Evaluate Lexical Richness Measures Using Coefficient of Variation

Wanwan ZHENG*, Mingzhe JIN**

(Received October. 23, 2017)

Although numerous lexical richness measures have been proposed, a positive evaluation method has not been established to select measures independent of text length. As an existing evaluation method, it is common to view the transition curves of the measure's original data or standardized data. However, this method is mostly judged visually and cannot sufficiently capture the change of measures. In other words, this method cannot compare and evaluate lexical richness measures directly by viewing transition curves of either original data or standardized data. In this paper, evaluation statistic CV (coefficient of variation) is proposed as a possible method to evaluate lexical richness measures. CV overcomes the drawback of previous research and make it possible to compare the stability of measures by visual observation. A total of 11 measures of *TTR*, *K*, *R*, *S*, *Uber*, *C*, *s*, *LN*, *k*, *M* and *m* are compared and evaluated using CV. Meanwhile, Japanese, Chinese, and English corpora are used to avoid the possible influence of the languages. Analysis results indicate that *s* is the measure with the smallest influence of text length and language.

Key words: lexical richness measure, coefficient of variation, comparison, evaluation

キーワード：語彙の豊富さを表す指標，変動係数，比較，評価

変動係数を用いた語彙の豊富さ指標の比較評価

鄭　弯弯，　　金　明哲

1. はじめに

近年，語彙の豊富さは，言語学や文章学だけではなく，神経病理，言語習得，法医学など幅広い領域に用いられている[1)]．語彙の豊富さを表す指標は複数提案されている．

語彙の豊富さの測定における最も基本的な考え方は，述べ語数 N の中に占める異なり語数 V(N) の割合 *TTR*（Type Token Ratio）である[2)]．この値が大きければ多様な単語が使われていて語彙が豊富であり，小さければ同一語の繰り返しが多く，語彙が乏しいと解釈する．しかし，文章が長くなるにつれて，異なり語数の増加率は徐々に小さくなるのに対し，延べ語数の影響が顕著になる．つまり，文章の長さに依存して *TTR* は小さくなり続けていく．

$$TTR = \frac{V(N)}{N} \tag{1}$$

TTR は文章の長さに依存していることが多くの研究で検証された．しかし，これまでの語彙の豊富さに関する研究において，その重要性は無視することはできない．延べ語数の *TTR* の値に対する影響を抑えるため，平方根と対数の特性を利用し，*TTR* を改良した指標が多く提案されている．平方根を用い

*Graduate School of Culture and Information Science, Doshisha University, Kyoto
Telephone : +81-080-9121-7928, E-mail : diq0015@mail4.doshisha.ac.jp
**Faculty of Culture and Information Science, Doshisha University, Kyoto

た改良型の指標としては R [3]，$CTTR$（Corrected Type-token Ratio）[4] がある．

$$R = \frac{V(N)}{\sqrt{N}} \tag{2}$$

$$CTTR = \frac{V(N)}{\sqrt{2N}} \tag{3}$$

対数を用いた改良型の指標としては C[5,6]，s，M，$Uber$ [7,8]，LN，k [8] が挙げられる．

$$C = \frac{\log V(N)}{\log N} \tag{4}$$

$$s = \frac{\log(\log V(N))}{\log(\log N)} \tag{5}$$

$$M = \frac{\log N - \log V(N)}{(\log N)^2} \tag{6}$$

$$Uber = \frac{(\log N)^2}{\log N - \log V(N)} \tag{7}$$

$$LN = \frac{1 - V(N)^2}{V(N)^2 \log N} \tag{8}$$

$$k = \frac{\log(V(N))}{\log(\log(N))} \tag{9}$$

このような指標の共通点は，異なり語数と延べ語数の関係に基づいて計算されたことである．しかし，異なり語数と延べ語数の比を基とする指標は語彙の豊富さを表す妥当な指標にはならないと主張されている[9]．また，これらの改良型はサンプルサイズの影響からは逃れられないと指摘されている[10,11]．

この問題を解決するため，述べ語数と異なり語数のほかに語彙が用いられている回数（頻度スペクトル）を加えた，K特性値が提案された[12]．K特性値は，値が小さいほど語彙が豊富であることを示す指標である．しかし，K特性値もテキストの長さやデータ構造に依存していることが指摘された．

$$K = 10^4 \frac{[\sum_{i=1}^{all} V(i,N)(i/N)^2] - N}{N^2} \tag{10}$$

Kと類似する指標として，次に示す m と S [13] が提案されている．

$$m = \frac{V(N)}{V(2,N)} \tag{11}$$

$$S = \frac{V(2,N)}{V(N)} \tag{12}$$

また，単語の使用頻度と順位との関係から導き出した Zipf 法則を拡張し，異なり語数，延べ語数，文章の中に最も多く出現する単語の相対頻度（p^*）と関わるパラメータ Z [14] が提案されている．

$$V(N) = \frac{Z}{\log(p^* Z)} \frac{N}{N - Z} \log(\frac{N}{Z}) \tag{13}$$

ここで，p^* は述べ語数 N によらない定数であると仮定している．Z 値が大きくなるにつれて，異なり語数 V(N) の値も大きくなるため，Z は文章の語彙の豊富さを表す指標と解釈できる．しかし，Z については，一人の著者によって書かれた英語の文章においては値が一定となったが，複数の著者によって書かれた大規模な文章では一定とならなかった．また英文以外の文章では値が一定とならなかったと結論づけている[15]．

近年，D [16] と $MTLD$ [17]（Measure of Textual Lexical Diversity）のようなソフトウェアベースの語彙の豊富さの指標が提案されている．D は専用のツール vocd[18] と D_Tools[16] で算出できる．D は一連の無作為に抽出したテキストのサンプリングを対象にして計算した結果である．TTR と延べ語数の関係式に D というパラメータを加えたモデルを作り，TTR 曲線と最も当てはまりがいいモデルを求める．次にその式を示す．D の値は 10 から 100 までの区間内にあり，値が大きければ大きいほど語彙が豊富である．

$$TTR = \frac{D}{N}[(1 + 2\frac{N}{D}) - 1] \tag{14}$$

一方，D に関して，D は単語の出現頻度に大いに

影響され，また計算は複雑な曲線にあてはめることが必要であるが， 計算がより簡単な *CTTR* と非常に似ている指標であると報告されている[1,19]．Vermeer（2004）は明らかに語彙の違いのあるグループが判別できていないことを指摘し，*D* は妥当な指標にならないと結論づけた[9,20]．また，超幾何分布を使って直接計算できる *HD-D* と *D* の間には高い相関があることを示した上で，*HD-D* を *D* の代わりの指標として用いることが提案されている[21]．

