語彙の豊富さの新しい指標

鄭　弯弯

同志社大学文化情報学研究科
diq0015@mail4.doshisha.ac.jp

1 はじめに

言葉を用いる以上，人間の思考はその人間の所有する語彙の範囲を超えられるものではない．情緒力と論理的思考力を根底で支えるのが語彙力である(第 50 回大阪教育大学教育学会)．語彙力は文章の内容や，難易度などにもかかわっていると考えられ，言語処理分野では文章分類や難易度判定の基準としてそれを捉える工学的方法が考案されてきた．20 世紀前半より語彙の豊富さ指標が研究されており，現在に至るまで様々なモデルに基づいて数多くの指標が提案された．その中に，既存の語彙の豊富さ指標の逆数を取る方法で指標を改良したケースが二つある．それぞれ m (Michea, 1966 & 1971)と S (Sichel, 1975), M (Mass, 1966)と Uber (1978 & 1979)である．本研究は，語彙の豊富さ指標の逆数を取ることで指標を改良できるかという問題点を提起し，語彙の豊富さの新しい指標を提案する．

2 コーパス

今回の実験で用いたコーパスは表 1 の通りである．語彙の豊富さ指標に言語が影響する可能性を避けるため，日本語 32 人 32 篇，中国語 40 人 186 篇，英語 60 人 59 篇の作品を用い，その中に文学的文章ジャンル：小説，伝記，日記，脚本，随筆，紀行文が含まれている．

表 1　実験で用いたコーパス

言語	作家数	作品数	延べ語数	データの大きさ
日本語	32 人	32 篇	5,444,305 語	22.7MB
中国語	40 人	186 篇	4,517,617 語	17.1MB
英語	60 人	59 篇	3,353,832 語	17.7MB

3 先行研究の問題点

語彙の豊富さの測定における最も基本的な考え方は，延べ語数 N の中に占める異なり語数 V(N)の割合 TTR(Type Token Radio)である (Templin, 1957). TTR は文章の長さに依存すると指摘されているが，これまでに提案された語彙の豊富さ指標はほぼ TTR の改良型である．しかし，Vermeer (2004) は異なり語数と延べ語数の比を基に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した．少数の TTR をベースとしない語彙の豊富さ指標の中で, S は文章の長さに影響されにくい指標の一つとして推奨されている．次に示す m と S の式からわかるように, S は出現頻度 2 回の異なり語数割る延べ語数であり，m の逆数になっている．

$$m = \frac{V(N)}{V(2,N)} \quad (1) \qquad S = \frac{V(2,N)}{V(N)} \quad (2)$$

同じ例として, M と Uber が挙げられる．

$$M = \frac{\log N - \log V(N)}{(\log N)^2} \quad (3)$$

$$Uber = \frac{(\log N)^2}{\log N - \log V(N)} \qquad (4)$$

本節では，語彙の豊富さ指標の逆数を取ることで指標を改良できるかという先行研究の問題点について，検証実験を行った．

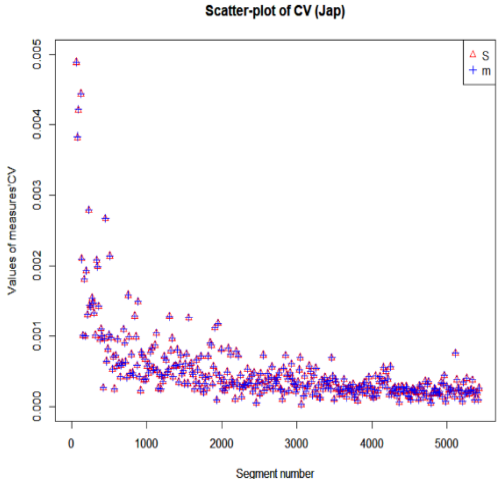
3.1 分析対象と分析データ

分析対象は m と S，M と Uber である．日本語，中国語，英語ごとにすべての文章を一つにまとめ，文章をシャッフルした．文章をシャッフルするとは，単語の順番をランダムに入れ替えることを指す．このように単語の順番をランダムに入れ替えるのは，各作家の影響を低減させ，大域的特性を概観するためである．シャッフルした文章について，先行研究 Tweedie and Baayen (1998)と木村・田中 (2011)を参考し，文章をチャンクごとに分けた．今回はチャンクのサイズを 1000 単語に設定し，一つずつ累加しながら，m と S，M と Uber の値を求めた．これにより，テスト文章が徐々に長くなっていったときの m と S，M と Uber の安定性を考察できる．

