英語テキストに含まれる単語の出現頻度に付随する不定性の評価

国立国語研究所論集 (NINJAL Research Papers) 30: 109–120 (2026)
ISSN 2186-1358

英語テキストに含まれる単語の出現頻度に付随する不定性の評価

田窪洋介[a]　　窪田　葵[b]

[a] 新居浜工業高等専門学校／高エネルギー加速器研究機構／国立国語研究所 共同研究員
[b] 新居浜工業高等専門学校 専攻科

要旨

自然言語の普遍的な性質の代表的なものとして Zipf 則がある。自然言語のデータを任意の関数でフィットし，Zipf 則の妥当性を定量的に評価する場合，その事前準備として正確な誤差を見積もっておく必要がある。本稿では Brown Corpus の英語データを用いて，英単語の出現数分布を作成し，その標準偏差から単語の出現数に付随する誤差（修正誤差）を評価した。その結果，修正誤差はポアソン誤差よりも有意に大きい値を持つことが分かった。また，英単語の出現順位と出現数のプロットのデータ点に修正誤差とポアソン誤差を付与し，Zipf 則の関数でフィットを行った。そして，データとフィットの一致度合いを χ^2 検定と KS（Kolmogorov-Smirnov）検定を用いて定量的に比較した。その結果，ポアソン誤差は英単語の出現数の不定性を過少評価していることが明らかになった。一方，フィット関数の形状はどちらの誤差を使用しても概ね同様であることが分かった *。

キーワード：Brown Corpus，統計的不定性，Zipf 則，χ^2 検定，KS 検定

1. はじめに

現在までに，自然言語に関する様々な普遍的性質が提唱されてきた。そのうちの 1 つに，Zipf 則が広く知られている（Zipf 1935）。Zipf 則は，テキスト中に出現する単語の出現回数は，出現順位に反比例して減少するという経験則である。すなわち，最も出現する単語が 2 番目に多い単語の約 2 倍，3 番目に多い単語の約 3 倍出現することを意味する。この経験則は英語や日本語のような言語の種類や，話し言葉や書き言葉のような文章の種類に依存せず，普遍的に成立することが知られている（田中 2021，田窪他 2024）。

このような統計的経験則の妥当性を定量的に検証するには，評価対象のデータを理論モデルの関数でフィットし，その一致度合いを統計的に評価する必要がある。一般的に，データのフィットには χ^2 フィットが用いられる。図 1 のようなプロットを関数 $f(x_i)$ でフィットする場合を考える。χ^2 は (1) 式のように定義され，$y(x_i)$ は x_i におけるデータの値，$f(x_i)$ は理論モデルの予想値，σ_i は $y(x_i)$ の誤差（不定性），n はデータ点の数を表す。

* 本研究を進めるにあたり，国立国語研究所の浅原正幸さんと山崎誠さんから有益な助言をいただいた。心より感謝申し上げる。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（JP23K17512），国立国語研究所・共同利用型共同研究（C）「統計的指標を用いた日本語テキストの数理的解明」（研究代表者：田窪洋介）の研究成果である。また，本稿の内容は 2025 年 3 月に開催された情報処理学会全国大会「英語テキストに含まれる単語の出現頻度に付随する不定性の評価」（窪田・田窪 2025）の発表成果も踏まえている。

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(y(x_i) - f(x_i)\right)^2}{\sigma_i^2} \qquad (1)$$

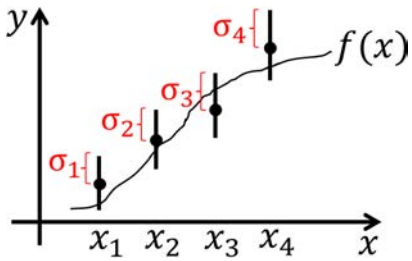

図 1　χ^2 フィットの概念図

χ^2 フィットは，χ^2 の値が最小化するようにフィット関数 $f(x_i)$ を決める手法である。χ^2 の式にデータの不定性 σ_i があるため，正しいフィットを行うためには，その事前準備として各データ点に付随する不定性を正確に評価しておく必要がある。

自然言語は人間の思考や，文法構造，文脈的な制約などの要因によるバイアスを受ける。したがって，自然言語は，ある事象が決まった確率で独立に起こる場合に用いられるポアソン分布には従わないと考えられる。そのため，単語の出現数に対する不定性としてポアソン分布を仮定した誤差（以下，「ポアソン誤差」と呼ぶ）を用いるのは不適切で，単語の出現数の標準偏差から正確な誤差（以下，「修正誤差」と呼ぶ）を評価する必要がある（窪田・田窪 2025，田窪他 2025）。本研究では，英語テキストについて修正誤差を評価した。そして，修正誤差をデータに適用することで，Zipf 則との一致度合いを定量的に評価した。