MTLD は，専用ツール The Gramulator で算出できる．*MTLD* では，まずテキストの先頭から単語を１つずつ累加しながら *TTR* を求め，*TTR* の値が 0.72 になると終了する．ここで得られたサブテキストを一つのサンプルとする．その後，終了時点から上記の作業を繰り返し，一つのテキストを複数に分割して，一つのテキストから *n* 個のサンプルを作成する．*L* を元文章の延べ語数，*n* をサンプル数にすると *MTLD* は次の式で定義されている．

$$MTLD = \frac{L}{n}, \qquad (n>1) \qquad (15)$$

MTLD は文章の長さに影響されにくいが，語彙の豊富さの指標 *M* よりは文章の長さに影響されやすいと報告されている[22]．

また，文章ごとに *TTR* 値を求めるのではなく，それを同じサイズに分割し(通常は 100 単語)，それぞれのセグメントに対して *TTR* を求め，文章全体の *TTR* を各セグメントの *TTR* の平均値 *MSTTR* (Mean Segmental Type-token Radio) を用いることが提案された．*MSTTR* の式を次に示す．*MSTTR* の基本的なアプローチは *MTLD* と類似している．

$$MSTTR = \frac{\sum_{i=1}^{n} TTR_i}{n}, \qquad (n>1) \qquad (16)$$

このような数多くの指標の中から文章の長さに対する依存性の低い指標を見つけ出すため，いくつかの評価実験が行われている．しかし，そのほとんどの研究は個別の指標しか用いなかった．Hout and Vermeer（2007）は *TTR*，*R*，*C* 三つの指標について評価実験を行い，*R* が最も優れていると述べた[23]．McCarthy and Jarvis（2010）は，*MTLD*, *M*, *D*, *HD-D*, *K*，*TTR* の六つの指標を比較し，*MTLD*，*M*，*D*，*HD-D* を推奨した．Koizumi and In'nami（2012）は，*TTR*，*R*，*M*，*MTLD*，*D*，*HD-D* について比較分析し，*MTLD* は最も文章の長さに依存しないが，延べ語数が 50～150 語，50～200 語の区間内で不安定になることを示した[24]．Torruella and Capsada（2013）は，*TTR*，*R*，*STTR*，*M*，*MSTTR*，*MTLD*，*HD-D* の七つの指標の中で *M*，*MSTTR*，*MTLD*，*HD-D* が文章の長さに影響されにくく，*M* が最も安定していると報告した．Kimura and Tanaka（2010）は，*K*，*Z*，*VM*，*H*，*r* の五つの指標の中で *K* と *VM* が文章の長さに最も依存しないことを示した．しかし，以上の先行研究が用いた指標の数が少なく，数個の指標の中で，ある指標が相対的に安定しているとしか説明できない．さらに，指摘された文章の長さに依存しにくい *MTLD* と *VM* の算出は非常に複雑であり，ある程度の専門知識とツールを持たなければならない．また複雑な計算が必要であるため，計算のスピードも遅い．

計算が複雑な指標 *D*，*MTLD*，*VM* について，Mizuki（2008）と McCarthy and Jarvis（2010）は，*D* の代わりに *CTTR* と *HD-D* の使用することを主張し，Torrulla and Capsada（2013）は，*MTLD* は *M* より劣っていると結論づけている．また Kimura and Tanaka（2010）は，*VM* は *K* と共に文章の長さに依存しにくいが，*VM* は *K* より安定性に欠けていると報告した．このようなことから，計算が複雑であっても，必ずしもよい指標であるとは言えず，*D*，*MTLD*，*VM* ともにそれなりの効果が得られなかった．

Tweedie and Baayen（1998）は 12 個の指標について評価実験を行った[25]．用いた評価方法は元データで描いた各指標の遷移曲線をそれぞれ考察することである．この方法では各指標のスケールが異なる条件下におけるバラツキを用いて指標の安定性を比較しているため，その妥当性が問題となる．また，指標間の比較が難しいため，改善案としてデータを標準化し，各指標を一つの散布図にまとめる方法が提案された[19,24]．しかし標準化を行っても，データの全体的な分布は変わらないため，元データを用いた場合と同じである．このような先行研究の問題点を踏まえて，本研究は変動係数を用いた評価方法を提

案する．

本研究の分析対象は簡単に計算でき，応用しやすい指標である *TTR*, *K*, *R*, *S*, *Uber*, *C*, *s*, *LN*, *k*, *M*, *m* の計 11 個の指標について，提案した方法を用いて文章の長さへの依存性について比較評価実験を実施する．また，語彙の豊富さの指標が言語に依存しているかについて，日本語，中国語，英語のコーパスで確認を行う．

2. コーパス

分析結果の一般性を示すため，各指標の評価に用いるコーパスを作成するにあたり，複数の著者の作品を選んだ．また，執筆の時期によって文章の語彙の豊富さが変わる可能性が十分考えられるため，長篇小説を対象とする．今回は日本語（84, 058～393, 706 語），中国語（67, 511～257, 811 語），英語（71, 737～193, 372 語）のコーパスを作成した．コーパスのリストを Table．1 に示す．

3. 問題点と改善法による分析結果

用いたデータが元データでも標準化したデータでも，目視による評価では各指標の安定性を十分に正しく比較評価することは困難である．そこで，目視によって各指標の安定性を直接比較するために，変動係数で考察する方法を提案する．

先行研究[15,25]では，文章をチャンクごとに分けて分析を行った．先行研究は一つの文章を 20 チャンクに分け，チャンクサイズを 1000 語ぐらいにした[25]．本研究では，各指標の変化を細かく考察するため，チャンクのサイズを 100 語に設定し，文章をいくつかのチャンクに分け，チャンクを一つずつ累加し，それぞれの指標の値を求める．

