語彙多様性指標の可視化と単回帰分析によるTTRの補正

著者	今田 水穂
雑誌名	言語資源活用ワークショップ発表論文集
巻	3
ページ	519-530
発行年	2018
URL	http://doi.org/10.15084/00001686

語彙多様性指標の可視化と単回帰分析による TTR の補正

今田 水穂 (文部科学省) *

Visualization of Lexical Diversity Indices and Adjustment of TTR by Single Regression Analysis

Mizuho Imada (MEXT)

要旨

語彙多様性を評価する既存の指標には、延べ語数 N と異なり語数 V(N) を入力とするもの、単語別の頻度を入力とするもの、単語列を入力とするものなどがある。本発表では、これらの指標の特徴を整理し、「現代日本語書き言葉均衡コーパス」(BCCWJ) を使用して指標値の分布を可視化する。N と V(N) を入力とする指標のいくつかは、両者の間に冪乗則 $V(N) = aN^b$ を仮定している。TTR は $b = 1$、R は $b = 0.5$ として a を指標値として利用するが、1 では大きすぎ、0.5 では小さすぎる。そこで V(N) と N の対数を単回帰分析して b の最適値を推定し、TTR を補正することを考える。実際には冪乗則は成立しないため、この補正は近似的だが、比較的簡単によい補正を得ることができる。この補正値を他の指標と比較し、テキストサイズが指標値の平均やばらつきに及ぼす影響を評価する。また、BCCWJ の 12 のサブコーパスについて b の値を推定し、一覧で示す。

1. はじめに

テキストの語彙多様性を評価する指標として、タイプ-トークン比 (TTR) が知られる。しかし TTR は延べ語数の影響を受け、テキストが長くなるほど指標値が小さくなる特徴がある。これを補正した指標の 1 つに R があるが、R は補正が強すぎて TTR とは逆に指標値が大きくなる特徴がある。他に多くの指標が提案されているが、計算の容易さ、テキスト長による平均やばらつきの変動など指標ごとに特徴がある。様々な観点から各指標の有用性を検討している最近の研究としては、木村・田中 (2010) や鄭・金 (2018) がある。

本稿では既存の主要な指標を計算に使用する入力データの形態から 3 種類に分類し、それらの特徴を検討する。また、延べ語数 N と異なり語数 $V(N)$ の間にべき乗則 $V(N) = aN^b$ が成り立つという仮定に基づき、両者の対数を単回帰分析して残差 ε を指標値として利用することで TTR を補正する方法を試みる。また「現代日本語書き言葉均衡コーパス」(BCCWJ) を用いて各指標の値を実際に計算し、その分布をグラフで確認するとともに、指標値と延べ語数を単回帰分析することでテキストの長さが各指標の平均や分散に及ぼす影響を評価し、単回帰分析による補正法が平均や分散の変動を受けにくいことを確認する。この補正値は延べ語数と異なり語数から簡単な式で計算することができるが、あらかじめ単回帰分析を行ってサンプル全体の

* imadamizuho.ac@google.com

分布の傾きを示すパラメータ値 b を計算しておく必要がある。そこでレジスタや語の集計単位を様々に変えて同補正法を試み、条件ごとのパラメータ値の一覧を示す。

2. 語彙多様性指標

語彙多様性を表す指標は、(1) 異なり語数と延べ語数を入力とするもの、(2) 単語別の頻度を入力とするもの、(3) 単語列を入力とするものがある。

表 1　データの種類と手法

種類	例	手法
(1) 総語数	{異なり語数: 11966, 延べ語数: 209326}	TTR, R, C, etc.
(2) 単語別頻度	{吾輩: 481, は: 6501, 猫: 237, . . .}	HD-D, Yule's K, etc.
(3) 単語列	{吾輩, は, 猫, である, . . .}	MSTTR, MATTR, etc.

(1) は異なり語数と延べ語数の関係を表す式を仮定して、その式の係数を指標として使うもので、TTR、R(Guiraud 1954)、C(Herdan 1960)、S(Somers 1966)、a^2(Maas 1972)、Uber(Dugast 1979) などがある。(2) は部分集合の特徴量を反復実測の代わりに各語の頻度に基づく確率計算で推定する HD-D(McCarthy and Jarvis 2007)、Yule's K、Simpson's λ、Shannon's H' や、頻度分布の形状を特徴量として利用するものなどがある。(3) は単語列の部分集合の特徴量 (n 語あたりの異なり語数など) を反復実測によって推定するもので、MSTTR、MATTR、voc-d(Mckee et al. 2000)、MTLD、MTLDMA などがある。入力データの情報量は (3) が最も大きく (1) が最も小さいが、その分、計算量も (3) が最も大きく (1) が最も小さい。