2. 評価手法

本研究では，評価対象の英文テキストに Brown Corpus を使用した。Brown Corpus は，1964 年に発表された世界最古のコーパスであり，今日に至るまで英語コーパス言語学における先駆けとして，その在り方を規定し続けている（後藤 1995，吉野 2000）。1961 年にアメリカ英語で書かれ，15 のレジスタ（例：報道，学術，フィクションなど。詳細は表 1 参照）に分類された，2000 語程度のテキスト・サンプルが 500 個収録されている。

データ解析の準備として，Python の自然言語処理ライブラリである NLTK（Natural Language Tool Kit）に含まれる WordNetLemmatizer を用いて，Brown Corpus の全単語に対してレンマ化（lemmatization）を実施した。レンマ化とは，単語の屈折形や派生形を基本形に変換する自然言語処理の手法であり，原形化（stemming）と異なり，文脈や品詞を考慮し，辞書的に正しい形に変換できる点が特徴である。具体的なレンマ化の変換例は次の通りである。

1. 動詞の活用形（例：“am”, “was”, “are”）は，全て原形（例：“be”）に統一される。
2. 名詞の複数形（例：“books”, “children”）は，全て単数形（例：“book”, “child”）に変換される。

3. 形容詞の変化形（例：“bigger”, “best”）は全て，原形（例：“big”, “good”）に変換される。

表 1　Brown Corpus のレジスタと単語数の内訳

レジスタ名	単語数	レジスタ名	単語数
Reportage	84792	Learned	154400
Editorial	52077	General Fiction	56323
Reviews	34034	Mystery and Detective Fiction	46636
Religion	32973	Science Fiction	11729
Skills and Hobbies	70034	Adventure and Western Fiction	56853
Popular Lore	92730	Romance and Love Story	56789
Belles Lettres, Biography, Memoirs, etc.	144774	Humor	17669
Miscellaneous	59120		

もしレンマ化を実施しない場合，同一の意味であるはずの “runs”，“ran”，“running” などが全て異なった単語として解釈される。レンマ化によって，このような意味的に等価であっても形態の異なる語が，すべて一意に統一されるため，実質的に語彙の重複が減少する。

レンマ化を行った後，Python コードを用いて，レジスタ別に各単語の出現数とその出現順位を計算した。そして，横軸に単語の出現順位，縦軸に出現数のプロットを作成し，Zipf 則の式（$y = \alpha + \beta / x$）でフィットを行った。プロットの作成やフィットは，欧州原子核研究機構（CERN）で開発された数値計算ソフトである ROOT を用いて作成した。ROOT は C++ と Python で用意されており，本研究では Python で用意されたもの（PyROOT）を使用した。ROOT の中には，数字の配列からグラフを作成したり，それをフィットするための関数が用意されている。

図 2 に「Reportage」の結果を示す。さらに，これと同様の手法で Brown Corpus に収録されている全てのレジスタに対してプロットとフィットを行った結果，全てのレジスタで一貫して単語の出現数は出現順位の冪乗に従うことが確認できた。これにより，Zipf 則が Brown Corpus に含まれる多様なジャンルの英文テキストに対して広く適用可能であることが示唆される。

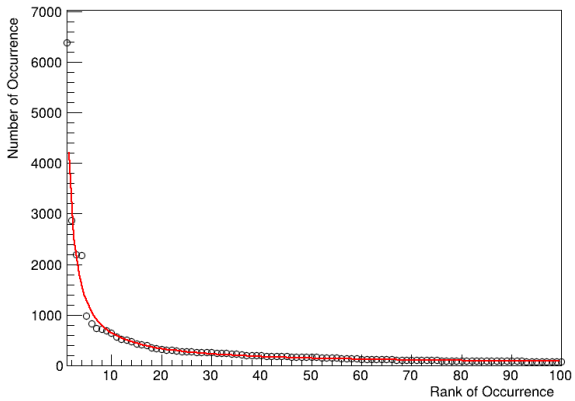

図 2　Zipf 則でのフィット（レジスタ：Reportage）

3. 単語の出現数に付随する誤差の調査

単語の出現数に付随する誤差を調査するために，Brown Corpus 内の全テキストを対象に，単語の出現回数のランキングを作成した。一例として，各テキスト・サンプルに現れる「the」（出現回数第 1 位）の出現回数のヒストグラムを図 3 に示す。横軸がテキスト・サンプル中の「the」の出現回数，縦軸がテキスト・サンプル数であり，「the」の出現回数が概ね 50 〜 250 回の範囲で変動していることが読み取れる。同様の手法で，出現回数が多い上位 40 語に対して，それぞれヒストグラムを作成した。ヒストグラムにおいて「Mean」が単語の出現数の平均値 $\langle N_{word} \rangle$ を，「Std Dev」が標準偏差 σ_{word} を示している。同様の手法で，出現回数上位 40 語までの単語に対して $\langle N_{word} \rangle$ と σ_{word} をそれぞれ取得した。

