語彙指標数値と文章主観評価の関係

――日本人大学生による２種類の書き言葉コーパスを使った実証研究――

田島　ますみ
深田　　淳
佐藤　尚子
玉岡　賀津雄

〈目　次〉

1　はじめに

1.1　言語パフォーマンスを測定する指標

外国語学習においてコミュニケーション能力の重要性が強調されるにつれ，評価方法も従来の客観テストに加えて言語パフォーマンスそのものを対象とする流れが出てきている．大規模なパフォーマンステストも開発されており，第二言語学習者と直接インタビューして口頭運用能力のレベルを判定する Oral Proficiency Interview （OPI） や，文章を書かせて採点する TOEFL のライティング・セクション，日本留学試験の日本語記述問題など，学習者がコミュニケーションを前提として産出する発話や文章がそのまま評価されることも増えている．一方，外国語教育や第二言語習得の研究分野では，学習者が産出した発話や文章を計量的に評価して，教育効果を示したり理論やモデル構築のための基礎的な分析に使用したりしている．コンピュータの活用によって，電子ファイル化した言語資料から数値的な情報が比較的簡単に多量に得られるようになっている状況も計量的研究の進展に寄与している．

このような流れの中で，言語パフォーマンスを測定する目的で計量的指標が用いられるようになってきているが，果たしてそれらの指標は適切な使われ方をしているのだろうか．言語パフォーマンスという複雑な事象を数値で単純化してとらえるのだから目安程度にと考えられて安易な使われ方がされていないだろうか．実際の計量的な研究を見ると，様々な指標が使用され，指標の使い方に研究者間で統一した見解や基準などは見当たらない．

これまでに指標の妥当性に焦点をあてた研究としては Wolfe-Quintero, Inagaki, & Kim (1998) の報告がある．この研究では，第二言語のライティングに関する39の量的研究で使用された108に及ぶ指標が，中間言語の発達を示す指標として妥当かという観点から検討された．結論として，流暢さ，

正確さ，複雑さの三つの範疇の中で最も適切な指標が暫定的に示されている．しかし多くの指標は第二言語としての英語を対象とした測定に使われたものであり，またスピーキングの測定は視野に入れていない研究であった．

本研究では，日本語を測定の対象とする際の語彙の豊富さを表す指標を取り上げる．言語学習において語彙の習得の重要性は論を待たない．また，昨今の日本人大学生の日本語力低下が問題視される中で，第二言語としての日本語を指導する日本語教師が，大学で日本人大学生の日本語力向上を目指す授業を担当することも多くなってきた．その第一言語としての日本語の教育実践の場では語彙の力がないという指摘をよく耳にする．問題の正確な把握のために，実際の言語資料を語彙の観点から数値化してとらえることも一つの具体的な方法となると考える．第一言語，第二言語の違いにかかわらず，語彙についての指標は言語パフォーマンス測定において重要なものである．本研究ではこの指標に絞って日本語のパフォーマンス測定に妥当なものであるのかを検討する．

1.2　語彙の豊富さを表す指標

語彙の豊富さを表す指標として最もよく知られたものは，延べ語数における異なり語数の割合（Type/Token Ratio, 以下 TTR）である．産出の全体の語数に対して異なる語がどの程度を占めるのかという割合であり，この値が大きければ多様な語が使われていて語彙が豊富であるということになり，小さければ同一語の繰り返しが多くて語彙が乏しいという解釈ができる．一般的に普及しよく知られた指標ではあるが，批判もあり，有効な指標であるかどうかには議論がある．

Wolfe-Quintero, et al（1998）は，この指標では短い作文で異なり語数が少ない場合と長い作文で異なり語数が多い場合で同じ値が出てしまい，両者を区別できないという問題を挙げた．彼らは第二言語のライティングに関する量的研究の結果を統合・検討した上で，Carroll（1967）の提案した改良型の指標がよりよい指標であると述べている．Carroll の指標は，TTR の分

母である延べ語数を操作し，産出された全体量への考慮を反映する．具体的には，分母を延べ語数の 2 倍の平方根とする．Wolfe-Quintero, et al (1998) が目的としたのは第二言語習得において中間言語の発達を測定するための最適な指標を特定することであった．中間言語が発達してくれば当然産出可能量が増えて長く書けるようになるが，TTR では延べ語数100語，異なり語数10語の作文と延べ語数1000語，異なり語数100語で同じ値が出る．Carroll の指標は，平方根を使うことで，長く書くと語彙の豊富さを表す指標が小さくなってしまう問題を緩和できるものである．