(1) のタイプの指標の代表的なものを以下に示す。N は延べ語数、V(N) は N 語あたりの異なり語数である。TTR は異なり語数を延べ語数で割ったものである。R と CTTR は平方根を用いた TTR の変種だが、CTTR は R の定数倍なので実質的に同一の指標である。それ以外のものは対数を用いた TTR の変種である。C、S、k は V(N) および N の対数、あるいは対数の対数を使用する。a^2 は $\log TTR = -a^2(\log N)^2$ と変形することができ、TTR と N の関係式と見なすことができる。Uber は a^2 の逆数なので、実質的に a^2 と同一の指標である。[(1)]

$$TTR = \frac{V(N)}{N} \qquad CTTR = \frac{V(N)}{\sqrt{2N}} \qquad k = \frac{\log V(N)}{\log(\log N)}$$

$$C = \frac{\log V(N)}{\log N} \qquad a^2 = \frac{\log N - \log V(N)}{(\log N)^2}$$

$$R = \frac{V(N)}{\sqrt{N}} \qquad S = \frac{\log(\log V(N))}{\log(\log(N))} \qquad Uber = \frac{(\log N)^2}{\log N - \log V(N)}$$

(1) 鄭・金 (2018) は他に $LN = \frac{1-V(N)^2}{V(N)^2 \log N}$ を挙げている。LN は $LN = -\frac{V(N)^2-1}{V(N)^2}\frac{1}{\log N}$ と変形でき、V(N) がごく小さい値のとき以外は $-\frac{1}{\log N}$ と近似する値になる。本稿では LN は扱わない。

次に、(2) のタイプの指標の代表的なものを以下に示す。n_i は語 w_i の頻度、p_i は語 w_i の生起確率 (n_i/N) である。V(n, N) は長さ N のテキストにおける頻度 n の語の異なり語数である。HDD は延べ語数 N のテキストから無作為に M 語を非復元抽出したときの異なり語数の期待値である。λ と ℓ は Simpson 指数と呼ばれるもので、テキスト中から無作為に 2 語を抽出したとき同じ語である確率に相当し、λ が復元抽出、ℓ が非復元抽出である。K は $K = 10^4 \frac{\sum_{i=1}^{all} n_i(n_i-1)}{N^2}$ と同値であり、N が十分大きければ ℓ の 10^4 倍と近似した結果を返す。H' は Shannon 指数、エントロピーなどと呼ばれるもので、$H' = -\ln\left[\prod_{i=1}^{all} p_i^{n_i}\right]^{\frac{1}{N}}$ と変形することができ、直観的には延べ語数 N のテキストから無作為に N 語を復元抽出したときに元のテキストと同一の単語列が得られる確率、あるいは各語が元のテキストの同じ位置の語と同一である確率の幾何平均と関係がある。m と S は異なり語数 V(N) と頻度 2 の語 (dis legomena) の異なり語数 V(2, N) の比である。Z は異なり語数 V(N)、延べ語数 N、最頻語の頻度 p を関係付ける係数である。m、S、Z は頻度スペクトルの分布形状を語彙多様性指標として利用するものと考えることができる。

$$HDD = \sum_{i=1}^{all}\left(1 - \frac{\binom{N-n_i}{M}}{\binom{N}{M}}\right)$$

$$H' = -\sum_{i=1}^{all} p_i \ln p_i$$

$$\lambda = \sum_{i=1}^{all} p_i^2$$

$$m = \frac{V(N)}{V(2,N)}$$

$$\ell = \frac{\sum_{i=1}^{all} n_i(n_i-1)}{N(N-1)}$$

$$S = \frac{V(2,N)}{V(N)}$$

$$K = 10^4 \frac{\left[\sum_{n=1}^{all} V(n,N)\left(\frac{n}{N}\right)^2\right] - N}{N^2}$$

$$V(N) = \frac{Z}{\log(pZ)}\frac{N}{N-Z}\log\left(\frac{Z}{N}\right)$$

最後に (3) のタイプの指標について説明する。このタイプの指標は、テキストから一定の条件を満たす単語列を抽出する処理を繰り返し、その特徴量の平均などを指標値として使用する。テキストから一定の長さの単語列を抽出したときの異なり語数を特徴量とするもの (MSTTR、MATTR、vocd) と、TTR が一定の値に達するときの単語列の長さを特徴量とするもの (MTLD、MTLD-MA) がある。ただし前者は異なり語数をそのまま特徴量として使うのではなく、MSTTR と MATTR は TTR、vocd は D という指標に換算する。D は長さ n と異なり語数 V(n) の関係を次の式でモデル化したときの係数であり、多数回の測定で得られたデータからフィッティングによって推定される(2)。

$$TTR = \frac{D}{N}\left[\left(1 + 2\frac{N}{D}\right)^{\frac{1}{2}} - 1\right]$$

テキストから一定の条件を満たす単語列を抽出する方法としては、テキストを一定の条件を満たすセグメントに分割して平均を求める方法、一定の条件を満たすウィンドウをテキス

(2) この式は D について $D = \frac{V^2}{2(N-V)}$ と解くことができる。

トの先頭から末尾まで移動させて平均を求める方法、無作為に単語を抽出する方法がある。MSTTR と MTLD はセグメント平均法、MATTR と MTLD-MA は移動平均法、vocd は無作為抽出法である。前述の HDD は、無作為抽出法と同様の計算を実測ではなく理論値として確率的に計算するものである。