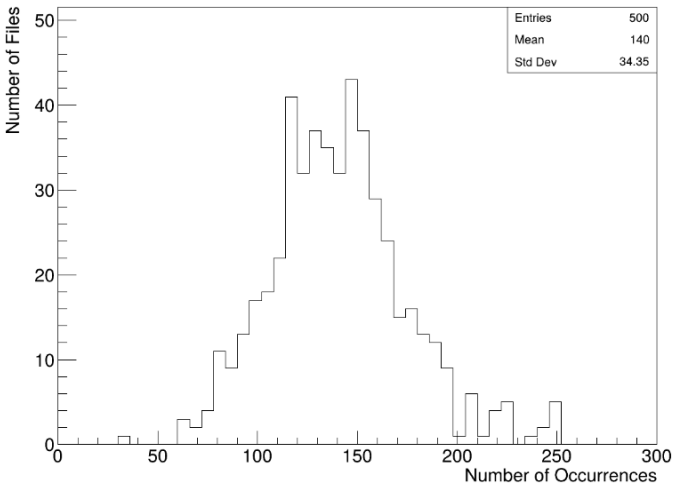

図 3　テキスト・サンプル中の単語の出現回数とテキスト・サンプル数の分布
（出現回数 1 位の単語「the」の場合）

次に，出現回数上位 40 語までの単語について，横軸が $\langle N_{word} \rangle$，縦軸が $\sigma_{word} / \langle N_{word} \rangle$ のプロットを作成した（図 4）。もし，単語の出現数がポアソン分布に従うとすると，ポアソン分布の標準偏差 σ_{word} は $\sqrt{\langle N_{word} \rangle}$ となる。そのため，相対誤差 RE（Relative Error）を $\sigma_{word} / \langle N_{word} \rangle$ のように定義すると，ポアソン誤差の RE は (1) 式のようになる（図 3 の点線）。

$$\mathrm{RE}\left(\mathrm{Poisson}\right) = 1/\sqrt{\langle N_{word} \rangle} \qquad (2)$$

一方，図 4 において実際の誤差はポアソン誤差よりも有意に大きいことが分かる。誤差の大きさを定量的に比較するために，ポアソン誤差の RE（(2) 式）を一般化すると (3) 式のようになる（ポアソン誤差の RE は α が零，β が零の場合）となる。この式を用いて，図 3 のデータ点のフィットを行った（図 3 の実線）。

$$\mathrm{RE}\left(\text{修正誤差}\right) = \alpha + \beta / \sqrt{\langle N_{word} \rangle} \qquad (3)$$

フィットの結果，最適解（修正誤差）として α = 0.06，β = 1.8 を得た。このことから，ポアソン誤差の RE は定数項（α）が零であるのに対し，修正誤差は 0.06（6%）であることが分かった。そのため，ポアソン誤差では単語の出現数回数が大きくなると RE は零に近づくのに対し，修正誤差では 6% の誤差が残る。また，$1/\sqrt{\langle N_{word} \rangle}$ の項（β）については，修正誤差はポアソン誤差に比べて 1.8 倍大きいという結果になった。