異なる作家と作品に対して，ある語彙の豊富さを表す指標が異なる変化を示す可能性は十分考えられるため，本研究では，各指標 I に対してチャンク i ごとに 10 篇の文章の平均値を求めて分析する．

$$\bar{v}_{Ii} = \frac{1}{10}\sum_{j=1}^{10} v_{Ii,j} \tag{17}$$

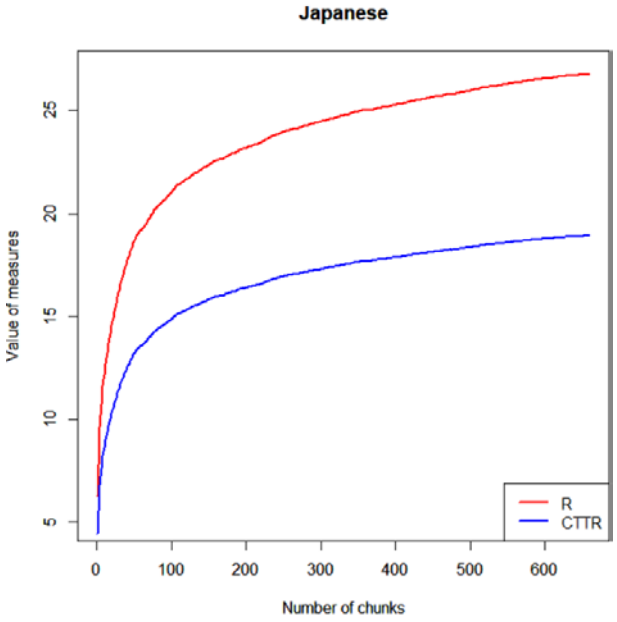
3.1 変動係数

3.1.1 *R* と *CTTR* の関係

平均値からのバラツキの度合を表す標準偏差は，データの平均値と単位が異なると直接比較できない．平均値および単位が異なったデータのバラツキを比べる統計量として変動係数（CV，coefficient of variation）がある．式を以下に示す．

$$CV = \frac{\sigma}{x} \tag{18}$$

語彙の豊富さの指標 *CTTR* は *R* を $\sqrt{2}$ で割ったものである．ここで，1 より大きい数値で割ることで，指標が改良できるかが問題となる．*R* と *CTTR* の元データで描いた遷移曲線を考察するため，Fig．1 に日本語におけるチャンクを累加しながら得られた指標値の曲線を示す．横軸はチャンクであり，縦軸が得られた指標値である．Fig．1 を考察すると，*CTTR* は *R* より安定しているように見える．*R* と *CTTR* の標準偏差を求めると，*R* の標準偏差は *CTTR* の $\sqrt{2}$ 倍になっているが，*R* と *CTTR* の変動係数は同じである．つまり，*CTTR* は *R* に対しての改良はできていない．



Fig．1．Transition curves of the original data of *R* and *CTTR*.

3.1.2 標準化データの遷移状況と変動係数の遷移状況の比較

TTR と比べて文章の長さに影響されにくいと指摘された *M*, *K*, *R*, *C* を標準化したデータの遷移状況と変動係数の遷移状況を比較してみる．

標準化を行うと各指標の分布を変えず，各指標を一つのプロットにまとめて比較できるため，まず次の式を用いて得られた $\bar{v}_{li}$ を標準化する．

$$x' = \frac{x - mean}{sd} \times 10 + a \tag{19}$$

Table 1.　The list of corpus.

	ID	Works	Published year	Characters	Tokens
Chinese	C1	Long River (C. W. Shen)	1938	353,251	67,511
	C2	Camel Xiangzi (Laos)	1939	433,207	88,437
	C3	The Last Quarter of the Moon (Z. J. Chi)	2005	335,309	126,167
	C4	Looks Great (S. Wang)	2004	636,334	125,959
	C5	The Lotus Lake (L. Sun)	1945	668,849	257,811
	C6	Fortress Besieged (Z. S. Qian)	1947	392,484	150,785
	C7	White Deer Plain (Z. S. Chen)	1992	393,692	148,979
	C8	Spring Fever (Yudafu)	1923	428,909	167,750
	C9	Soul Mountain (X. J. Gao)	1990	457,294	175,285
	C10	Happy accuser (H. S. Zhang)	2004	563,350	220,996
English	E1	When All The Woods are Green (S. W. Mitchell)	1894	426,724	103,334
	E2	The Sheik (E. M. Hull)	1919	396,149	87,825
	E3	The Life of Charlotte Bronte (E. Gaskell)	1857	334,024	78,407
	E4	White Nights and Other Stories (F. Dostoevsky)	1848	394,721	89,632
	E5	King Coal (U. Sinclair)	1917	326,589	71,737
	E6	Wastralls (C. A. D. Scott)	1918	433,266	93,266
	E7	The Essays of George Eliot (G. Eliot)	1883	506,459	102,280
	E8	Eve(S. B. Gould)	1891	521,975	118,014
	E9	Sister Carrie (T. Dreiser)	1900	726,711	155,980
	E10	The Financier (T. Dreiser)	1912	906,612	193,372
Japanese	J1	The Makioka Sisters (J. Tanizaki)	1949	301,402	102,729
	J2	Kozakura Hime Story (A. Asano)	1985	271,379	93,059
	J3	Aphrodisiac of play (Y. Mishima)	1997	259,988	100,667
	J4	The paradox of Youth (S. Oda)	1976	323,332	111,485
	J5	Female genealogy (K. Izumi)	1907	340,477	131,230
	J6	Heralds (S. Natsume)	1942	290,976	105,917
	J7	A Certain Woman (T. Arishima)	1919	409,403	131,993
	J8	Thirst for Love (Y. Mishima)	1950	246,992	84,058
	J9	I am a Cat (S. Natsume)	1987	637,240	213,319
	J10	September Affair (R. Yokomitu)	1937-1946	1,147,270	393,706

式の中の x は元データ，x' は標準化データ，a は変換後にマイナスの値にならないようにするための定数である．続いて，五つのチャンクごとの $\bar{v}_{li}$ に対して，右に一つずつシフトしながら変動係数を求める．