Malvern & Richards (2002) も，違うサンプルサイズ，すなわち長さの違う発話や文章において計算された TTR を比較することは妥当ではないと主張した．さらに，Carroll の指標も含めて，サンプルサイズの影響からは独立しているとして考案された改良型 TTR であっても，それらはやはり延べ語数の関数である点を指摘した．改良型であってもサンプルサイズの影響を逃れられないということである．Malvern & Richards は問題点克服のため，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使う指標を提案した．この D が語彙の豊富さを表す指標になる．実際の D を求めるには複雑な手順が必要だが，Meara & Miralpeix は D の数値を出すためのコンピュータソフトウェアを開発し，無償でダウンロード可能のフリーウェアとして公開している（http://www.lognostics.co.uk/tools/index.htm.）．50語以上ある英語のテキストであれば，簡単な手順で一瞬ではないがそれほど時間もかからず D を算出する．

Vermeer (2004) はこの D に関し，Jarvis (2002) の研究結果を引用して，明らかに語彙の違いのあるグループを識別できていない，語彙テストの点数と総合的評価の点数との相関が弱い，あるいは中程度しかないという問題点を指摘し，結局 D にしても TTR を使う指標であり，Type/Token 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結論づけた．Vermeer の TTR 系の指標に対する批判は，第一に語の難易度が考慮されていないということにある．新たな指標として考案されたのは，MLR (Measure

of Lexical Richness）で，難易度に従って分けられた 9 つの語彙リストを利用し，どのリストの語をどのくらいの割合で使っているのかをもとに数値を計算する指標であった．対象言語はオランダ語であり，9 つの語彙リストの難易度は日常生活の言語的インプットにおける頻度に基づいて分類された．成人の第二言語習得では語の頻度と習得順序の関係は年少者に比べて複雑であるが，初等教育段階の年少者であれば日常のインプットにおける頻度は習得順序と強く相関しているというのが，Vermeer の主張である．実際にオランダ語が母語である16人，第二言語である16人，計32人の年少者（平均年齢7.11歳）を被験者として指標の妥当性を検証し，MLR がテキストの長さに影響されない語彙の豊富さを表す指標であり，TTR 系の指標よりも妥当性が高い指標であることを示す結果を報告している．

語彙の豊富さを表す指標はこれらだけに限らない．Jarvis （2002）は，TTR を変形させた指標の中で広く使用されている三つを，Herdan の *C*，Guiraud の *R*，Uber 指標として挙げている．延べ語数と異なり語数を用いた計算式で算出する指標は，他にも Maas，Tuldava，Dugast らによって考案されている（金，2008）．Johnson によって推奨された MSTTR（Mean Segment Type/Token Ratio）は，テキスト全体を一定数の延べ語数を含む部分（segment）に分割してぞれぞれの TTR を出し，すべての部分の平均 TTR を出すというものである．Malvern & Richards (2002) は，この指標の長所を認めながらも問題点を指摘した．また，部分の平均をとらなくても，標準化した TTR を使うという方法もある（水本，2008）．さらに，延べ語数と異なり語数のみを使うのではなく，単語が用いられている回数（頻度スペクトル）を用いた指標がある．こちらにも TTR 系の指標同様にいくつかの指標があるが，代表的な指標は Yule の K 特性値というものである (Yule, 1944)．

このように語彙の豊富さを表す指標は数多く提案されているが，この状況はもっとも基本的な TTR が妥当な指標とはなりがたいという問題から発している．現状では多くの指標の中で特にこれといった決定的な指標は確立さ

