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

田窪洋介 [1,2]　浅原正幸 [3,4]　山崎誠 [3]
[1] 新居浜工業高等専門学校 電気情報工学科
[2] 高エネルギー加速器研究機構 素粒子原子核研究所
[3] 国立国語研究所　[4] 総合研究大学院大学
Yosuke.Takubo@cern.ch　{masayu-a,yamazaki}@ninjal.ac.jp

概要

自然言語と言語モデルの一致度合いを定量的に評価するためには，事前準備として自然言語のデータ点に付随する統計的不定性を正確に評価しておくことが前提となる．本研究では，日本語話し言葉における形態素の出現数に対する不定性（修正誤差）を評価した．そして，修正誤差とポアソン誤差をデータに付与した場合の Zipf 則との一致度合いについて，χ^2 検定と KS (Kolmogorov-Smirnov) 検定で定量的に比較した．

1　はじめに

自然言語のもつ普遍的な性質は，さまざまな言語モデルによって検証されている．自然言語のデータと言語モデルを詳細に比較するためには，テキスト中に現れる形態素の出現頻度の統計的ゆらぎなど，自然言語のデータに付随する統計的不定性を正確に見積もっておく必要がある．一方で，自然言語は人の思考や文法によるバイアスを含むため，形態素の出現数はポアソン分布には従わないことが予想され，その不定性の評価は単純ではない．

自然言語の普遍的な性質の代表的なものとして，Zipf 則 [1] がある．Zipf 則はテキスト中に出てくる単語（本研究では「形態素」を用いる）の出現数が出現順位の冪乗に比例するという法則である．自然言語のデータと Zipf 則の関数をフィットする際，形態素の出現数に付与する誤差によってフィット結果が大きく変わってくる．また，誤差を正しく付与しなければ，χ^2 検定などを用いてデータとフィットの一致度合いを定量的に評価する際に，正しい一致確率が得られない．以上の理由から，形態素の出現数に対する不定性を正確に評価しておく必要がある．

表 1　本研究で使用した CSJ の各レジスタに含まれる形態素数のまとめ

レジスタ	形態素数
学会講演 (A)	1,805,532
模擬講演 (S)	1,834,028
その他 (M)	151,881
対話音声 (D)	60,297
朗読音声 (R)	99,969

著者らは先行研究として，国語研究所の『現代日本語書き言葉均衡コーパス』(BCCWJ) に収蔵されている日本語テキストを用いて，書き言葉に含まれる形態素の出現数に対する統計的不定性について研究を行った [2]．本研究では，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（ここでは「修正誤差」と呼ぶ）を評価した．そして，修正誤差を用いてデータと Zipf 則をフィットし，一致度合いを χ^2 検定と KS 検定を用いて定評的に評価した．

2　評価方法

日本語話し言葉のデータとして，国語研究所の『日本語話し言葉コーパス』(CSJ) に収蔵されている「MORPH/SDB/noncore」のテキストを使用した．CSJ の話し言葉は，学会講演 (A)，模擬講演 (S)，その他 (M)，対話音声 (D)，朗読音声 (R) のレジスタに分類されている．表 1 に本研究で使用した各レジスタの形態素数をまとめた．

日本語話し言葉における修正誤差を評価するため，各形態素について出現数 (N_{word}) の標準偏差 (σ_{word}) を調査した．σ_{word} の評価には，形態素数の多い A と S のみを使用した．各レジスタの話し言葉を 1 万形態素のサンプルに分割し，それぞれのサンプルに含まれる各形態素の N_{word} のヒストグラムを作成した．比較のために，テキスト中の形態素を