3. 単回帰分析による TTR の補正

前節で示した指標のうち、(1) のタイプの指標は V(N) と N から計算される。V(N) と N の関係を理解するために、BCCWJ コアデータを用いて両者の関係を確認する。図 1 の左は横軸を N、縦軸を V(N) とした散布図である。$N > 12000$ の 1 サンプルを外れ値として除外した。N が大きくなるに従って V(N) も大きくなるが、データポイントの分布は直線的ではなく、V(N) と N が比例しているわけではないことが分かる。図 1 の右は同じデータを両対数グラフにプロットしたものである。両対数グラフでは、データポイントが直線的に分布する。

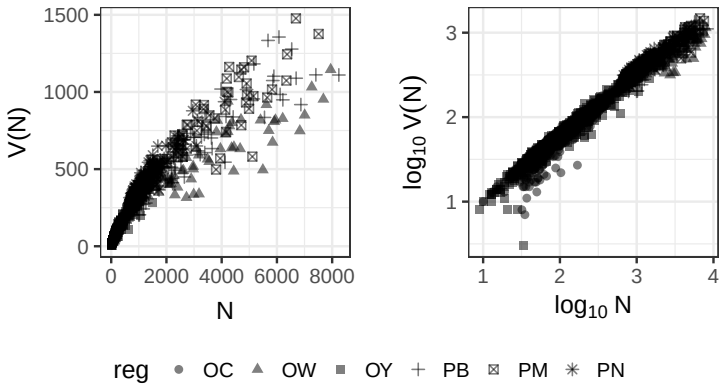

図 1　異なり語数と延べ語数

両対数グラフが直線的に分布するということは、V(N) と N の間にべき乗則 $V(N) = kN^{\beta}$ が成立することを意味する。この関係はヒープスの法則と呼ばれる。ここでは、V(N) と N の関係は次の式で近似できると仮定する。

$$V(N) = 10^{a} \times N^{b}$$

$$\log_{10} V(N) = a + b \times \log_{10} N$$

前節で述べた語彙多様性指標のうち、TTR、R、C は、V と N の間にべき乗則を仮定するモデルと考えることができる。

$$V(N) = TTR \times N \qquad V(N) = R \times N^{0.5} \qquad V(N) = N^{C}$$

これは両対数グラフ上において、TTR、R、C の値が等しいサンプルは、それぞれ直線上に並ぶことを意味する。実際に、TTR、R、C それぞれの上位 5%、下位 5% にあたる値を示す直線を両対数グラフに重ねたものを図 2 に示す。破線は回帰直線である。

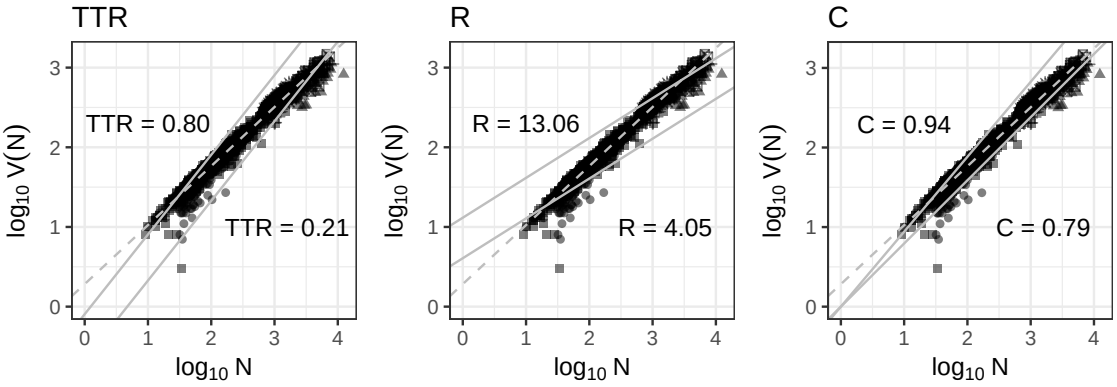

図 2　語の頻度と語彙多様性指標

TTR が等しいデータポイントは傾き 1、R が等しいデータポイントは傾き 0.5 の直線上に並ぶ。しかし実際のデータと比べて TTR は傾きが大きすぎ、R は傾きが小さすぎるため、N が大きくなるほど TTR の実測値は小さくなり、R の実測値は大きくなる。C が等しいデータポイントは、原点を通る傾き C の直線上に並ぶ。しかし実際のデータの分布は原点を通らないため、N が大きくなるほど C の実測値は小さくなる。

そこで実際のデータを単回帰分析して、各データポイントの残差 ε を指標値として利用することを考える。ただし ε をそのまま指標値として使うのではなく、TTR と形式を合わせるために次の式で計算する。

$$ETTR = \frac{V(N)}{N^b}$$

この式は回帰直線の傾き b をパラメータとして、切片 a と残差 ε の和を指標値として利用するもので、$ETTR = 10^{a+\varepsilon}$ である。R が平方根、C が対数を使用するのに対して、この指標は N のべき乗 (exponentiation) を使用するので、指標名は ETTR としておく。b の値はデータを単回帰分析することで推定することができる。BCCWJ コアデータの場合は、$b \approx 0.74$ である。