Fig. 2 に日本語コーパスにおける *TTR*, *M*, *K*, *R*, *C* の五つの指標値を標準化したデータのプロットを示し，Fig. 3 にその変動係数のプロットを示す．縦軸は標準化したデータと変動係数，横軸はチャンクの数とセグメントナンバーである．Fig. 2 の右の部分では *TTR* が *C*, *M*, *R* より安定し，*K* が最も安定しているように読み取られる．一方，Fig. 3 に示した変動係数を考察すると，*TTR* の変動係数が最も大きい．変動係数は単位と平均値が異なるデータのバラツキを比べるための統計量であるので，*TTR* は *M*, *K*, *R*, *C* より文章の長さへの依存性が高いと言える．また，*C* は *K* より安定している．

同じ方法で中国語を Fig. 4 と Fig. 5 に，英語を Fig. 6 と Fig. 7 に示す．中国語の場合，標準化データを用いた Fig. 4 から，*M*, *C* より *R*, *K* が文章の長さに依存しにくいように読み取れるが，変動係数を用いた Fig. 5 を考察すると，逆になっている．英語でも同様の傾向が見られた．

ちなみに，元データと標準化データの遷移状況を考察した先行研究では *C*, *M* をよい指標として評価できなかった[24,25]．

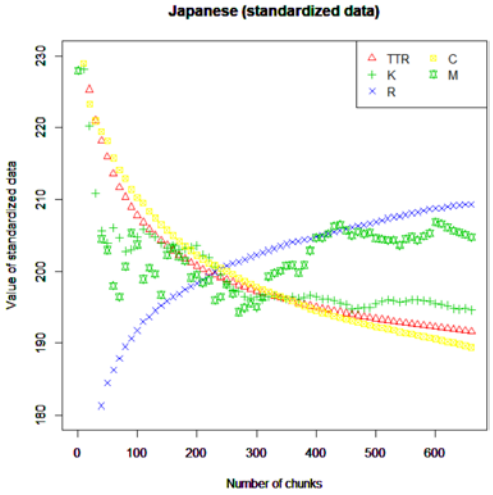


Fig. 2. Transition of Japanese standardized data.

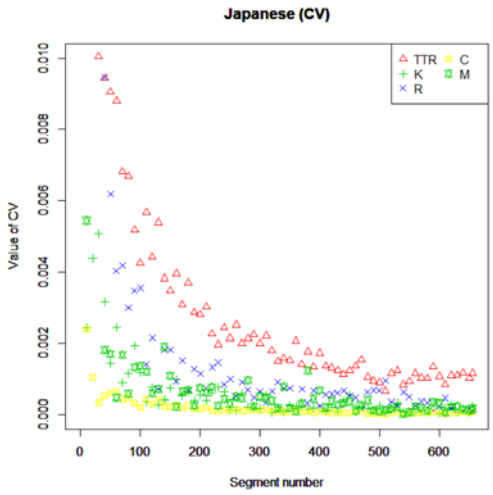


Fig. 3. Transition of Japanese CV.

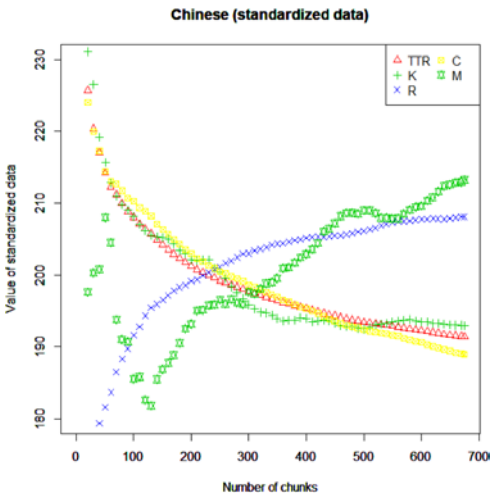


Fig. 4. Transition of Chinese standardized data.

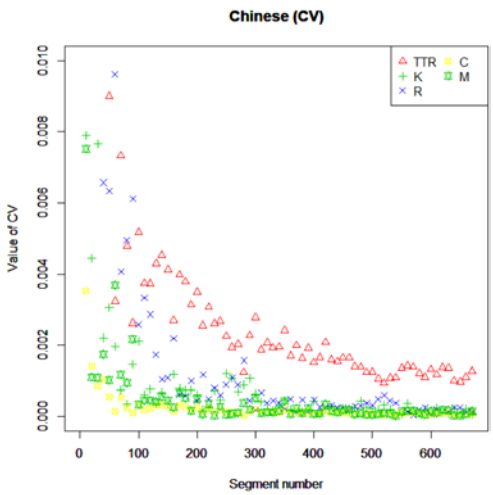


Fig. 5. Transition of Chinese CV.

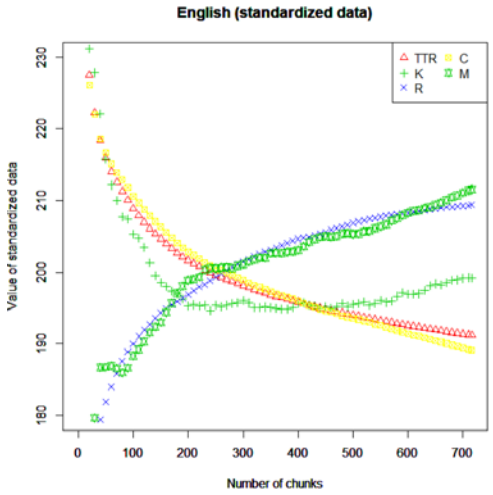


Fig. 6. Transition of English standardized data.

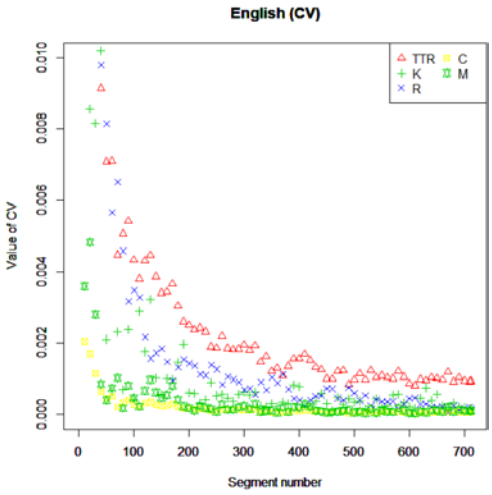


Fig. 7. Transition of English CV.