れていない．

1.3　日本語での実証研究

日本語での語彙指標の妥当性に関する研究には，田島（2002），田島・深田・佐藤（2008）がある．田島（2002）は，第二言語としての日本語の発話を文字化したKYコーパスを資料として実証研究を行った．KYコーパスは，中国語・英語・韓国語を母語とする日本語学習者のOPIテープを文字化したもので，初級から超級まで4レベル，さらに各級で上中下のサブレベルが3段階あって，合計12段階のレベル判定がついているコーパスである（詳細は，「OPIを利用したコーパス」http://opi.jp/shiryo/ky_corp.html）．そこから各級15人，計60人分の発話データを用いて日本語学習者の話し言葉における語彙指標の値を検討した．この研究においてもTTRはOPIの級が上がるにつれて値が小さくなる傾向が見られ，先行研究で指摘されている問題が確認された．また，TTRの数値が有意な違いを示すのは，初級とそれ以外の級との間だけであり，しかも初級のほうがそれ以外の級よりも有意に指標数値が高く，語彙が豊富であることを示してしまうという結果であった．Carrollの指標の出す値のほうが級と連動して値が大きくなり，各級の違いも中級と上級，上級と超級の間の有意差は見られなかったが，他の級間では有意差が見られた．よって，単純なTTRよりもCarrollの修正TTRのほうがある程度識別力のある，より妥当な指標であることの統計的証左を提示したといえる．

田島・深田・佐藤（2008）では，第一言語としての日本語をデータとした．日本人大学生が授業の課題で提出した文章をデータとし，文章の主観評価と語彙指標の値の相関関係を検討した．傾向としては田島（2002）とほぼ同じで，TTRと主観評価との相関係数は小さく，ほとんど相関を示さなかったが，係数の方向としては負であった．Carrollの指標のほうが主観評価と中程度の相関が見られた．この研究は第一言語としての日本語をデータとして検証した．研究の動機として，Carrollの修正TTRが中間言語の発達

を測定・提示することには有効な指標であっても第一言語において文章の優劣を示す指標として機能するのかということへの疑問があったからである．語彙量，その中でも産出できる語彙について言えば，中間言語の発達段階においてはかなりの差を示す．初級学習者と超級学習者では明らかな違いがある．このような大きな違いは識別できても，ある程度の語彙量が期待できる成人母語話者が書いた文章の優劣を示す指標として機能できるのかには疑問の余地がある．実際，田島（2002）の結果でも，Carroll の修正 TTR が有意差を示したのは初級と各級間，および中級と超級の差であり，中級と上級，上級と超級の有意差は示さなかった．言語習得の比較的初期の段階にあっては妥当な指標であっても，習得が進むにつれて数値の個人差が小さくなり，ある一定以上のレベルに達すれば指標として機能しないという可能性も考えられる．このような研究動機から行った調査の結果は，データが第一言語としての日本語の文章であっても，第二言語としての日本語の発話データである KY コーパスを使った先行研究とほぼ同様の傾向を示して，Carroll の指標のある程度の妥当性を保証するものとなった．

1.4　本研究の目的

これまでの研究結果の示すところは，日本語の話し言葉や書き言葉の測定や評価においても，TTR は妥当な指標ではなく分母への操作を加える改良型の指標のほうが適切であり，測定や評価の際は後者を使うべきだということである．1.2 で述べたように，TTR は問題点が指摘されて新たな指標が続々と提案されてきている．しかし，新しい指標の難点としては複雑な手順を要し，数値を出すまでに労力と時間がかかるものが多いということである．Malvern & Richards の D はコンピュータで算出できるが，英語のテキストが対象である．Vermeer の提案した MLR もオランダ語の語彙リストに基づいて計算される数値であり，提案者自身も時間がかかる欠点は指摘している．日本語を対象として測定することを現実的に考えれば，比較的容易に値が得られる TTR 系の指標を使用することは利便性のあることであ

り，どの程度に妥当な指標で，どのような場合に有効なのかに関しての知見を得ておくことは言語パフォーマンス測定のための基礎となるだろう．

本研究では，前述の二つの研究に引き続き，日本語における TTR と Carroll の修正 TTR の妥当性を検討することを目的として，それらの指標数値と主観評価による点数との関係を分析した．田島・深田・佐藤（2008）では，「何かを紹介する文章」という記述的な文章をデータとした．今回はそれとは性質の異なる 2 種類の論述的な文章を扱うこととした．2 種類とも日本人大学生の書いた文章であり，第一言語としての日本語のデータである．まず一つは，大学の授業の学期末試験の答案として書かれたテキストである．試験の答案であることから，平易にわかりやすく記述することを期待される「何かを紹介する文章」に比べて，論述力が問われる種類の文章である．論理的な展開がされたり日常語彙を超える範疇の語彙が入ってきたりするという特徴が考えられる．もう 1 種類の文章は量の面での特徴を有する．ある意見に対して自分はどう考えるかを書く課題で，これも論述的な文章である．課題の中に下限・上限の字数制限があり，全体量がコントロールされる．となれば当然延べ語数もある程度範囲が決まってくる．TTR の問題点がテキストの長さの影響を受けるということであったので，この点，初めからテキストの長さが一定の範囲内である文章であれば，TTR の妥当性も変わってくるのではないだろうか．これら 2 種類の文章は，第一言語としての日本語の産出であるけれども，田島・深田・佐藤（2008）で扱った文章とは性質の異なる文章であると考え，性質の違いが結果に影響を与えるかどうかを調べることを目的として本研究の分析対象とした．