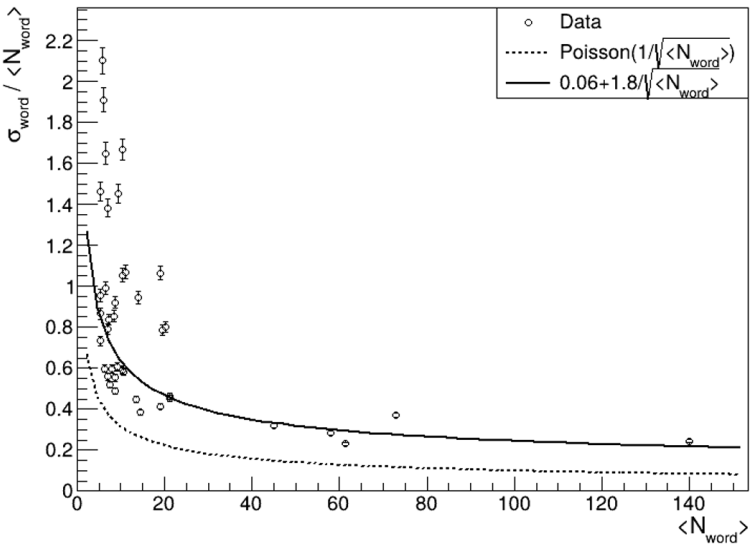

図 4　$\langle N_{word} \rangle$ と $\sigma_{word} / \langle N_{word} \rangle$ の相関

4. Zipf 則への応用

データ点に付与する誤差の違いによるフィットへの影響を評価するために，前章で導出した修正誤差を用いて Zipf 則（$y = \alpha + \beta / x$）のフィットを行った（図 5）。図 5 の横軸は単語の出現順位を，縦軸は単語の出現数を示している。図 5(a)(b) は「Humor」に対して，図 5(c)(d) は「Adventure and Western Fiction」に対してフィットを行った結果である。各データ点に対して図 5(a)(c) はポアソン誤差を，図 5(b)(d) は修正誤差を適用した。修正誤差を適用した場合（図 5(b)(d)）は，ポアソン誤差を適用した場合（図 5(a)(c)）に比べ，各点に付随する誤差バーが長くなっており，不定性が増大していることが定性的に確認できる。これは，修正誤差では定数項によって常に 6% 以上の不定性が残ることが理由であり，自然言語の本質的な揺らぎを正しく反映している。

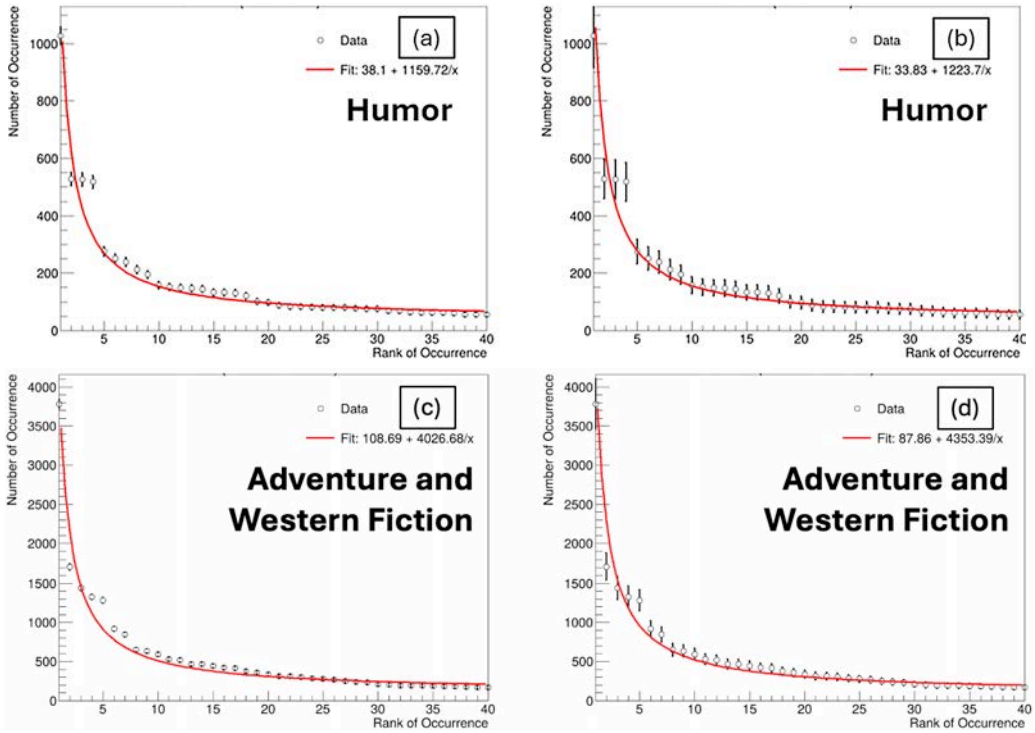

図 5　Zipf 則でフィットした結果（(a) (c)：ポアソン誤差，(b) (d)：修正誤差）

5. Zipf 則との一致度合いの確認

5.1 χ^2 検定

Zipf 則と実際の出現頻度分布との一致度合いを定量的に確認するために，χ^2 検定（Bevington and Robinson 2023）をレジスタ別に実施した（図 6）。これは統計モデルの適合度を評価する代表的な手法で，観測値と理論値の間にどの程度の差があるかを測るものである。$i = 1 \sim n$（n：順位の最大値）において，出現順位が i 番目の単語の出現数を N_i^{data}，Zipf 則の予測出現数を N_i^{Zipf}，各単語に付随する誤差を σ_i とすると，χ^2 は (4) 式のように表される。N_i^{Zipf} と N_i^{data} の値が近い場合，$\left(N_i^{data} - N_i^{Zipf}\right)^2 \cong 0$ であるため，$\chi^2 \cong 0$ となる。反対に，N_i^{data} と N_i^{Zipf} の値が離れていると χ^2 値が増大する。

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(N_i^{data} - N_i^{Zipf}\right)^2}{\sigma_i^2} \qquad (4)$$

この χ^2 値を用いることで，N_i^{data} と N_i^{Zipf} の間の乖離の程度を数値化し，(4) 式より得られる理想的な χ^2 分布と比較することで一致確率（一般的に「p 値」と呼ばれる）を計算できる。図 6 は各レジスタに対して χ^2 検定を実施した結果である。横軸は検定に使用する単語の最大値（第何位までの単語に対して χ^2 検定を実施するか），縦軸は χ^2 値を示す。全てのレジスタにおける χ^2 検定の結果は付録の図 9 を参照のこと。