4. 指標についての比較評価

4. 1 評価指標と分析データ

本節では提案した変動係数を用いて，よく使われている *TTR*, *K*, *R*, *S*, *Uber*, *C*, *s*, *LN*, *k*, *M*, *m* の計 11 個の指標について，どの指標が最も文章の長さに影響されにくいかを分析する．本研究では，$\bar{v}_{Ii}$ の変動係数を用いて，バラツキと傾きの視点から収束性と安定性について目視による考察と定量的な分析で語彙の豊富さの指標を評価する．

日本語，中国語，英語における 11 個の語彙の豊富さ指標について，五つのチャンクごとの $\bar{v}_{Ii}$ に対し，右にシフトしながら各指標の変動係数を求め，その散布図で考察することにする．Fig. 8 に示したものは求めた変動係数について 10 を間隔にして取ったデータである．つまり，変動係数の 1 番目，10 番目，20 番目…のデータである．Fig. 8 (a) が日本語，Fig. 8 (b)が中国語，Fig. 8 (c)が英語の結果である．

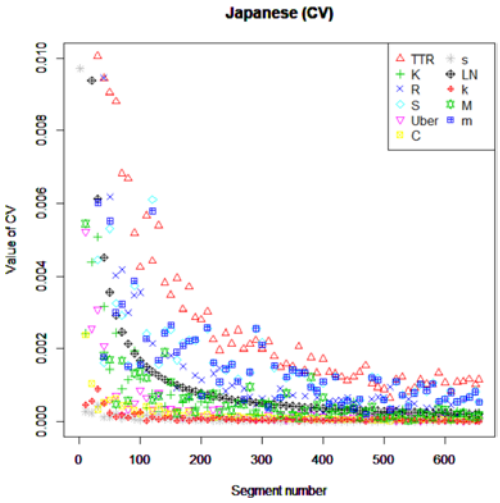
図の左側は 11 個の指標が激しく変化し，チャンクの累加につれて分散が徐々に小さくなっている．三つの言語とも明らかに *TTR* が最も文章の長さの影響を受けている．*R* と *m* も劣っていることが読み取れる．これらのプロットを考察しながら明らかによくない指標を除いていくと指標 *s* と *k* が最も文章の長さに依存せず，その次に *C* が文章の長さに依存しにくいことがわかった．このような目視の結果を踏まえ，次の節ではバラツキと傾きを用いて，収束性と安定性二つの側面から定量的に評価を行う．

Fig. 8 より，言語に関わらず，各指標の変化が最初は激しく，チャンクを累加するにつれて徐々に緩やかになっていることがわかる．個別指標を除くと 200 番目の変動係数前後（200 チャンク前後）で収束している．これをさらに検証するために，以下の式を用いて，始点からの i 番目の変動係数の変化率を求めて考察する．

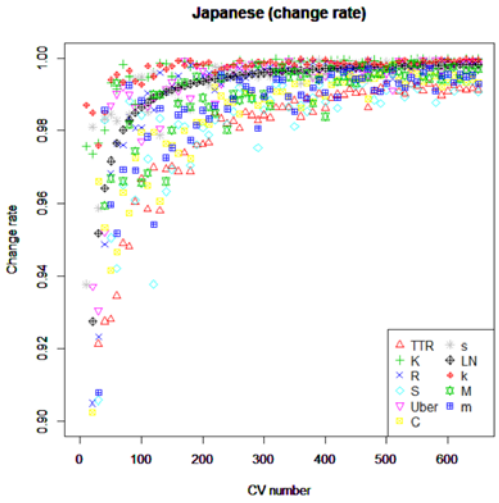
$$\Delta CV_{Ii} = |CV_{Ii} - CV_{I1}|/CV_{I1}, i = 2,3,4... \quad (20)$$

式から分かるように ΔCV_{Ii} は変動係数の相対的な変化の大きさを表す．求めた変化率の散布図を Fig. 9 に示す．考察をしやすくするため Fig. 9 に示しているデータは求めた変化率に対し，10 チャンク毎の値である． Fig. 9 (a) が日本語，Fig. 9 (b)が中国語，Fig. 9 (c)が英語の結果である．

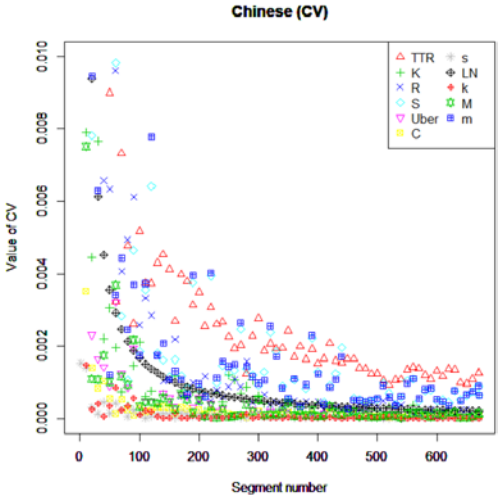
日本語，中国語，英語において語彙の豊富さ指標はほぼ 200 番目の変化率前後（200 チャンク前後）で遷移曲線は前より安定になっている．これは Fig. 8 の考察結果と一致している．200 チャンク以降，個別の指標に大きな変化が起こっていても，それは他の指標より文章の長さに依存すると考え，収束性に対しての議論は 1～200 チャンクまでとして，200 チャンク以降は指標の安定性の問題として扱う．