2　方　法

2.1　二つのコーパス

一つ目の資料は，国立 A 大学の2005年度に一般教養科目「生命科学」で

学期末に実施された論述試験の答案，40人分である．薬剤耐性菌に対する研究・技術開発の歴史的経緯と現状についての記述を求める問題であった．書いた学生は日本語母語話者で，１年生37人，２年生２人，３年生１人で，学部は多様である．手書きで大学指定の答案用紙に書かれたもので，答案の中に図が入っていたり箇条書きが混じっていたりして文章以外の要素があるものは除き，文章だけで論述しているものを40人分選んだ．

二つ目の資料は，2008年度に日本人大学生の日本語表現力向上を目的とした授業を受講した学生に，授業時間内に課題を指示して書かせた文章で48人分である．48人は全員１年生で，そのうち24人は国立Ａ大学の様々な学部に所属し，あとの24人は私立Ｂ大学の法学部に所属する．課題は日本語文章能力検定に過去に出題された問題を利用した．ある意見に対して自分はどのように考えるかを500字以上，760字以内で書けというものである．字数だけでなく構成についても条件があり，事実の報告，意見，意見の理由・根拠，異なる意見への反論の順に四つの部分で構成することが指示された．課題に対する文章は指定の原稿用紙に手書きで書かせ，40分を制限時間として提出させた．

これら２種類の手書きの文章をすべて電子ファイル化し，一つ目の論述試験の答案のファイル群をコーパス１，二つ目の意見文のファイル群をコーパス２とした．主観評価にはこれらのワードファイルを印刷したもの，語数のカウントにはテキストドキュメントのファイルを用いた．主観評価は印刷したものを用いたので，手書きの文章では評価に影響を与えるような表記上の観点，例えば，漢字が正確に書けていない，段落冒頭の１マス空けがはっきりしない，句読点やかぎかっこがマスの中の正しい位置に書かれていないなどの問題は今回の評価には含まれていない．

2.2　主観評価

各コーパスの評価は３人の評価者が行った．評価者は大学レベルでの日本語教育に経験のある教員４人で，コーパス１を評価者Ａ，Ｂ，Ｃ，コーパス

２を評価者 A，C，D が担当した．評価者の日本語教育歴は2008年 9 月現在の概算で，A が11年，B が25年，C が5.5年，D が11年で平均13.13年であった．これらの年数は第二言語としての日本語教育の経験年数であるが，このうち評価者 A は 4 年，B は 2 年，それぞれ大学で日本人学生の日本語教育も担当している．

評価は，総合的絶対評価で 3 段階の点数をつけた．試験問題と課題指示という要求に対して書かれた文章として，全体の質を検討し「優れている」ものを 3，「普通」であるものを 2，「劣っている」ものを 1 とした．まず評価者が各自で点数をつけ，その後各自の結果を突き合わせて 3 人で集まって討議するという二つの段階を設けて，点数が一致しないものも最終的には単一の点数を決定した．

各自で行った評価の点数の一致度を記しておく．コーパス 1 では，評価者 A と B，B と C，C と A の相関がそれぞれ0.68，0.44，0.67であった． 3 人の一致度は Spearman-Brown の公式を使う係数で0.71であった．コーパス 2 では，評価者 A と C，C と D，D と A の相関がそれぞれ0.53，0.52，0.53であった． 3 人の一致度は Spearman-Brown の公式を使う係数で0.67であった．高くない評価者間信頼性を補うため，両コーパスとも第二段階の討議の場では十分に時間をかけ，各自の評定根拠や基準を明らかにし， 3 人が納得した上で最終の点数（Score）を決定した．