図 6 は，ポアソン誤差では一致確率が概ね零になっており，Zipf 則が実際のデータと一致し

ていないことを表している。一方，修正誤差ではレジスタによって変化の様子に違いがあるものの，一部のレジスタを除いて一致確率が一定の値を持っている。これにより，ポアソン誤差は単語の出現数に対する不定性を過小評価しており，Zipf 則との一致度合いを評価するためには修正誤差を用いるべきであることを示すことができた。また，いずれの誤差を適用した場合でも一致確率が概ね零である「Mystery and Detective Fiction」と「Romance and Love Story」（付録：図 9）は，統計的にはデータは Zipf 則と異なっていると結論付けられる。

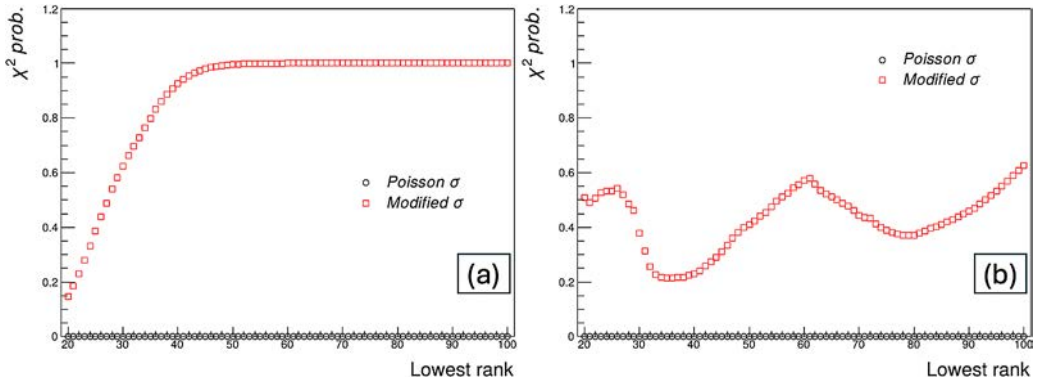

図 6　ポアソン誤差と修正誤差を用いた χ^2 検定の結果
（レジスタ：(a) Reportage，(b) Adventure and Western Fiction）

5.2 KS 検定

次に，Zipf 則と実際の出現頻度分布との一致度合いを，それぞれの分布の形状そのものに着目して評価するために，KS 検定（Kolmogorov 1933）を実施した。KS 検定は 2 つの母集団の確率分布の間に，統計的な差異があるかどうかを評価する検定手法である。KS 検定の概念図を図 7 に示す。KS 検定を用いた評価には，図 5 のプロットから得られる理論曲線を使用する。理論曲線の横軸は図 5 における各点の縦軸の値（N_i^{Zipf}）である。N_i^{Zipf} の昇順に $i = 1 \sim n$ とし，縦軸に累積確率 $f_i = i / n$ を付与する。検定対象の標本の場合は，上位 n 位までの N_i^{data} を横軸に，累積確率を縦軸に取り，プロットした。そして，両者の縦軸の値の差の最大値を D_{max} とすると，標本と理論分布が等しいという帰無仮説において，その差が $z = \sqrt{n}D_{max}$ よりも大きくなる確率は (5) 式のように表される。

$$\mathrm{P}\left(z = \sqrt{n}D_{max}\right) = 2\sum_{i=1}^{\infty}\left(-1\right)^{i-1} e^{-2i^2z^2} \qquad (5)$$

図 8 は，レジスタ別に KS 検定を実施した結果である。横軸は順位の最大値を，縦軸は一致確率を示す。全てのレジスタにおける KS 検定の結果は付録の図 10 を参照のこと。いずれのレジスタにおいても，ポアソン誤差と修正誤差の両方で一致確率が一定の値を持っている。また，ポアソン誤差と修正誤差の間で一致確率が完全に一致している点が多く見られた。この結果は，ポアソン誤差と修正誤差の間でフィットの形状が概ね一致していることを示唆しており，KS 検定が

