比べて有意に減少している。これは単語をシャッフルすることによって文字の出現頻度の偏りが元のテキストよりも縮小したためだと考えられる。そのため、テキスト中の文字の偏り度合いの指標として標準偏差を使用することが可能である。それに加えて、シャッフル・テキストでは、標準偏差がレジスタによらず同程度の値になっている。このことから、シャッフル・テキストでは、どのレジスタも文字の多様性が同じ程度であることが分かる。

5. まとめ

本稿では、日本語テキストについて乱数が満たすべき性質の 1 つである正規数性について議論した。用いた手法としては、文字エンコーディングを用いて日本語テキストをビット列に変換し、正規数性の指標 ($^{\max}\psi_x$) を計算した。そして、レジスタ間での系統的な違いを評価した。その結果、文字エンコーディングによらず、各レジスタの $^{\max}\psi_x$ の相対的な違いに系統的な傾向が見られた。このことから、正規数性はレジスタを特徴づける指標として使用できる可能性があることが分かった。加えて、単語をシャッフルしたテキストの $^{\max}\psi_x$ の標準偏差は、元のテキストと比較して有意に減少していた。これはシャッフル・テキストでは、文字の出現頻度の偏りが一様になったためだと考えられる。そのため、テキスト中の文字の偏り度合いの指標として、標準偏差を使用することが可能である。

最後に、本研究の将来の展望を述べておく。ビット列の乱数性を評価する指標には正規数以外にも様々ある。その 1 つである Lempel-Ziv 複雑性は、文章のアルゴリズム複雑性の指標として使用されている (Lempel and Ziv 1976)。これらの新たな指標を応用することが次の目標となる。加えて、今回の手法では実現できなかったが、隣り合う文字間の相関を評価することも重要な課題だと考えている。

謝　辞

本研究は 2023 年度 IU-REAL 異分野融合・新分野創出プログラム　スタートアップ (IU-REAL23P03)、JSPS 科研費 (JP20H01906、JP23K17512)、国立国語研究所「共同利用型共同研究（C)」の助成を受けたものである。

文　献

Alastair A Abbott, Cristian S Claude, Michael J Dinneen, and Nan Huang (2019). “Experimentally probing the algorithmic randomness and incomputability of quantum randomness.” *Physica Scripta*, 94:4, p. 045103.

Cristian S Claude (2002). *Information and Randomness: An Algorithmic Perspective.*: Springer Berlin, Heidelberg.

D. G. Champernowne (1933). “The Construction of Decimals Normal in the Scale of Ten.” *Journal of the London Mathematical Society*, s1-8:4, pp. 254–260.

A. Lempel, and J. Ziv (1976). “On the complexity of finite sequences.” *IEEE Transactions*

on Information Theory, 22, p. 75.