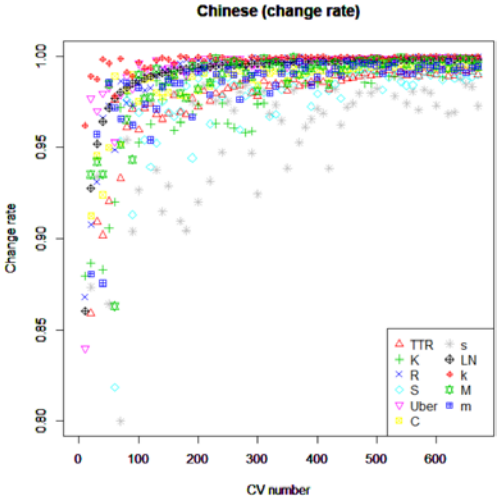
(a) Japanese



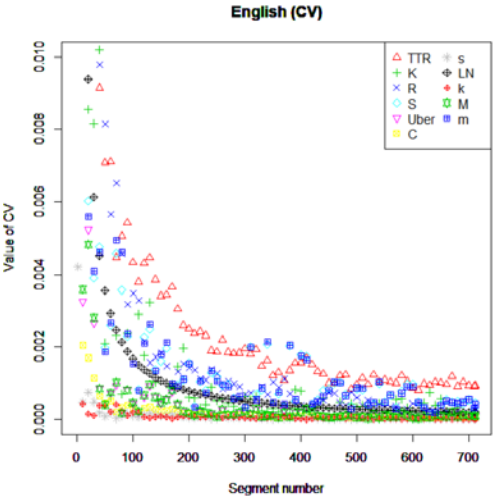
(a) Japanese



(b) Chinese

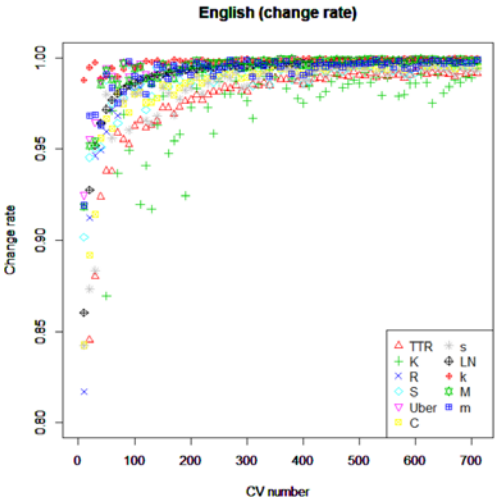


(b) Chinese



(c) English

Fig. 8. Transition of 11 measures' CV.



(c) English

Fig. 9. Transition of CV' s change rate.

4. 2　バラツキ

本節では，バラツキから各指標の収束性と安定性を議論する．Fig. 8 において，各指標の最初の値は異なるが，文字数を増やすことにつれて，各指標の遷移曲線は徐々に近くなっている．また，安定するとバラツキと傾きは共に小さくなる．

日本語における 11 指標値について 1～50 チャンク，51～100 チャンク，101～150 チャンク…それぞれの移動変動係数の平均値を求め，その棒グラフを Fig. 10 に示す．1～50 チャンクの変動係数の平均値が非常に大きいため，ほかの区間の結果が読み取りにくくなっている．中国語と英語も同様であった．そこで，1～50 チャンクのデータを除いた棒グラフを Fig. 11 に示す．Fig. 11 (a)，(b)，(c) それぞれは，日本語，中国語，英語の結果である．Fig. 11 から，言語に関わらず，*s*, *k*, *C* が他の指標より優位性を示し，*s* の変動係数の平均値が最も小さいこと読み取れる．一方，*TTR*, *R*, *S*, *LN*, *m* は文章が長くなるにつれて，他の指標より変化がやや激しいこともわかった．棒グラフによる考察結果と Fig. 8 の考察結果は一致する．

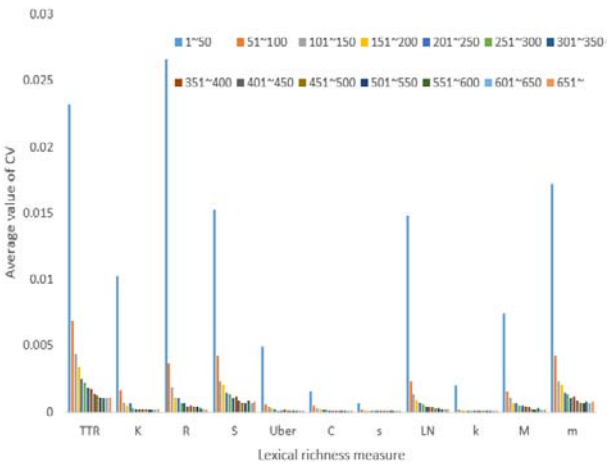
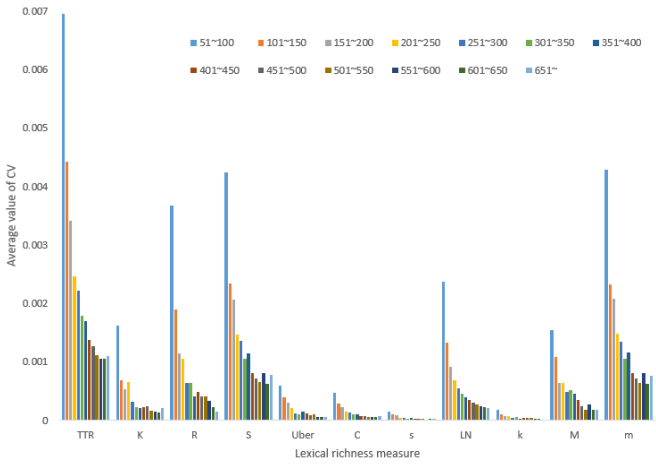
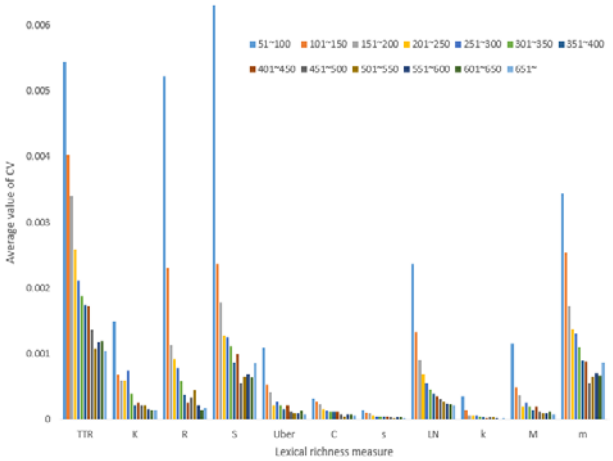