表 2　単回帰分析

	(Intercept)	log10(N)	R^2	Adj. R^2	Num. obs.	RMSE
log10(V(N))	0.28***	0.74***	0.97	0.97	1980	0.08
	(0.01)	(0.00)				

$^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05$

この指標を両対数グラフに重ねた図を以下に示す。TTR、R、C と比べて、データの分布によく当てはまっていることが確認できる。

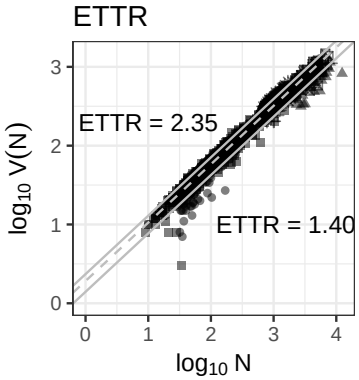

図 3　語の頻度と語彙多様性指標

4. 指標の評価

前節までに挙げた指標がテキストの長さによってどのような影響を受けるか、BCCWJ コアデータを用いて確認する[(3)]。最初に散布図を確認する (図 4)。図の横軸は $\log_{10} N$、縦軸は各指標の値である。CTTR は R の定数倍なので省略した。データポイント数は 1980 だが、一部のデータを外れ値として除外した ($C < 0.7$、$S > 1$、$S < 0.5$、$k < 1$、$a^2 > 0.2$、$U > 100$、$m > 30$、$s < 0.4$)。λ、ℓ、K、Z は値のばらつきが大きいため対数で示した。MSTTR、MATTR、HDD は $N = 42$、MTLD、MTLDMA は $TTR = 0.77$ で計算し、$N \leq 42$ のサンプルを除外した。また MSTTR、MATTR は HDD との比較のため TTR ではなく V(42) に換算して示した。データポイントの形は知恵袋 (OC)、白書 (OW)、ブログ (OY)、書籍 (PB)、雑誌 (PM)、新聞 (PN) の 6 つのレジスタを表す。図中の直線は、回帰直線である。

類似の手法である λ、ℓ、K のうち、ℓ と K はほぼ同様の分布を示しているのに対し、λ は N が小さいときに ℓ や K とは異なる分布を示す。MSTTR、MATTR、HDD は、それぞれ計算の方法が違うものの、似た分布を示す。MTLD と MTLDMA はそれほど似ておらず、MTLD の方がばらつきが大きい。対数関係にある a^2 と Uber、および m と s は、それぞれ対数化した場合に上下対称の分布となる。

次に、N が指標値の平均や分散に及ぼす影響を確認する。平均については、各指標を $\log_{10} N$ で単回帰分析し、その決定係数 R^2 で評価する。分散については、単回帰分析によって得られた残差の絶対値を単回帰分析し、その R^2 で評価する。いずれも R^2 が小さいほど N の影響が小さいと考えられる。結果を図 5 に示す。横軸は指標値の R^2、縦軸は残差の R^2 である。データポイントの形は表 1 で示した各指標の入力種別である。

(3) TTR、R、C、S、a^2、Uber、K、MSTTR、MATTR、HDD、MTLD、MTLDMA は、R の koRpus パッケージを使用して計算した。それ以外の指標は独自に計算した。