2.3　語彙指標

本研究で取り上げる語彙指標は，延べ語数における異なり語数の割合（Type/Token Ratio ; TTR），延べ語数の 2 倍の平方根における異なり語数の割合（Type/$\sqrt{2\times\text{Token}}$; Carroll の修正 TTR）である．以降，便宜のため Carroll の指標を TTR2と記す．その他に指標数値を算出する際に必要となる延べ語数（Tokens），異なり語数（Types）についても結果を報告する．延べ語数，異なり語数の算出は，日本語形態素解析システム「茶筌」（奈良先端科学技術大学院大学松本研究室開発．http://chasen.naist.jp/hiki/

ChaSen/）を使用した形態素分析をもとに計算機処理によって行った．

2.4　統計分析

まずは主観評価の点数と語彙指標数値の相関関係を調べた．その後，回帰分析を行って指標が文章評価の点数をどの程度予測できるか検討した．

3　結果と考察

3.1　基本統計量

はじめに，コーパス１の主観評価の点数と指標数値に関する基本統計量を表１に，コーパス２のものを表２に示す．

表１．コーパス１の評点と指標の記述統計量（N＝40）

	平均値	標準偏差
Score	1.78	0.73
Tokens	249.93	75.64
Types	116.33	26.74
TTR	0.48	0.07
TTR2	5.21	0.57

表２．コーパス２の評点と指標の記述統計量（N＝48）

	平均値	標準偏差
Score	1.79	0.77
Tokens	292.40	44.69
Types	115.06	17.29
TTR	0.40	0.05
TTR2	4.76	0.54

基本統計量での顕著な違いは，延べ語数（Tokens）の項である．コーパス２はコーパス１に比べて延べ語数の平均値が大きいけれども，標準偏差は小さくなっている．コーパス２には字数制限があり，しかも「500字以上，

760字以内」という下限，上限を指定するものであったので標準偏差が小さくなったものと考える．その他の語彙指標ではコーパス２のほうでコーパス１よりも平均値がやや小さく，標準偏差も小さめとなっている．

3.2　コーパス１における評点と指標の相関関係

次に，主観評価の点数と指標数値との相関行列を示す．コーパス１の結果が表３である．評点と延べ語数，異なり語数，TTR との間に中程度の相関が認められる．しかし，評点と TTR の相関係数は負であり，評点が高くなるほど TTR が小さくなるという方向への相関である．また，評点と TTR2 の間の相関はほとんど見られなかった．

表３．コーパス１における評点と指標の相関行列（N＝40）

		Score	Tokens	Types	TTR	TTR2
Pearson の相関	Score	—				
	Tokens	0.43**	—			
	Types	0.30*	0.92**	—		
	TTR	-0.47**	-0.77**	-0.52**	—	
	TTR2	0.05	0.58**	0.84**	-0.03	—

*：５％水準で有意。　**：１％水準で有意。

今回の結果の中で TTR2に関しては，これまでの研究結果で示唆されてきたこととは異なる傾向を示している．TTR に関しては負の相関があるということで今までの結果の方向性と変わらないが，TTR2と主観評価の間に相関が見られないという結果は予想外であった．田島（2002）も田島・深田・佐藤（2008）も TTR を改良した TTR2は，日本語において第一言語，第二言語の違いにかかわらず，ある程度の妥当性があることを示してきた．第二言語の運用能力が上であったり文章として評価が高かったりすればTTR2も同様に高い数値となる傾向が確認されてきたわけである．それに反する結果となったが，これは何に起因するのか．

これまでに分析対象としてきたデータとコーパス１を比べてみると，トピ

ックの多様性についての違いが考えられる．先行の研究で扱ったのはOPIの発話と何かについて紹介する文章である．OPIではインタビュアーが様々なトピックの質問をして会話が進んでいくので，学習者1人分の発話の中にも数多くのトピックがある．紹介する文章のトピックは，「紹介するもの」であって単一であるが，何を選ぶかは書く側に委ねられている．実際，データとした文章も紹介しているものは様々であった．これらのデータに比べれば，コーパス1は論述試験の問題に対して書かれた文章であって，トピックは単一であり同一である．模範答案という形で期待される文章の理想形もある程度出題者の中にあるだろう．トピックはかなり限定された種類の文章であり，そのため使用される語彙もある程度決まってくるということが考えられる．ただちに一般化することはできないが，今回の結果は，トピックが限定される種類の文章では，語彙の豊富さを表すTTR2の値と文章評価の点数は相関しない可能性があるということを示している．