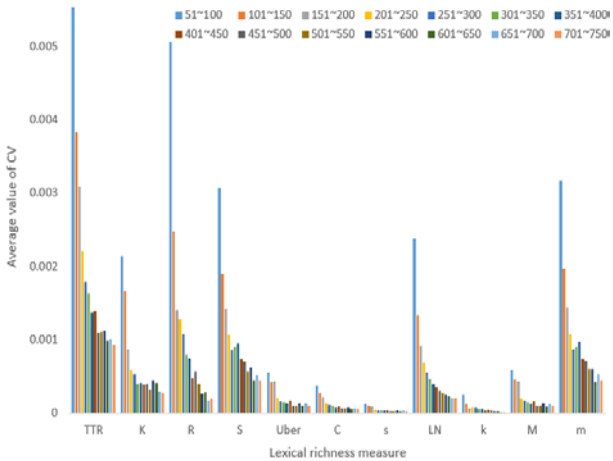
Fig. 10. Bar plot of Japanese CV.



(a) Japanese



(b) Chinese



(c) English

Fig. 11. Bar plot of CV.

4. 3 遷移曲線の傾き

各指標の収束状況を詳しく把握するため，遷移曲

線について一定区間を区切り，区間内で線形単回帰分析を行う．回帰式 $y=a+bx$ の b は回帰直線の傾きであり，直線の傾斜具合を表す．この傾斜具合は指標の変動係数の傾きを示すものである．Fig. 8 に示す各指標の遷移曲線における変動係数の回帰直線の傾きを詳しく分析するため，ここでは移動平均のアイディアを借用し，移動回帰を行う．移動は最初の変動係数から隣接している 10 番目までの変動係数を一つの区切りにし，右に一つシフトしながら回帰モデルを作成する．得られた傾きの箱ひげ図を Fig. 12 に示す．Fig. 12 (a)，(b)，(c) は日本語，中国語，英語の 1～200 チャンクの傾きであり，(d)，(e)，(f) は 200 チャンク以降の傾きである．

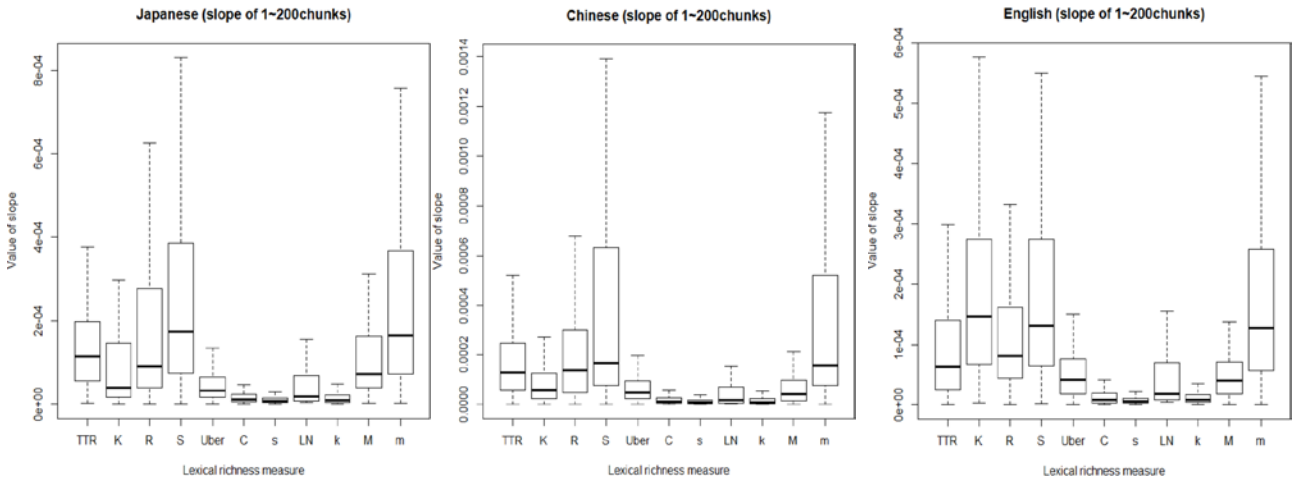
Fig. 12 により， 前節で小さなバラツキを示した s, k, C の傾きは他の指標より小さく，s が最も優れている．200 チャンク以後の LN の傾きは相対的に小さいが，前節で検証した変動係数が大きい．よって，総合的に判断すると s より良い指標であるとは判断し難い．

5. まとめ

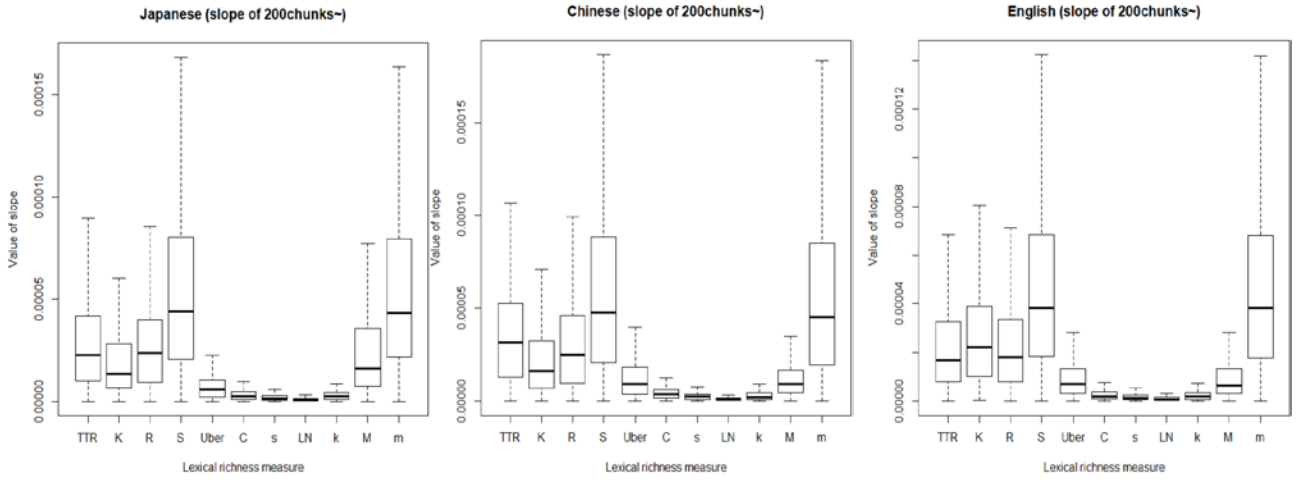
本研究では，先行研究における元データもしくは標準化データの遷移曲線のみを考察する方法は適切性に欠けているという問題点を提起し，変動係数による指標の評価方法を提案し，日本語，中国語，英語のコーパスを用いて比較分析を行った．その結果，提案した方法では，標準化したデータでは見られなかった文章の長さへの依存性と指標の改良効果などが明確に考察できた．