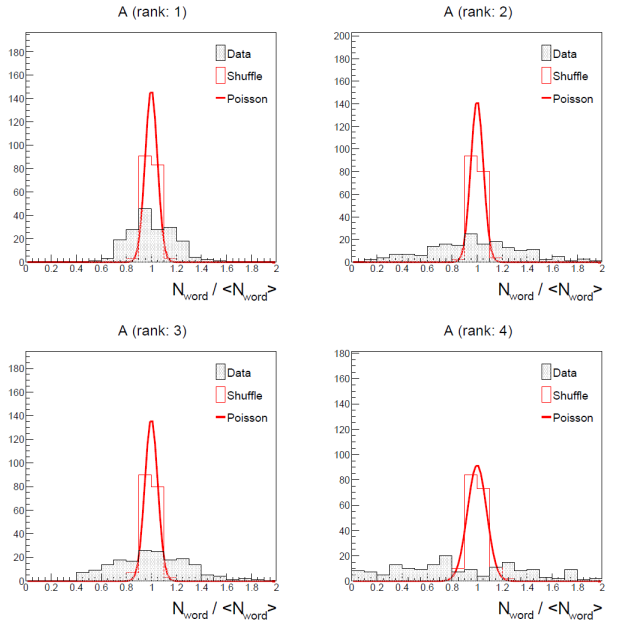

図 1　A における出現順位が 4 位までの形態素について，出現数 (N_{word}) をその平均値 ($\langle N_{\mathrm{word}}\rangle$) で規格化した分布．データとシャッフル・テキストの分布と，ポアソン分布を載せている．

シャッフルしたサンプル（シャッフル・テキスト）も作成した．シャッフル・テキストでは，サンプル間の形態素の出現数の偏りがなくなり，ポアソン分布に近づくと期待される．これらのヒストグラムを用いて各形態素の出現数の平均値 ($\langle N_{\mathrm{word}}\rangle$) と σ_{word} を求め，それらの相関を評価した．

データと Zipf 則との比較を行うために，話し言葉に出てくる形態素の出現順位と出現数を知る必要がある．これは，修正誤差の評価に使用した同じデータを用いて，各レジスタ（A, S, M, D, R）について形態素の出現順位と出現数のテーブルを作成した．ただし，Zipf 則に従いにくい助詞，助動詞，記号はテーブルから除いた．そして，形態素の出現数にポアソン誤差と修正誤差を付与し，データと Zipf 則のフィット結果の違いを比較した．

3　統計的不定性の評価

日本語話し言葉における形態素の出現数 N_{word} とその標準偏差 σ_{word} の関係を評価する．もし，N_{word} がポアソン分布に従うとすると，σ_{word} は $\sqrt{\langle N_{\mathrm{word}}\rangle}$ になるため，$\frac{\sigma_{\mathrm{word}}}{\langle N_{\mathrm{word}}\rangle}$ は $\frac{1}{\sqrt{\langle N_{\mathrm{word}}\rangle}}$ となる．シャッフル・テキストは，ポアソン分布のように振舞うと予想されるので，この式に従うと期待される．一方，データがどの程度ポアソン分布と異なっているかが，本章の主眼となる．

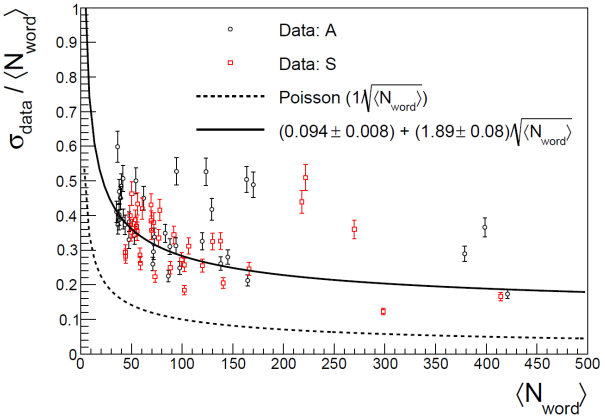

図 2　日本語話し言葉における形態素の出現数 ($\langle N_{\mathrm{word}}\rangle$) と，誤差 ($\sigma_{\mathrm{word}}$) を $\langle N_{\mathrm{word}}\rangle$ で規格化した値の相関．点線がポアソン分布の場合で，実線が本研究の修正誤差の最適解 $0.094 + \frac{1.9}{\sqrt{\langle N_{\mathrm{word}}\rangle}}$．

図 1 は，A について出現順位が 4 位までの形態素について出現数 (N_{word}) を平均値 ($\langle N_{\mathrm{word}}\rangle$) で規格化し，ヒストグラムにしたものである．元のテキストはポアソン分布に比べて分布の幅が広がっており，有意に統計的不定性が大きくなっていることが分かる．一方，シャッフル・テキストはポアソン分布とよい一致を示している．