もう一点，注意を促しておきたい結果はTTRとTTR2の相関である．係数は-0.03でほとんど相関は見られない．二つの指標はどちらも語彙の豊富さを示す指標であるのにもかかわらず，この二つに相関がないというのは意外な結果であった．数値の意味から考えれば二つとも語彙の豊富さを表すはずの指標であるが，まるで違うものを測っているかのような相関のなさであった．

3.3　コーパス2における評点と指標の相関関係

コーパス2では，またさらに異なる結果が出た．表4がコーパス2における評点と指標数値の相関行列である．評点と延べ語数の間に相関は見られず，評点と異なり語数，TTR，TTR2の間にそれぞれ弱い相関が見られた．字数制限があるコーパス2では，今まで負の相関を示す傾向にあったTTRも正の方向での相関を示した．ただし，TTRもTTR2も相関の強さは弱めであった．

トピックという観点から言えば，コーパス2もKYコーパスや紹介する

表 4．コーパス 2 における評点と指標の相関行列（N＝40）

		Score	Tokens	Types	TTR	TTR2
Pearson の相関	Score	—				
	Tokens	0.08	—			
	Types	0.25*	0.70**	—		
	TTR	0.24*	-0.32*	-0.44**	—	
	TTR2	0.29*	0.27*	0.88**	0.82**	—

*：5%水準で有意。　**：1%水準で有意。

文章よりは限定的である．課題は「サービス業の人にお礼を言う必要はない」という意見に対しての自分の考えを書けというもので，意見を書く際にどのような具体的事実をもってくるかで語彙の違いが出るが，文章の主要トピックはやはり限定されている．コーパス 1 では，TTR2と評点との間に相関が見られなかった原因をトピックが限定されているためではないかと考察した．しかしながら，コーパス 2 の結果では，比較的トピックが限定されている文章でも，字数制限があって文章の全体量がコントロールされる場合は，TTR2と文章評価の点数との間に相関関係が期待できることを示した．

さらに興味深い結果は，これまで負の相関の傾向を見せて指標としては不適当という結果を出してきた TTR も，評点との間に正の弱い相関が見られたということである．コーパス 2 でこのような結果が出たことは，TTR の持つ問題が文章の長さからの影響であることをあらためて示している．文章の長さがある程度揃っている場合，TTR が指標として使える可能性が出てきたことになる．しかし，相関係数の値から言えば TTR2 のほうがわずかではあるが強めの値を出しているので，使う際の選択肢としては TTR2 を優先するべきであることに変わりはない．

また，コーパス 2 においては，コーパス 1 で相関のなかった TTR と TTR2に強い相関（0.82）が見られた．これらの結果から考えられるのは，トピックの限定性よりも文章の長さのほうが TTR 系の指標と主観評価との相関には影響を与えるということである．字数がコントロールされる場合，

TTR も指標としての妥当性が否定されない可能性が示唆された．

3.4　回帰分析

評点と指標数値の相関関係を見た上で，語彙指標が評点を予測できるかどうかを検証するため，語彙指標を説明変数，主観評価の点数を目的変数とする重回帰分析を行った．表 5 がコーパス 1 での結果である．

評点を予測する有意な説明変数はなかった．TTR も TTR2も評点の予測には有効ではないという結果だった．回帰式の決定係数は0.27で，かなり低くこのモデルを評点の予測に使うことは難しいと言える．また，説明変数間の強い相関がある場合に起こる多重共線性が存在していることがわかる．多重共線性の目安はトレランスが0.1以下，VIF は10以上とされており，これに相当する変数は分析から外したほうがよいとされる．今回の結果は TTR 以外のすべての説明変数で多重共線性が存在した．

表 5．コーパス 1 における評点を目的変数とした重回帰分析の結果

	非標準化係数	標準誤差	標準化係数	*t* 値	*P* 値	トレランス	VIF
定数項	-1.52	4.24		-0.36	0.72		
Types	-0.08	0.07	-3.00	-1.11	0.28	0.00	365.64
Tokens	0.02	0.02	2.35	1.33	0.19	0.01	156.70
TTR	-2.11	4.00	-0.19	-0.53	0.60	0.15	6.62
TTR2	1.57	1.72	1.21	0.91	0.37	0.01	86.79
R^2	0.27						
調整済み R^2	0.19						
F 値	3.27（$p=0.02$）						