また，バラツキと回帰モデルの傾き係数を用いて *TTR*, *K*, *R*, *S*, *Uber*, *C*, *s*, *LN*, *k*, *M*, *m* の計 11 個の指標における収束性と安定性について比較評価を行った．目視と統計分析の方法により考察することで相互検証した結果，日本語，中国語と英語に関わらず，バラツキと回帰モデルの傾き係数において，他の指標より，s がより小さい値が得られた．以上により，文章の長さと言語の依存性が最も低い指標は s であることを明らかにした．語彙の豊富さ指標 s が文章の長さに最も影響が小さいのは 2 重対数をとったことが一つの理由であると考える．対数を n 重(n>2)とった場合どのような振る舞いをするかなどに関しては別紙に譲る．

本研究の一部は，ハリス理化学研究所の研究助成金によって行った．ここに記して感謝する．



(a) Japanese (b) Chinese (c) English



(d) Japanese (e) Chinese (f) English

Fig. 12. Box plot of CV's slope.

参考文献

1) P. M. McCarthy, S. Jarvis, "Vocd: a Theoretical and Empirical Evaluation", *Language Testing*, **24**, 459-488 (2007).
2) M. Templin, *Certain Language Skills in Children: Their Development and Interrelationships*, (The University of Minnesota Press, Minneapolis, 1957).
3) H. Guiraud, *Les Caracteres Statistiques du Vocabulaire*, (Universitaires de Fance Press, Pairs, 1954).
4) J. B. Carroll, on Sampling from a Lognormal Model of Word-frequency Distribution, *Computational analysis of present-day American English*, (Brown University Press, Providence, 1967), pp. 406-424.
5) G. Herdan, *Type-Token Mathematics: A Textbook of Mathematical Linguistics*. (Mouton & Co, The Hague, The Netherlands, 1960).
6) G. Herdan, *Quantatative Linguistics*, (Butterworth, London, 1964).
7) D. Dugast, "Sur Quoi se Fonde la Notion D'etendue Theoratique du Vocabulaire?", *Le Francais Modern*, **46**[1], 25-32 (1978).
8) D. Dugast, *Vocabulaire et Stylistique. I: Theatre et Dialogue*, Travaux de Linguistique Quatitative, (Slatkine-Champion, Geneva, 1979).
9) A. Vermeer, "The Relation between Lexical Richness and Vocabulary Size in Dutch L1 and L2 Children", *Vocabulary in a Second Language*, 173-189 (2004).
10) D. Malvern, B. Richards, "Investing Accommodation in Language Proficiency Interviews Using a New Measure of Lexical Diversity", *Language Testing*, **19**[1], 85-104 (2002).
11) M. Tajima, J. Fukada, H. Satou, K. Tamaoka, "Relationship between Lexical Index and Subjective Evaluation of Texts", *Chuo Gakuin University Human and Natural Discussion*, **19**, 57-77 (2009).
12) G. U. Yule, *The Statistical Study of Literary Vocabulary*, (Cambridge University Press, Cambridge, 1944).
13) H. S. Shichel, "On a Distribution Law for Word Frequencies", *Journal of the American Statistical Association*, **70**, 542-547 (1975).
14) Y. K. Orlov, *Ein Modell der Haufigkeit Struktur des Vokabulars*, In Studies on Zipf's Law, (Brockmeyer, Bochum, 1983), pp. 154-233.
15) D. Kimura, K. Tanaka, "A Study on Constants of Natural Language Texts", *Journal of Natural Language Processing*, **18**[2], 119-137 (2011).
16) P. M. Meara, I. Miralpeix, D_Tools (version 2. 0; lognostics: Tools for Vocabulary Researchers: Free Software from Lognostics) [Computer Software], University of Wales Swansea (2007).
17) P. M. McCarthy, "An Assessment of the Range and Usefulness of Lexical Diversity Measures and the Potential of the Measure of Textual", Lexical Diversity (MTLD) (Unpublished PhD dissertation). *University of Memphis*, (2005).
18) G. Mckee, D. Malvern, B. Richards, "Measuring Vocabulary Diversity Using Dedicated Software", *Literary and Linguistic Computing*, **15**[3], 323–337 (2000).
19) A. Mizuki, "Relation between Lexical Indexes and Holistic Scoring by Raters in an English Essay", *The Institute of*

*Statistical Mathematics Cooperative Research Repor*t *215*, 15-28 (2008).

20) S. Jarvis, “Short Text, Best-fitting Curves and New Measures of Lexical Diversity”, *Language Testing*, **19**[1]**,** 57-84 (2002).

21) P. M. McCarthy, S. Jarvis, “MTLD, vocd-D, and HD-D: a Validation Study of Sophisticated Approaches to Lexical Diversity Assessment”, *Behavior Research Methods*, **42**, 381-392 (2010).

22) J. Torrulla, R. Capsada, “Lexical Statistics and Tipological Structures: A Measure of Lexical Richness”, *Procedia-Social and Behavioral Sciences*, **95**, 447-454 (2013).

23) R. V. Hout, A. Vermeer, “Comparing Measures of Lexical Richness”, *Modeling and Assessing Vocabulary Knowledge*, 93-115 (2007).

24) R. Koizumi, Y. In’nami, “Effects of Text Length on Lexical Diversity Measures: Using Short Texts with Less Than 200 Tokens”, *System*, **40**, 554-564 (2012).

25) F. J. Tweedie, R. H. Baayen, “How Variable May a Constant be? Measures of Lexical Richness in Perspective”, *Computers and the Humanities*, **32**, 323-352 (1998).