図 2 は A と S について，横軸に N_{word} を，縦軸に σ_{word} と $\langle N_{\mathrm{word}}\rangle$ の比 $\frac{\sigma_{\mathrm{word}}}{\langle N_{\mathrm{word}}\rangle}$ を取ったものである．A と S ともに，出現順位 36 位までの結果を表示している．図の点線はポアソン分布の場合を示している．図 2 によると，どの $\langle N_{\mathrm{word}}\rangle$ についても標準偏差がポアソン分布よりも大きくなっている．また，A と S によって $\frac{\sigma_{\mathrm{word}}}{\langle N_{\mathrm{word}}\rangle}$ の値の偏りがないことも分かる．

図 2 から出現数に付随する不定性は A と S に依存しないことが分かった．そのため，この 2 つのレジスタのデータを合わせ，ポアソン分布の誤差 $\frac{1}{\sqrt{\langle N_{\mathrm{word}}\rangle}}$ と比較するために，$\alpha + \frac{\beta}{\sqrt{\langle N_{\mathrm{word}}\rangle}}$ の関数でフィットを行った．そして，最適解として $(0.094 \pm 0.008) + \frac{(1.89\pm0.08)}{\sqrt{\langle N_{\mathrm{word}}\rangle}}$ を得た．この結果から，ポアソン分布では N_{word} が無限大に近づけばゼロになるが，修正誤差では 9.4%の誤差が残る．また，修正誤差では $\frac{1}{\sqrt{\langle N_{\mathrm{word}}\rangle}}$ の項につく係数は，ポアソン誤差に比べて 1.9 倍大きい．

先行研究において，日本語書き言葉の修正誤差の最適解として $(0.102 \pm 0.003) + \frac{(1.64\pm0.03)}{\sqrt{\langle N_{\mathrm{word}}\rangle}}$ を得ている．従って，日本語書き言葉と話し言葉における形態素の出現数の不定性は，定数項については誤差の範囲で一致している．この結果は，出現数が多い形

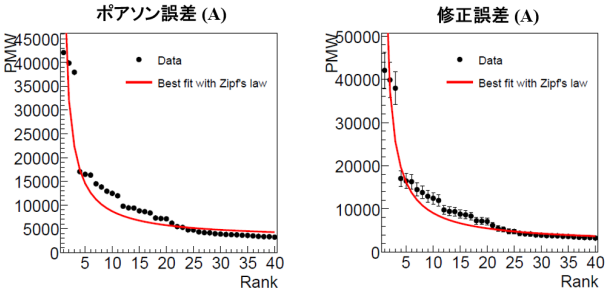

図 3　A における形態素の出現順位と PMW の相関で，左はデータ点にポアソン誤差を右は修正誤差を付与した場合．実線は Zipf 則の関数でフィットした結果．

態素の不定性は，書き言葉と話し言葉によらない普遍的な性質である可能性を示唆している．一方，$\frac{1}{\sqrt{\langle N_{\mathrm{word}}\rangle}}$ の係数については，話し言葉と書き言葉で約 3σ の乖離がある．

4　Zipf 則への応用

図 3 は，A における形態素の出現順位と PMW(Per Million Words) の相関である．その図に，データ点にポアソン誤差と修正誤差を付与したところ，修正誤差では出現順位が大きいデータ点で，誤差バーが大きくなっていることが分かる．これは，ポアソン誤差では形態素の出現数が大きくなればそれに伴って誤差が平方根で減少するが，修正誤差では常に 9%以上の不定性が残るためである．

次に，Zipf 則との一致度合いを定量的に評価するために χ^2 検定 [3] と KS 検定 [4] を行った．出現順位が i 番目の形態素の出現数を N_i^{data}，Zipf 則の予想出現数を N_i^{Zipf} としたとき，χ^2 は次のように定義される．

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

ここで，σ_i は N_i^{data} の標準偏差，n はデータ点の数になる．もし，N_i^{data} が標準偏差の不定性の範囲で N_i^{Zipf} の予想と一致していれば，χ^2 は平均値 n の分布となる．また，N_i^{data} が N_i^{Zipf} と完全に一致していれば χ^2 はゼロとなる．従って，χ^2 が大きい値を持つほど，Zipf 則からのずれが大きいことになる．一方で，χ^2 がゼロに近すぎると，データの不定性よりも Zipf 則と一致していることになる．その場合は，標準偏差を過大評価しているか，データにバイアスがあるかも知れない．