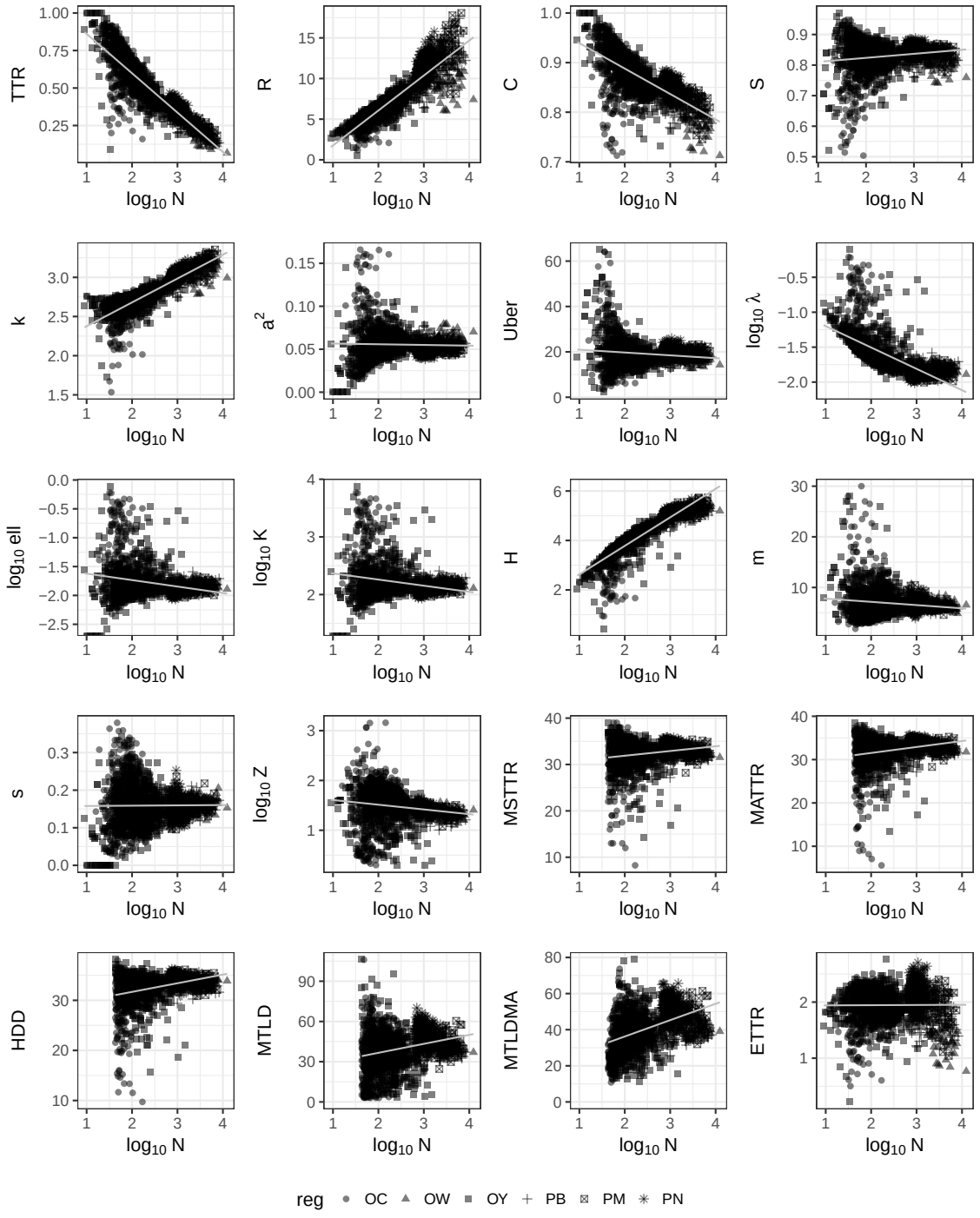

図 4　語の頻度と語彙多様性指標

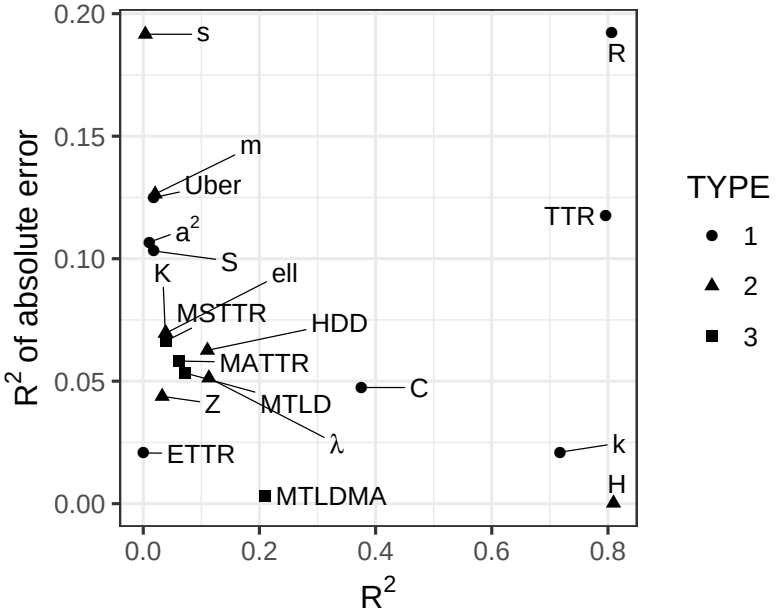

図 5　N が指標値の平均や分散に及ぼす影響

ETTR はどちらの R^2 も小さい値を取り、平均と分散がいずれも N の影響を受けにくいことが分かる。TTR と R はいずれの R^2 も大きな値を取り、N の影響を受けやすい指標だと言える。また R の方がより大きな値を取ることから、TTR より R の方が N の影響が小さいとは言えない。TTR と R 以外では、H や k は平均値が N の影響を強く受けるため、長さの異なるデータの比較には適さないと考えられる。それ以外の指標の多くは、平均が N から受ける影響を小さくすることには成功しているが、分散が N から受ける影響は比較的大きい。図 4 を見ると多くの指標において N が小さいときに指標値のばらつきが大きくなっている。データが小さいときに結果が安定しないことは自然なことではあるが、そのような指標値について N が小さいデータと大きいデータを比較する際には、N が小さいデータの方が極端な値を取りがちであることを考慮する必要がある。