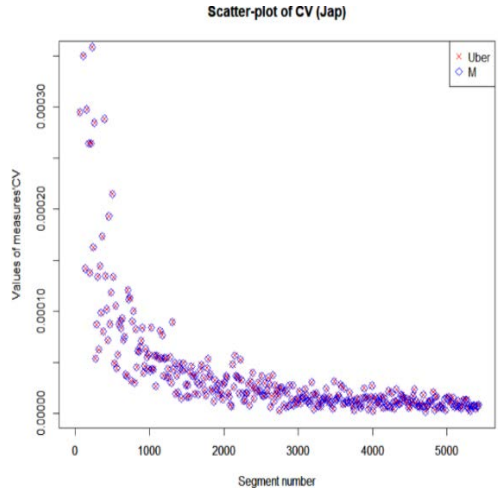
m と S，M と Uber の安定性を考察するため，用いたのは変動係数 (CV, coefficient of variation)である．変動係数は，平均値および単位の異なるデータのバラツキを比べる統計量であり，式を以下に示す．

$$CV = \frac{\sigma}{\bar{x}} \qquad (5)$$

変動係数を求める際，5 チャンクを一つの区切りとして一つずつ右にシフトした．紙面の都合上，中国語と英語の結果を省略し，日本語において，m と S，M と Uber の変動係数を散布図にした結果を図 1 に示す．その考察結果，言語にかかわらず，m と S，M と Uber の変動係数は重なっている．これについて数式上でも証明できる．したがって，語彙の豊富さ指標が文章の長さから受ける影響は逆数を取ることでは改良できない．



(a)



(b)

図 1　m と S，M と Uber の移動変動係数散布図

4 語彙の豊富さの新しい指標

鄭・金(2017)は 11 個の指標の比較評価の結果より，最も文章の長さの影響が小さい指標は s であることを報告した．今回は s を指数分布に代入することで得られた es を語彙の豊富さの新しい指標として提案する．

$$s = \frac{\log(\log V(N))}{\log(\log N)} \qquad (6)$$

$$es = \lambda e^{-\lambda s} \quad (7)$$

指数分布の確率密度曲線により，λが小さいほど曲線は緩やかになる．しかし，現実にλが無限に小さい値を取ることは難しく，また，λが無限に小さい値を取る必要があるかについて疑問である．次の節でλの設定について議論する。

4.1 分析方法

前節で得られたシャッフルした文章の中から，100 単語をランダムに抽出する試行を 5000 回行い，それぞれの TTR，es，es0.1，es0.01，es0.001，es0.0001，es0.00001 を算出した．ここで es はλ=1 時の値，es0.1 はλ=0.1 時の値を表すこととする．得られた結果をさらに z-score 標準化した．続いて，TTR と es，TTR と es0.1…に対し，ユークリッド距離とピアソン相関係数を求めた．毎回抽出した延べ語数は 100 単語であることは変わらないため，算出した TTR は正しく語彙の豊富さを表し，値が大きいほど語彙が豊富である．本研究ではユークリッド距離とピアソン相関係数二つの方面から総合的にλの設定を決める．TTR とのユークリッド距離が小さいほど，ピアソン相関係数が大きいほど，その指標を語彙の豊富さ指標として良いと評価する．

さらに，考察のため，求めたユークリッド距離とピアソン相関係数に対してデータ標準化を行い，プロットした結果を図 2 に示す．

4.2 考察

es は TTR とのユークリッド距離は小さいが，TTR との相関も小さく，使うべきではない．es0.0001 と es0.00001 は TTR との相関は大きいが，ユークリッド距離も大きいため，適切ではない．es0.01 と es0.001 については，言語にかかわらず，TTR とのユークリッド距離，ピアソン相関において近似している．es0.001 が es0.01 と似ているならば，es0.009，es0.008…es0.002 は es0.01 とより似ているはずである．したがってλが 0.01 より小さい値を取る必要はない．

残った es0.1 と es0.01 の中から選ぶ．es0.1 と TTR のユークリッド距離は小さいが，TTR との相関も小さい．es0.1 と es0.01 のユークリッド距離は 0.1067105 であり，ピアソン相関は 0.9999989 であることから，λは 0.01 まで取る必要がない．es0.01 が es0.1 と近似するならば，es0.09，es0.08…es0.02 はより es0.1 と似ていることが考えられる．以上により，λを 0.1 に設定する．

$$es = 0.1e^{-0.1s} \quad (8)$$

4.3 es は有効な指標になれるか

4.1 節で得られた es と TTR のピアソン相関，es のバラツキと TTR のバラツキの間に共に高い相関関係があれば，es は有効な指標であると判断できる．es と TTR の値の相関係数を求める際，指標値そのままを用いた．es と TTR のバラツキの相関係数を求める際は，5 チャンクを一つの区切りとして最初から一つずつ右にシフトしながら求めた変動係数を用いた．