よって重回帰分析ではなく，TTR，TTR2 をそれぞれ説明変数とする単回帰分析を行ってみた．表 6 と表 7 がそれぞれの結果である．決定係数は TTR が0.22，TTR2で0.00（0.0029）であった．TTR2 の回帰式は，すべての係数を 0 とする帰無仮説も棄却されず，決定係数も 0 に近い．評点の予

表 6．コーパス 1 における TTR を説明変数とする単回帰分析の結果

	非標準化係数	標準誤差	標準化係数	t 値	P 値
定数	4.23	0.75		5.63	0.00
TTR	-5.10	1.55	-0.47	-3.30	0.00
R^2	0.22				
F 値	10.87 （$p<0.01$）				

表 7．コーパス 1 における TTR2を説明変数とする単回帰分析の結果

	非標準化係数	標準誤差	標準化係数	t 値	P 値
定数	1.41	1.10		1.29	0.21
TTR2	0.07	0.21	0.05	0.33	0.74
R^2	0.00				
F 値	0.11 （$p=0.74$）				

測には役立たないという結果であった．TTR の回帰式の帰無仮説は棄却されるが決定係数は低く，この程度の決定係数の値では予測式としては使えない．また，TTR の係数が負であることも注意を要する点である．意味的には語彙が豊富でないほど評価が上がることを示す式となってしまうからである．

表 8 はコーパス 2 での重回帰分析の結果である．コーパス 1 の結果と同様，評点を有意に予測する変数はなかった．回帰式の決定係数は0.13でこちらも非常に低い．こちらはすべての説明変数で多重共線性が存在した．表 9 と表10が，TTR，TTR2をそれぞれ説明変数とする単回帰分析の結果である．決定係数は TTR が0.06，TTR2が0.09でいずれも低い．TTR2のほうで係数 0 の帰無仮説は棄却されるが，決定係数が非常に低いためこれも評点の予測に有効とは言えない．

表８．コーパス２における評点を目的変数とした重回帰分析の結果

	非標準化係数	標準誤差	標準化係数	t 値	P 値	トレランス	VIF
定数	-8.91	7.50		-1.19	0.24		
Types	-0.32	0.22	-7.15	-1.45	0.15	0.00	1252.67
Tokens	0.03	0.03	1.84	1.19	0.24	0.01	122.64
TTR	-48.52	48.51	-2.92	-1.00	0.32	0.00	429.44
TTR2	12.04	8.75	8.44	1.38	0.18	0.00	1940.51
R^2	0.13						
調整済み R^2	0.05						
F 値	1.60（p＝0.19）						

表９．コーパス２における TTR を説明変数とする単回帰分析の結果

	非標準化係数	標準誤差	標準化係数	t 値	P 値
定数	0.18	0.99		0.19	0.85
TTR	4.07	2.38	0.24	1.71	0.09
R^2	0.06				
F 値	2.93 （p＝0.09）				

表10．コーパス２における TTR2を説明変数とする単回帰分析の結果

	非標準化係数	標準誤差	標準化係数	t 値	P 値
定数	-0.20	0.96		-0.20	0.84
TTR2	0.42	0.20	0.29	2.08	0.04*
R^2	0.09				
F 値	4.31 （p＝0.04）				

4　おわりに

4.1　結果と考察のまとめ

本研究では，第一言語としての日本語の文章において主観評価の点数と語彙指標の数値との相関分析と回帰分析を行った．相関は，2種類のデータにおいて異なる傾向を示した．トピックが限定される文章の場合，これまで妥当な指標と考えられてきたTTR2は主観評価との相関が見られなかった．また，これまで問題が指摘されてきたTTRに，字数制限のある文章において主観評価との間に弱い相関が見られた．TTRも場合によっては使える指標となることが示唆されたが，相関の強さから見ればTTR2のほうが若干ではあるが高い係数を示した．値の算出も簡単なので，言語パフォーマンスの質を測定する指標を選択する場合には，こちらを優先的に選ぶべきである．

TTRに対するTTR2，すなわちCarrollの修正TTRの優位は確定されつつあるが，今回の回帰分析の結果を見る限り，この程度の相関では主観評価の点数を予測する指標とは言い難い．本研究で，TTR2であっても，トピックが限定される場合は指標として機能しないおそれもあることが示された．前述した他言語での試みのように，TTR系の指標とは別の新しい指標の検討が必要である．