5. 種々の条件における補正値

ETTR はテキストの延べ語数と異なり語数だけで計算することができるが、パラメータ b の値を決定するために単回帰分析を行う必要がある。しかし、あらかじめ b が分かっていれば、その都度単回帰分析を行う必要はない。前々節では、BCCWJ コアデータ (短単位語彙素) において b が 0.74 程度の値になることを確認した。しかし、この値は語の単位やテキストのジャンルといった条件によって変化することが考えられる。そこで本節では、種々の条件下で b がどの程度の値を取るか確認する。コアデータより広い範囲のレジスタを調べるため、BCCWJ 非コアデータの出版サブコーパス (3 レジスタ) と特定目的サブコーパス (9 レジスタ) を調査範囲とした。パラメータとして、次の 3 つのカテゴリーを使用する。

表 3　パラメータ

カテゴリー	値
レジスタ	書籍 (PB) | 雑誌 (PM) | 新聞 (PN) | 白書 (OW) | 教科書 (OT) | 広報誌 (OP) | ベストセラー (OB) | Yahoo!知恵袋 (OC) | Yahoo!ブログ (OY) | 韻文 (OV) | 法律 (OL) | 国会会議録 (OM)
トークン単位	短単位 (suw) | 長単位 (luw)
タイプ単位	語彙素 (lemma) | 書字形基本形 (orthbase) | 書字形出現形 (orth)

これらの組み合わせごとに単回帰分析した結果を表 4 に示す。

表 4　単回帰分析

reg	token	type	a		b		R^2	adj. R^2	Num. obs.	RMSE
OB	luw	lemma	0.3121***	(0.01)	0.7355***	(0.00)	0.96	0.96	1390	0.06
		orthbase	0.2998***	(0.01)	0.7421***	(0.00)	0.96	0.96	1390	0.06
		orth	0.2980***	(0.01)	0.7534***	(0.00)	0.97	0.97	1390	0.05
	suw	lemma	0.3615***	(0.01)	0.7067***	(0.00)	0.95	0.95	1390	0.06
		orthbase	0.3482***	(0.01)	0.7149***	(0.00)	0.95	0.95	1390	0.06
		orth	0.3568***	(0.01)	0.7234***	(0.00)	0.96	0.96	1390	0.06
OC	luw	lemma	0.2039***	(0.00)	0.7845***	(0.00)	0.92	0.92	91445	0.07
		orthbase	0.1940***	(0.00)	0.7920***	(0.00)	0.92	0.92	91445	0.07
		orth	0.1540***	(0.00)	0.8231***	(0.00)	0.92	0.92	91445	0.07
	suw	lemma	0.2737***	(0.00)	0.7432***	(0.00)	0.91	0.91	91445	0.06
		orthbase	0.2599***	(0.00)	0.7538***	(0.00)	0.92	0.92	91445	0.06
		orth	0.2161***	(0.00)	0.7879***	(0.00)	0.92	0.92	91445	0.06
OL	luw	lemma	0.5436***	(0.03)	0.6041***	(0.01)	0.93	0.93	346	0.07
		orthbase	0.5421***	(0.03)	0.6064***	(0.01)	0.93	0.93	346	0.07
		orth	0.5183***	(0.03)	0.6205***	(0.01)	0.94	0.94	346	0.07
	suw	lemma	0.7770***	(0.03)	0.4868***	(0.01)	0.87	0.87	346	0.08
		orthbase	0.7754***	(0.03)	0.4905***	(0.01)	0.87	0.87	346	0.08
		orth	0.7648***	(0.03)	0.5020***	(0.01)	0.88	0.88	346	0.08
OM	luw	lemma	0.3801***	(0.03)	0.7009***	(0.01)	0.98	0.98	159	0.06
		orthbase	0.3704***	(0.03)	0.7045***	(0.01)	0.98	0.98	159	0.06
		orth	0.3401***	(0.03)	0.7203***	(0.01)	0.99	0.99	159	0.06
	suw	lemma	0.6239***	(0.03)	0.5997***	(0.01)	0.98	0.98	159	0.06
		orthbase	0.6099***	(0.03)	0.6057***	(0.01)	0.98	0.98	159	0.06
		orth	0.5828***	(0.03)	0.6213***	(0.01)	0.98	0.98	159	0.06
OP	luw	lemma	0.2099***	(0.02)	0.7973***	(0.01)	0.98	0.98	354	0.03
		orthbase	0.2069***	(0.03)	0.7994***	(0.01)	0.98	0.98	354	0.03
		orth	0.1914***	(0.03)	0.8079***	(0.01)	0.98	0.98	354	0.03
	suw	lemma	0.5775***	(0.04)	0.6548***	(0.01)	0.93	0.93	354	0.04
		orthbase	0.5811***	(0.04)	0.6574***	(0.01)	0.93	0.93	354	0.04
		orth	0.5679***	(0.04)	0.6655***	(0.01)	0.92	0.92	354	0.04
OT	luw	lemma	0.2360***	(0.04)	0.7456***	(0.01)	0.91	0.91	412	0.10
		orthbase	0.2308***	(0.04)	0.7489***	(0.01)	0.91	0.91	412	0.10
		orth	0.2553***	(0.03)	0.7526***	(0.01)	0.92	0.92	412	0.09
	suw	lemma	0.4023***	(0.04)	0.6723***	(0.01)	0.87	0.87	412	0.11