データがポアソン分布に従っていると仮定し，データの χ^2 が統計的ゆらぎによって起こる確率 (p 値と呼ばれる) を，データとモデルの一致度合いの評価に使用する．例えば，p 値が 0.05 ならば，データの χ^2 値がポアソン分布の統計的ゆらぎによって得られる確率は 5%ということになる．

一方，KS 検定は分布の形のみに注目する検定手法である．図 4 に KS 検定の概念図を示した．まず，N_i^{data} を x 軸上に並べ，小さい方から $y_j (j = 1, 2, ..., n)$ と定義し直す．そして，y 軸の値として，各 y_j に対して累積確率 $f(y_j) = j/n$ を与える．同様に，Zipf 則の理論曲線を作成し，$f(y_j)$ についてデータとのずれの最大距離を D_{max} とする．そうすると，データと理論分布が等しいという仮定（無帰仮説）において，その差が $z = \sqrt{n} D_{\mathrm{max}}$ より大きくなる確率は以下で計算できる．

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

従って，$P(z)$ が小さければデータとモデルの差が D_{max} よりも大きくなる確率が低いことに対応し，分布の一致度合いが悪いことを意味する．

今回は各データにポアソン誤差と修正誤差を付与し，各レジスタについて最下位の出現順位が 20 から 100 までの範囲で Zipf 則の関数 $\alpha + \frac{\beta}{x}$ でフィットを行った．図 3 は最下位の出現順位が 40 についてのフィット結果になる．最後に，χ^2 検定と KS 検定を用いてフィット結果を定量的に比較した．

図 5 は，各レジスタについて χ^2 検定を行った結果である．ポアソン誤差では，誤差を過少評価しているために，全てのレジスタで p 値がゼロになった．一方，修正誤差では各データ点の誤差が増加したため，M と D では p 値が評価できるようになっている．しかし，A，S，R では修正誤差を用いても p 値がゼロになっている．従って，これらのレジスタでは統計的には Zipf 則との一致は認められないという結論になる．また，M と D についても，フィットに用いた出現順位の範囲によって p 値が変化しており，考慮する出現順位の範囲によって Zipf 則との一致度合いが異なることを示している．