分布の形状のみに着目する検定であることから，個々のデータ点の誤差が検定結果に対して直接的な影響を及ぼさないためであると考えられる。

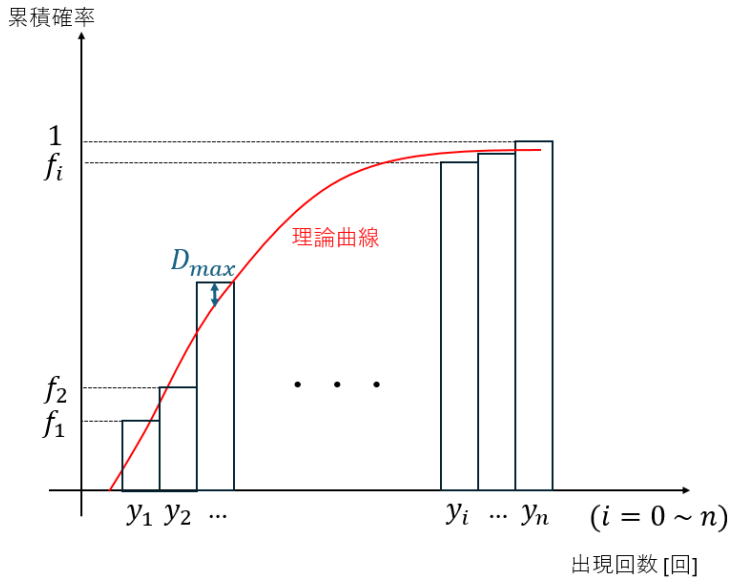

図 7　KS 検定の概念図

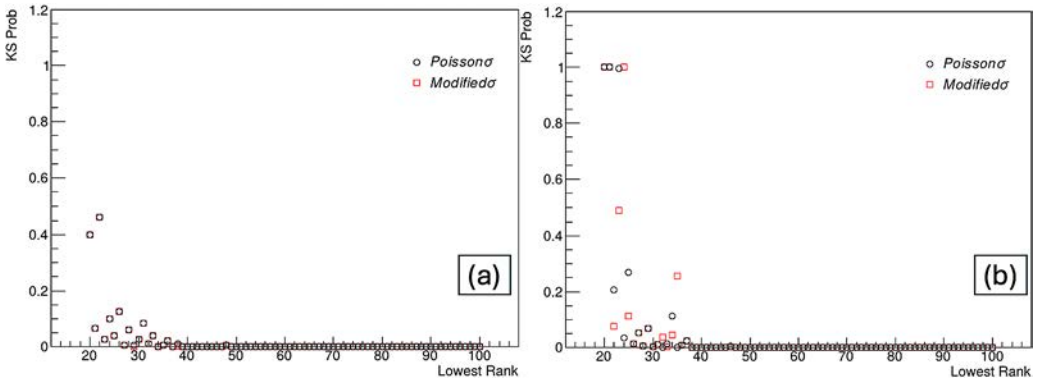

図 8　ポアソン誤差と修正誤差を用いた KS 検定の結果
（レジスタ：(a) Reportage，(b) Editorial）

6. まとめ

本研究では，Brown Corpus を対象として，英語テキストにおける修正誤差を評価した。その結果，修正誤差はポアソン誤差と比べて 6% の定数項があり，$1/\sqrt{\langle N_{word} \rangle}$ の項については 1.8 倍の違いがあることが分かった。このような差異は，自然言語特有のバイアスによるものであり，ポアソン誤差よりも正確に不定性を反映している。

修正誤差を使用する有意性を確認するため，データ点に修正誤差を付与し，χ^2 検定と KS 検定

によってデータと Zipf 則の一致度合いを判定した。χ^2 検定の結果，ポアソン誤差ではいずれのレジスタにおいても一致確率が概ね零となった。一方，修正誤差では一部のレジスタを除いて，一致確率は一定の値を持った。この結果から，ポアソン誤差では誤差を過少評価していることを明らかにした。KS 検定では，ポアソン誤差と修正誤差の両者で，χ^2 検定のような一致確率の大幅な変化はなかった。この結果は，ポアソン誤差と修正誤差を用いた場合でフィットの形状が概ね一致していることを示している。

参照文献

Bevington, Philip R. and D. Keith Robinson (2023) *Data reduction and error analysis for the physical sciences*. New York: McGraw-Hill.

Kolmogorov, Andrei Nikolaevich (1933) Sulla determinazione empirica di una legge di distribuzione. *Inst. Ital. Attuari, Giorn.* 4: 83–91.

Zipf, George Kingsley (1935) *The psychobiology of language*. Cambridge, Massachusetts: Houghton Mifflin Company.

窪田葵・田窪洋介（2025）「英語テキストに含まれる単語の出現頻度に付随する不定性の評価」『情報処理学会 第 87 回全国大会講演論文集』.

後藤斉（1995）「言語研究のためのデータとしてのコーパスの概念について―日本語のコーパス言語学のために―」『東北大学言語学論集』4: 71–87.

田窪洋介・浅原正幸・山崎誠（2024）「日本語テキストに含まれる単語の出現頻度に付随する不定性の評価」『計量国語学会』34(7): 529.

田窪洋介・浅原正幸・山崎誠（2025）「日本語話し言葉における形態素の出現数に対する統計的不定性の評価」『言語処理学会 第 31 回年次大会 発表論文集』59–64.

田中久美子（2021）『言語とフラクタル：使用の集積の中にある偶然と必然』東京：東京大学出版会.