		orthbase	0.3965***	(0.04)	0.6778***	(0.01)	0.88	0.88	412	0.11
		orth	0.4291***	(0.04)	0.6796***	(0.01)	0.89	0.89	412	0.11
OV	luw	lemma	1.3786***	(0.10)	0.4104***	(0.04)	0.35	0.35	252	0.07
		orthbase	1.3459***	(0.10)	0.4255***	(0.04)	0.37	0.37	252	0.07
		orth	1.1158***	(0.08)	0.5139***	(0.03)	0.59	0.58	252	0.05
	suw	lemma	1.6573***	(0.12)	0.3185***	(0.04)	0.20	0.20	252	0.07
		orthbase	1.6221***	(0.12)	0.3356***	(0.04)	0.22	0.21	252	0.07
		orth	1.3047***	(0.10)	0.4529***	(0.03)	0.44	0.43	252	0.06
OW	luw	lemma	0.3842***	(0.02)	0.7167***	(0.01)	0.91	0.91	1500	0.06
		orthbase	0.3800***	(0.02)	0.7188***	(0.01)	0.91	0.91	1500	0.06
		orth	0.3727***	(0.02)	0.7262***	(0.01)	0.92	0.91	1500	0.06
	suw	lemma	0.8427***	(0.03)	0.5361***	(0.01)	0.71	0.71	1500	0.09
		orthbase	0.8525***	(0.03)	0.5369***	(0.01)	0.72	0.72	1500	0.09
		orth	0.8533***	(0.03)	0.5424***	(0.01)	0.73	0.73	1500	0.09
OY	luw	lemma	0.2230***	(0.00)	0.7709***	(0.00)	0.95	0.95	52680	0.08
		orthbase	0.2159***	(0.00)	0.7765***	(0.00)	0.96	0.96	52680	0.08
		orth	0.2014***	(0.00)	0.7922***	(0.00)	0.96	0.96	52680	0.08
	suw	lemma	0.2822***	(0.00)	0.7420***	(0.00)	0.95	0.95	52680	0.08
		orthbase	0.2707***	(0.00)	0.7512***	(0.00)	0.95	0.95	52680	0.08
		orth	0.2558***	(0.00)	0.7674***	(0.00)	0.96	0.96	52680	0.08
PB	luw	lemma	0.3533***	(0.01)	0.7236***	(0.00)	0.89	0.89	10117	0.07
		orthbase	0.3405***	(0.01)	0.7296***	(0.00)	0.89	0.89	10117	0.07
		orth	0.3405***	(0.01)	0.7400***	(0.00)	0.91	0.91	10117	0.07
	suw	lemma	0.5011***	(0.01)	0.6586***	(0.00)	0.81	0.81	10117	0.09
		orthbase	0.4889***	(0.01)	0.6662***	(0.00)	0.81	0.81	10117	0.09
		orth	0.4988***	(0.01)	0.6737***	(0.00)	0.83	0.83	10117	0.09
PM	luw	lemma	0.4526***	(0.02)	0.7092***	(0.00)	0.92	0.92	1996	0.06
		orthbase	0.4389***	(0.02)	0.7152***	(0.00)	0.92	0.92	1996	0.06
		orth	0.4237***	(0.01)	0.7290***	(0.00)	0.94	0.94	1996	0.05
	suw	lemma	0.6380***	(0.02)	0.6362***	(0.01)	0.88	0.88	1996	0.07
		orthbase	0.6272***	(0.02)	0.6437***	(0.01)	0.88	0.88	1996	0.07
		orth	0.6157***	(0.02)	0.6562***	(0.01)	0.90	0.90	1996	0.06
PN	luw	lemma	0.2139***	(0.01)	0.7944***	(0.00)	0.95	0.95	1473	0.03
		orthbase	0.2009***	(0.01)	0.7999***	(0.00)	0.95	0.95	1473	0.03
		orth	0.1816***	(0.01)	0.8126***	(0.00)	0.96	0.96	1473	0.03
	suw	lemma	0.4117***	(0.02)	0.7149***	(0.01)	0.87	0.87	1473	0.04
		orthbase	0.4003***	(0.02)	0.7216***	(0.01)	0.88	0.88	1473	0.04
		orth	0.3937***	(0.02)	0.7298***	(0.01)	0.88	0.88	1473	0.04

表 4 の b の値を ETTR のパラメータとして使用することで、延べ語数の影響を補正した指標値を得ることができる。同時に、これらの値は各レジスタの語彙分布の特徴を示している。他の条件が同じであれば、a が δ 大きくなると V(N) は 10^{δ} 倍になり、b が δ 大きくなると V(N) は N^{δ} 倍になる。N が 10 以上のテキストであれば、b の方が V(N) の値により強く影響する。従って、基本的に b の値が大きいレジスタほど、語彙が多様だと考えられる。luw と suw の値の差が大きいレジスタは、複合語が多いことが予想される。lemma、orthbase、orth の値の差が大きいレジスタは、表記が多様であったり、活用語が豊富であることが予想される。各レジ