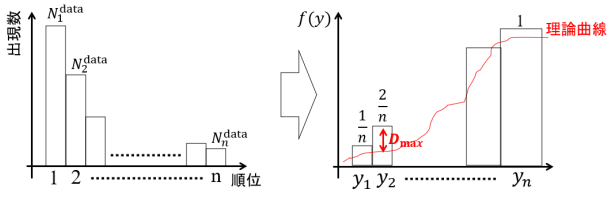

図 4　KS 検定の概念図．

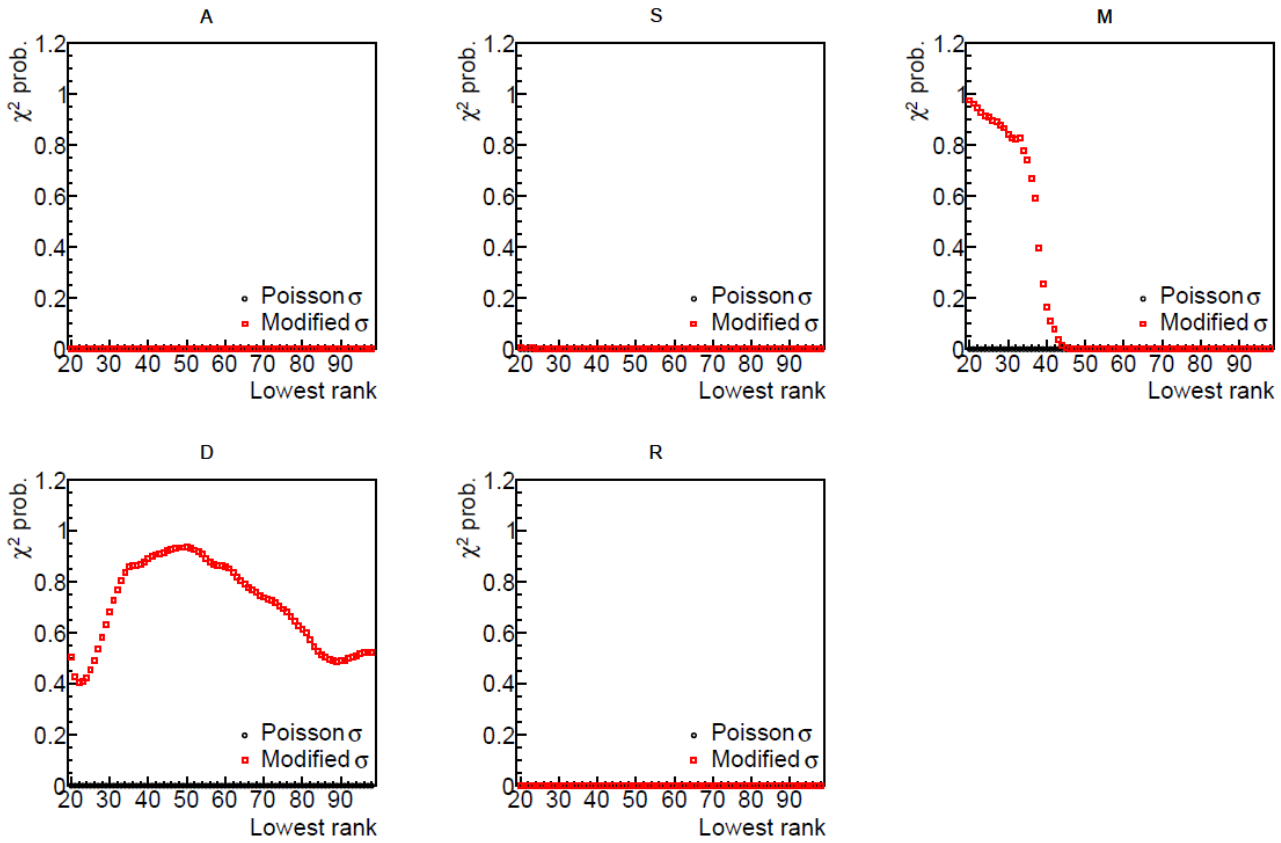

図 5　ポアソン誤差と修正誤差を用いた場合の χ^2 検定の結果．横軸は Zipf 則のフィットに用いた最下位の形態素の出現順位，縦軸は χ^2 検定で得られた一致確率 (p 値).

KS 検定についての結果を付録 A の図 6 に示す．ポアソン誤差を用いた場合でも，ゼロ以上の KS 確率を持っている．KS 検定ではデータと理論式の分布の形のみを比較しているため，データ点の誤差が KS 確率に直接影響しないためである．修正誤差を使用した方が，ポアソン誤差よりも KS 確率が高くなっている．従って，修正誤差を使用した方が，Zipf 則の形状により適合するようにフィットされていることになる．

5　まとめ

本研究では，日本語話し言葉に出てくる形態素の出現数に対する統計的不定性を評価した．そして，ポアソン誤差は不定性が過少評価となっていることが分かった．この結果は，日本語書き言葉についての先行研究と同様の結論であった．また，出現数の不定性における定数項は，日本語書き言葉と話し言葉で一致していた．一方，$\frac{1}{\sqrt{\langle N_{\mathrm{word}} \rangle}}$ の係数については，話し言葉の方が書き言葉よりも係数が大きく，約 3σ の違いがあることも分かった．

修正誤差の実用例として，データと Zipf 則の一致度合いを χ^2 検定と KS 検定を用いて，ポアソン誤差の結果と比較した．χ^2 検定では，ポアソン誤差では p 値がゼロとなったのに対して，修正誤差では M と D についてある一定の値を持った．また，KS 検定では，全てのレジスタにおいて，修正誤差を用いることで Zipf 則とのフィットが改善することを示せた．以上の結果から，自然言語におけるデータとモデルの一致度合いを評価する上で，修正誤差の有用性を示すことができた．

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 (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.

参考文献

[1] George Kingsley Zipf. **The Psychobiology of Language**. Houghton Mifflin Company, 1935.