吉野貴好（2000）「英語コーパス言語学の歴史的背景」『高崎経済大学論集』43(1): 97–107.

関連 Web サイト

CERN『ROOT: analyzing petabytes of data, scientifically. - ROOT』https://root.cern/（2025 年 5 月 5 日確認）

ICAME CORPUS MANUALS『Brown Corpus Manual』https://icame.info/icame_static/manuals/BROWN/INDEX.HTM（2025 年 8 月 5 日確認）

NLTK『NLTK :: Natural Language Toolkit』https://www.nltk.org/index.html（2025 年 5 月 6 日確認）

【付録】

Brown Corpus の全てのレジスタについて χ^2 検定を実施した結果を図 9 に，KS 検定を実施した結果を図 10 に示す。

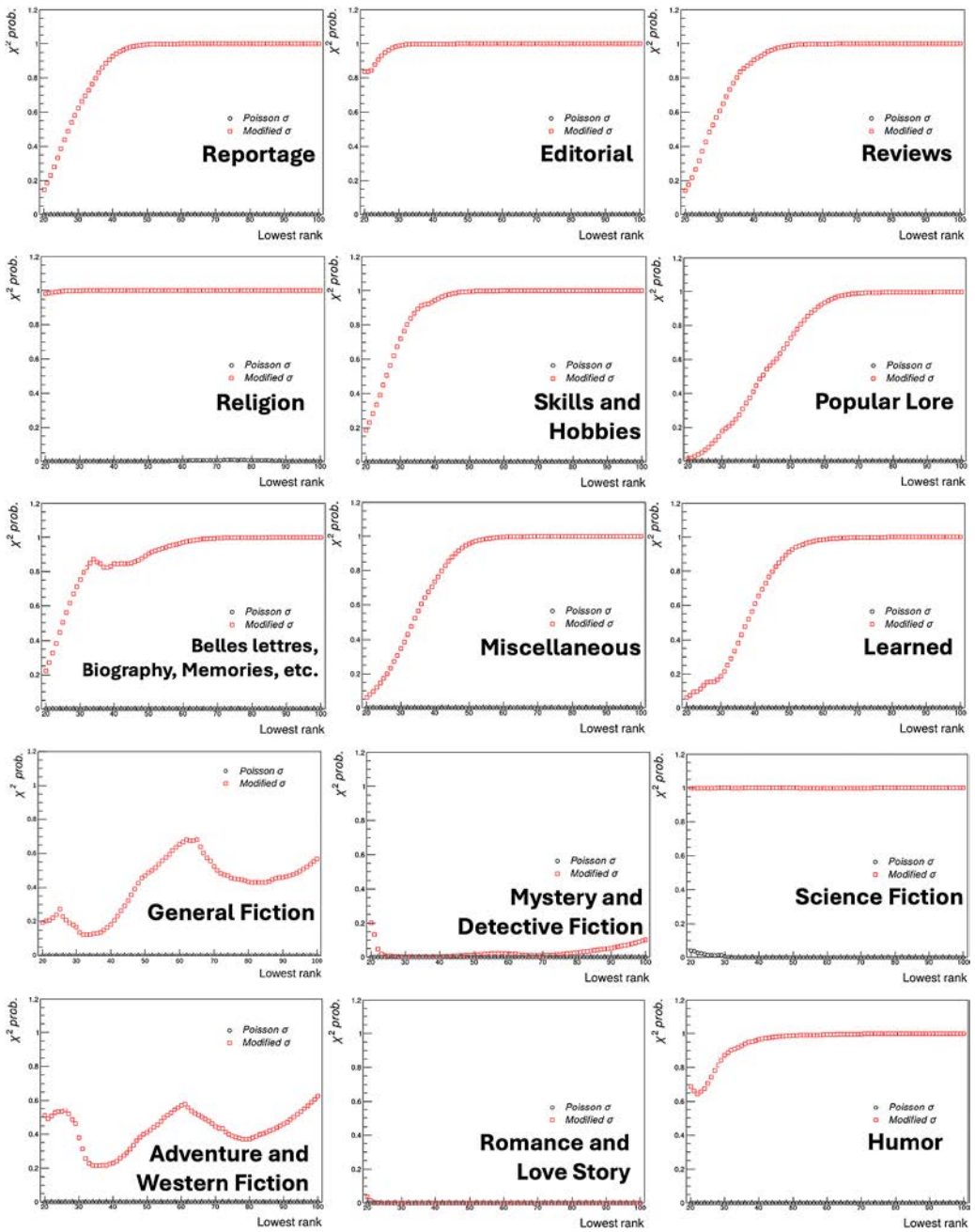

図 9　χ^2 検定の実施結果

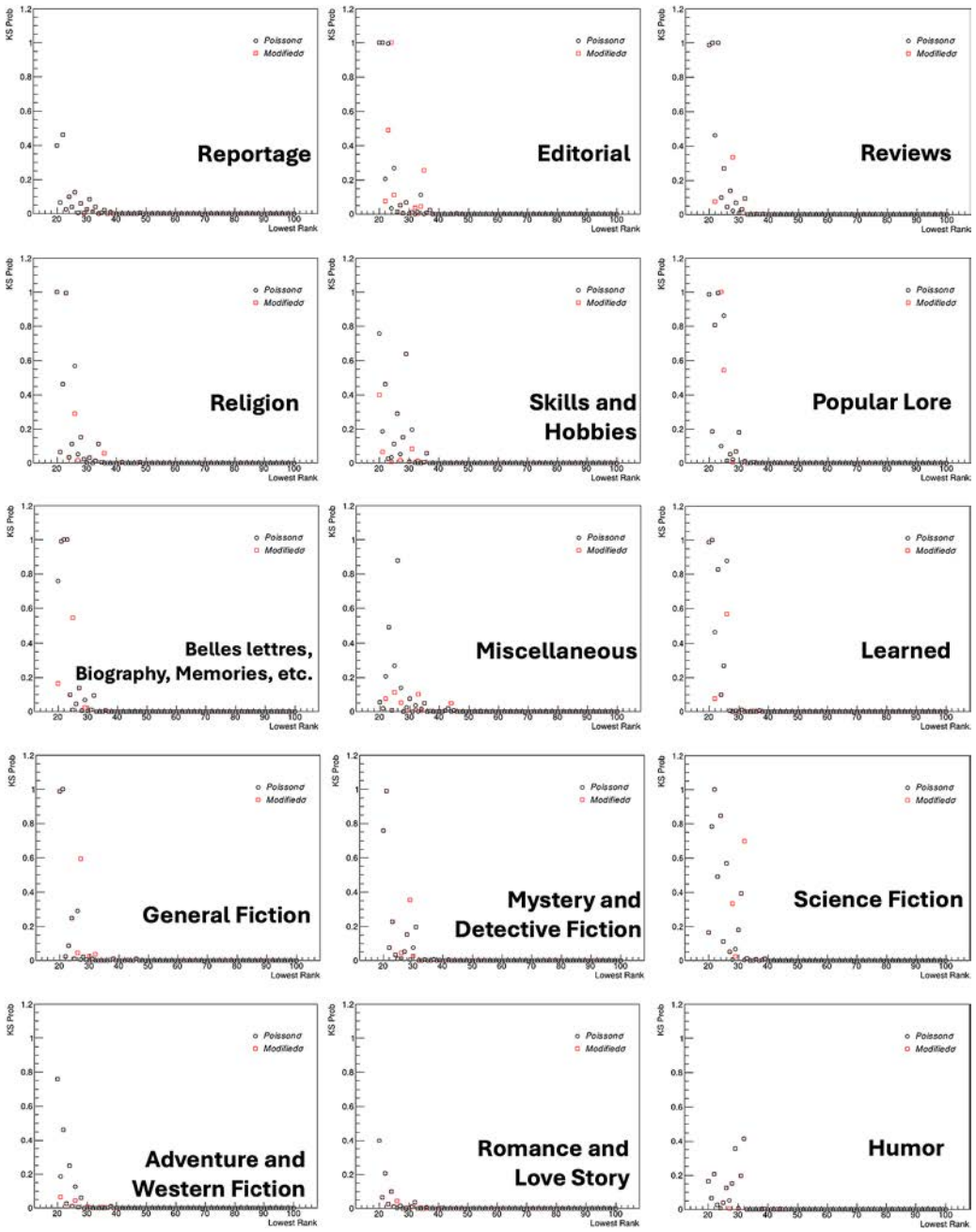

図 10　KS 検定の実施結果

Evaluating Uncertainty on Word Frequencies in English Texts

TAKUBO Yosuke[a]　　KUBOTA Aoi[b]

[a]National Institute of Technology, Niihama College /
High Energy Accelerator Research Organization / Project Collaborator, NINJAL
[b]Advanced Engineering Course Student, National Institute of Technology, Niihama College

Abstract
Zipf's law is one of the well-known universal characteristics of natural language. To evaluate the validity of Zipf's law quantitatively, it is necessary to evaluate the statistical errors of the word frequencies in a text (referred to as "true error"). In this study, true errors in English texts were examined using standard deviations of the word frequency distributions obtained from the Brown Corpus. It was found that the true error is significantly larger than the Poisson error. We performed fits of word frequencies as a function of the frequency rank using Zipf's law, assigning either the true or Poisson errors to the data points. We then applied the χ^2 and Kolmogorov-Smirnov tests to compare consistency between the data and Zipf's law. The results show that the error's size was underestimated when using Poisson errors. In addition, the overall shape of the fitted function was similar, regardless of the type of error used.

Keywords: Brown Corpus, Zipf's law, statistical uncertainty, χ^2 test, KS test