スタの語彙の構成は別に調査し検証する必要があるが、本稿では予測の提示に留める。

6. おわりに

TTR などの語彙多様性指標の特徴について検討し、単回帰分析の残差を利用することでテキスト長が指標値に及ぼす影響を補正する方法を示した。この方法はサンプル全体の分布を表すパラメータを単回帰分析によって計算する必要があるが、パラメータが分かっている場合には V(N) と N のみを入力として容易に計算することができ、指標値の平均と分散の変動について他の指標と比べて遜色ない補正を得ることができる。一方で、この手法には問題も存在する。最後に、この手法に関する既知の問題について言及する。

第 1 に、この手法は V(N) と N がべき乗則に従うことを前提とする。厳密には両者の間はべき乗則に従っておらず、V(N) と N の両対数グラフは直線ではなくややカーブする。そのため、N の大きさによってパラメータ b の値は変動する。図 6 は各テキストの先頭から一定の割合の部分単語列を取って単回帰分析したときの b の変動である。OC、OW、OY、PM では N が増加するほど b が減少し、PB と PN は単調に減少はしないが変動している。従って、便宜的には表 4 の b を使用することで TTR や R よりはよい補正を得ることができるが、なるべくよい補正を得るためには分析対象となるサンプルについて単回帰分析を行い、b の値を計算して使用することが望ましい。

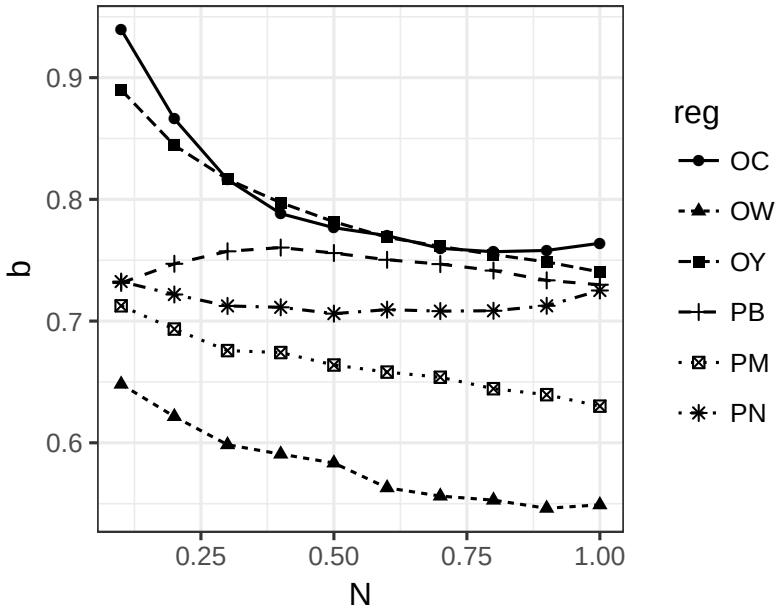

図 6　テキスト長によるパラメータの変動

第 2 に、V(N) と N の値によって計算する全ての手法に共通して言えることだが、語彙の豊かさは必ずしも V(N) と N の値のみから判断できるものではない。例えば、直観的には明らかに語彙の豊かさに差があるように見える小学生と中学生の作文が、語の頻度からみると V(N) と N のいずれも全く等しいということがあり得る。年齢の低い児童が話し言葉の語彙を書き

言葉でも使用するのに対して、年齢が上がるにつれて書き言葉の語彙を習得し使用するようになることを考えると、テキストに現れる語彙数が潜在的な語彙量の総体を正しく反映しているとは必ずしも言えない。語の豊かさについては、語の難易度、レジスタごとの語のふさわしさ、品詞構成比などで評価される語彙密度など、頻度以外の観点からも総合的に判断する必要がある。

文　献

Dugast, Daniel (1979) *Vocabulaire et stylistique*: Slatkine.

Guiraud, P (1954) *Les Charactères Statistiques Du Vocabulaire*: Presses Universitaires de France.

Herdan, Gustav (1960) *Type-Token Mathematics*: Mouton.

Maas, H. D. (1972) "Zusammenhang Zwischen Wortschatzumfang Und Länge Eines Textes," *Zeitschrift für Literaturwissenschaft und Linguistik*, Vol. 8, pp. 73-79.

McCarthy, Philip M. and Scott Jarvis (2007) "Vocd: A Theoretical and Empirical Evaluation," *Language Testing*, Vol. 24, No. 4, pp. 459-488.

Mckee, Gerard, D.D. Malvern, and Brian Richards (2000) "Measuring Vocabulary Diversity Using Dedicated Software," *Literary and Linguistic Computing*, Vol. 15, pp. 323-337.

Somers, H. H. (1966) "Statistical Methods in Literary Analysis," in Leeds, J. ed. *The Computer and Literary Style*: Kent State University Press, pp. 128-140.

木村大翼・田中久美子 (2010) 「文書長に依存しない文書定数」,『言語処理学会第 16 回年次大会発表論文集』, 1090-1093 頁.

鄭弯弯・金明哲 (2018) 「変動係数を用いた語彙の豊富さ指標の比較評価」,『同志社大学ハリス理化学研究報告』, 第 58 巻, 第 4 号, 230-241 頁.