[2] 山崎誠田窪洋介. 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, 2024. 計量国語学会第 68 回　大会発表予稿集.

[3] Philip R. Bevington and D. Keith Robinson. **Data Reduction and Error Analysis for the Physical Sciences**. McGraw-Hill Higher Education, 2003.

[4] Andrey Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. **G. Ist. Ital. Attuari**, Vol. 4, pp. 83–91, 1933.

A　KS 検定の結果

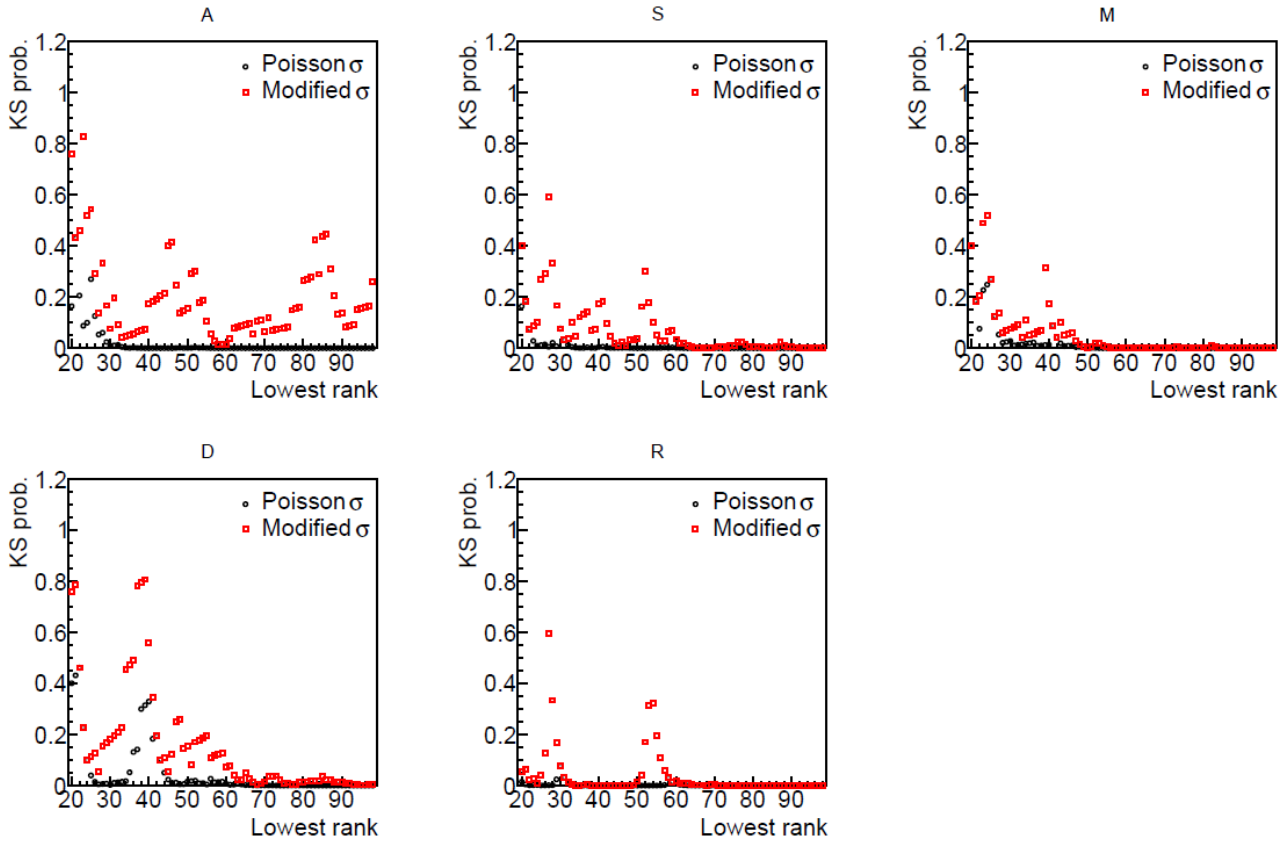

図 6　ポアソン誤差と修正誤差を用いた場合の KS 検定の結果．横軸は Zipf 則のフィットに用いた最下位の形態素の出現順位，縦軸は KS 検定で得られた一致確率．