その結果を表 2 に示す．es と TTR の値，es と TTR のバラツキ，共に高い相関が見られたため，es は有効な語彙の豊富さ指標であると説明できる．

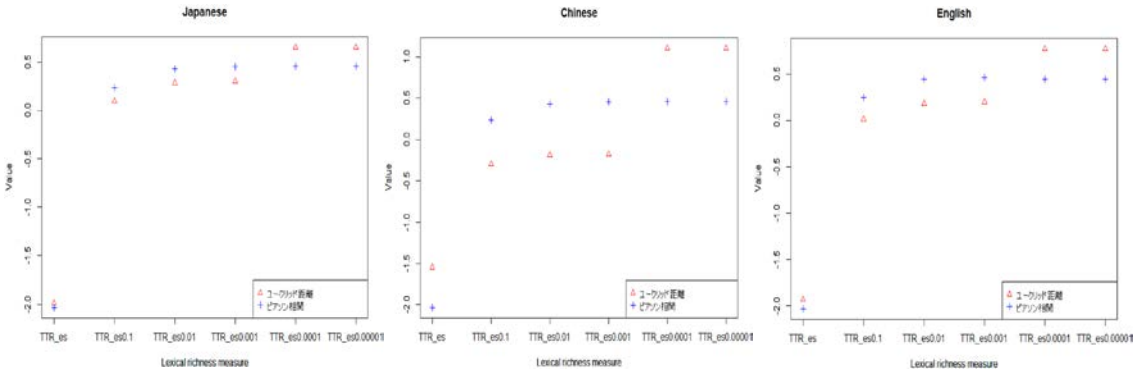


図 2　ユークリッド距離とピアソン相関係数の散布図

表 2　指標値とバラツキにおいて es と TTR の相関関係

	日本語		中国語		英語	
	指標値	変動係数	指標値	変動係数	指標値	変動係数
ピアソン相関係数	-0.9983	0.9965	-0.9989	0.9978	-0.9983	0.9965
p 値	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16

5 s と es を比較してみる

3.1 節の方法を用いて，チャンクのサイズを 1000 語に設定し，チャンクを一つずつ累加しながら s と es の値を求めた．さらに最初から一つずつ右にシフトし，5 チャンクごとに算出した変動係数の遷移曲線を図 3 に示す．図 3 で二つのことを考察する．まず，変動係数の大きさをみると，変動係数が小さいほどその指標のバラツキが小さく，文章の長さに影響されにくいといえる．次に，変動係数遷移曲線の傾きをみると，傾きが小さいほどその指標は一定のバラツキに収束している．図 3 の考察結果より，es の方が変動係数と変動係数の遷移曲線の傾きが共により小さく，es は顕著な改良効果が得られた．

6 まとめ

逆数を取ることで語彙の豊富さ指標を改良したという先行研究の問題点を提起し，新たな語彙の豊富さ指標 es を提案した．es の語彙の豊富さ指標としての有効性を検証したうえで，s と比較した結果から，es には一定の改良効果があることがわかった．

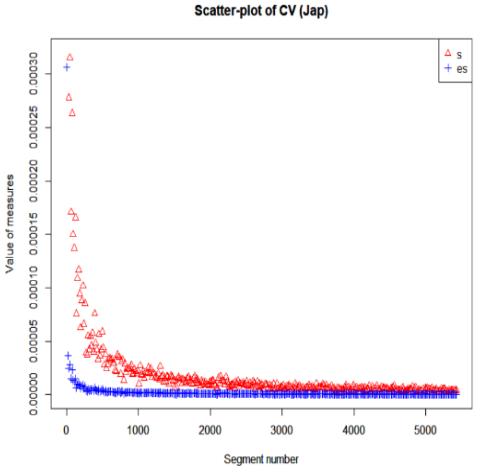


図 3　日本語において変動係数の遷移曲線

参考文献

[1] 木村大翼・田中久美子 (2011)，文章長に依存しない文章定数―Yule の K, Golcher の VM―，自然言語処理，Vol.18, No.2, 119-137.

[2] Tweedie, F. J. & Baayen, R.H. (1998), How Variable May a Constant be? Measures of Lexical Richness in Perspective, Computers and the Humanities, 32:323-352.

[3] 鄭　弯弯・金明哲 (2017)，語彙の豊富さの指標の評価に関する問題点と改善方法，　計量国語階学会第 61 回大会予稿集, 7-12.