4.2　今後の研究課題

TTR系の指標の問題点である，語の難易度を考慮していないということに対し，前述したVermeer（2004）は難易度に従って9つに分類した語彙リストを使う指標を提案した．この研究は年少者を対象として考えており，分類は日常生活での言語インプットの頻度を基にしている．Vermeer自身が

認めているように，成人の第二言語習得における語の難易度はもっと複雑である．語の難易度を反映する指標の開発には，信頼性の高い語彙の難易度の分類方法が必要となってくるが，それはたやすいことではない．

その方向とは別に，語の難易度を考慮する方法として，日本語の書き言葉を対象とする場合，漢字に関する数値に注目するということが考えられる．日本語の語彙の一般的な傾向として難しい語は漢字語であることが多い．語彙に関する指標として漢字に関する数値を利用することは一つの選択肢ではないだろうか．現在，本研究と同じグループで漢字指標についての研究も進めている．結果はまたあらためて報告したい．

また別の方向としては，単一の最適な指標を選び出すというよりも複数指標を組み合わせた回帰式を語彙の指標として用いることも考えられる．実際，第二言語としての英語を対象として，複数の語彙指標を説明変数とした回帰式が英作文の総合的評価を予測するかを検討した研究も出ている（水本，2008）．線形の回帰モデルで各説明変数の数値が出ていれば，複雑な式であっても数値の算出には時間はかからない．より妥当な数値を示す指標の選択肢として組み合わせ指標も視野に入れるべきであろう．

はじめにも述べたことであるが，言語パフォーマンスを測定することはコミュニケーション能力の重視とコンピュータの活用が進む状況の中で，ますます頻繁になり，重要視されていくであろう．その際に使用される指標の妥当性に対して無自覚であることは問題である．よりよい指標を選択することで評価や研究の精度は高まる．また妥当でない指標に基づく評価や研究は意味をなさない．これまでに述べた方向を含め，指標の妥当性に関する研究を今後も続けていく予定である．

謝　辞

本研究の文章評価に，今千春さん（千葉大学大学院生），藤原ゆかりさん（東京国際大学非常勤講師）のご協力を得ました．御礼申し上げます．

〈文献〉

金明哲（2008）．統計的テキスト解析（5）――統計法則と指標――『ESTRELA』172，60-65．

田島ますみ（2002）．KYコーパスを用いた語彙的複雑性の測定に関する研究――語彙的多様性及び密度と言語運用能力との関連――The tenth Princeton Japanese Pedagogy Workshop proceedings, 94-104．

田島ますみ・深田淳・佐藤尚子（2008）．語彙多様性を表す指標の妥当性に関する研究――日本人大学生の書き言葉コーパスの場合――『中央学院大学社会システム研究所紀要』9（1），51-62．

水本篤（2008）．自由英作文における語彙の統計指標と評定者の総合的評価の関係『学習者コーパスの解析に基づく客観的作文評価指標の検討』統計数理研究所共同研究リポート215，15-28．

Carroll, J. B. (1967). On sampling from a lognormal model of word-frequency distribution. In H. Kucera & W. N. Francis (Eds.), *Computational analysis of present-day American English* (pp. 406-424). Providence, RI : Brown University.

Hatch, E. & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics,* New York : Heinle & Heinle Publishers.

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing,* 19 (1), 57-84.

Malvern, D. & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing*, 19 (1), 85-104.

McNamara, T. (2000). *Language Testing*. Oxford : Oxford University Press.

Vermeer, A. (2004). The relation between lexical richness and vocabulary size in Dutch L1 and L2 children. In P. Bogaards & B. N. Laufer (Eds.), *Vocabulary in a Second Language* (pp. 173-189). Amsterdam : John Benjamins.

Wolfe-Quintero, K., Inagaki, S. & Kim, H-Y. (1998). *Second language development in writing : Measures of fluency, accuracy, and complexity*. Honolulu : University of Hawaii, Second Language Teaching & Curriculum Center.

Yule, G. U. (1944). *The statistical study of literary vocabulary.* Cambridge: Cambridge University Press.

Meara, P. M. & Miralpeix, I. D_Tools v.2. 0.
http://www.lognostics.co.uk/tools/index.htm.（2009年 3 月17日アクセス）