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

その他のタイトル	Relationship Between Lexical Indexes and Holistic Scoring by Raters in an English Essay
著者	水本 篤
雑誌名	統計数理研究所共同研究リポート215　「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」
ページ	15-28
発行年	2008
URL	http://hdl.handle.net/10112/12985

統計数理研究所共同研究リポート 215
「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」(2008), 15-28.

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

水本　篤
流通科学大学
E-mail:　atsushi@mizumot.com

あらまし　本研究では，自由英作文から得られるさまざまな語彙の指標のうち，初級レベルの大学生 EFL 学習者の自由英作文（エッセイ）の総合的評価と相関の強いものを調べ，重回帰分析により，総合的評価を予測する回帰式を作成し，その妥当性を検討することを目的とした。結果として，総語数以外では，使用語彙の難易度を示す指標や語彙の豊富さを示す指標などと，自由英作文の総合的評価の間に中程度以上の相関が見られた。重回帰分析では，各種指標を組み合わせることにより，総合的評価をある程度予測できる重回帰式が作成可能であることが明らかになった。

キーワード　自由英作文，語彙の統計指標，総合評価

Relationship Between Lexical Indexes and Holistic Scoring by Raters in an English Essay

Atsushi, MIZUMOTO
University of Marketing and Distribution Sciences (Assistant Professor)

Abstract　This study investigated the relationship between the holistic rating of English essays written by false-beginner level university EFL learners and the lexical indexes obtained from their essays. Multiple regression analysis was employed in order to create a predictive multiple regression equation with those lexical indexes. The results show that a moderately high correlation existed between the lexical difficulty and richness indexes and the essay score. It was also found that with multiple regression analysis, creating a predictive multiple regression equation by combining those lexical indexes would be possible to some extent.

Keyword　English essay, lexical indexes, holistic scoring

1. はじめに

1.1. 英語テストにおける自由英作文(エッセイ)

言語テスティングの歴史を見てみると，cloze テストをはじめ，受験者の統合的な英語習熟度を測定できると主張されている間接的な能力のテストが数多く開発されてきた。これは，ライティング，スピーキングなどのスキル（パフォーマンス）を測定するためには，採点者をトレーニングし，テストを実施するために場所を確保しなければならないなどといった，実用性（practicability）の観点からの打開策であったといえる。しかし，当たり前のことながら，測定したい能力があれば，実際にその能力を測定するテストを実施しなければならない。このような考え方から，日本では年間 140 万人以上が受験している TOEIC（Test of English for International Communication）でも，スピーキングとライティングのテストが，20007 年より，従来のリスニング・リーディングのみのテストに加えて実施されるようになっている。

ライティング能力の測定のためには，さまざまなタイプのタスクの中でも，自由英作文(エッセイ）を利用するテストが多い。大学入試問題でも，自由英作文が使用されるケースが多く見られるが，受験者が多くなればなるほど，採点に手間がかかってしまい，費用対効果が小さくなるため，自由英作文の評価（採点）をどうするかという観点は大きな問題になる。

1.2. 自由英作文の評価

自由英作文の評価は，大きく分類すると，総合的評価（holistic rating）と分析的評価（analytic rating）の 2 種類がある（山西, 2004)。このうち，比較的実用性が高く，多くのテストや評価で使用されるのは，総合的評価であり（Weigle, 2002)，1960 年代から始まった，自由英作文の自動採点システムにおいても，表面的な特徴（surface features）を用い，自由英作文の総合的評価の予測を行ってきた。石岡（2004）によると，現在では，自然言語処理（Natural Language Processing; NLP）やベイズ理論を利用した自動採点システムも開発されて，アメリカの経営大学院に進学する際に受験が必要な GMAT（Graduation Management Admission Test）の小論文の採点で使用されている e-rater は，人間の評価との一致度が 97%であると報告されている（Attali & Burstein, 2006)。自動採点システムでは，さまざまな説明変数を用いて採点を行う。その中でも最もよく使用される変数が石岡（2004）にまとめられているが，ほとんどが助動詞や代名詞などの単語の数であり，自動採点システムにおける説明変数でも語彙の占める役割は大きいことがうかがえる。また，総合的評価にしても，分析的評価にしても，語彙の観点が評価に含まれていることを考えると，自由英作文を評価する際には，どのような語彙が使用されているかに注目する必要があるといえる。

1.3. 使用語彙の豊富さを表す指標

自由英作文の評価では，総語数（Tokens），平均語長などの表面的な特徴を用い，自動化のために人間の採点の予測を行ってきたが，使用語彙の豊富さを表す指標によって，ライティング能力以外の能力を予測する研究は，これまでにそれほど多く行われていない。

石川（2005）は学習者の自由英作文（エッセイ）を，総語数（Tokens），語種数（Types），標準化 TTR（Standardized Type/Token Ratio），Guiraud（Type を Token の平方根で割った値），平均語長（Mean Word Length），センテンス数（Sentences），センテンスあたりの平均語数（Mean Words/Sentences）から分析し，これらの指標のうち，標準化 TTR が受容語彙力と中程度の相関を示すことを明らかにした。

小泉（2007）は，スピーキング・テストにおける発話を分析し，内容語の密度，TTR，Guiraud，D（TTR の値は語数が多いと低くなるので，その限界点を TTR はモデリングが可能であることから克服した指標； Malvern & Richards, 2002）などの使用語彙の豊富さの指標を調査し，D と内容語の密度が語彙の多様性を表す妥当な指標であることを証明した。

このように，使用語彙の豊富さを表す指標はさまざまな能力に関連があると考えられるため，自由英作文の評価を自動化すると考えた場合にも使えるものがあるのではないかと思われる。

1.4. 本研究の目的

自由英作文の自動採点システムを構築するには，専門的知識のみならず，開発・実施のコストも莫大なものとなるだろう。そこで，本研究では，自由英作文から比較的簡単に得られるさまざまな客観指標のうち，初級レベルの大学生 EFL 学習者の自由英作文（エッセイ）の総合的評価と相関の強いものは何であるのかを調査し，教室レベルやプレイスメント・テストなどでの適応を考え，重回帰分析によって，総合的評価を予測する回帰式を作成し，その回帰式がどの程度使えるものであるのかを検証することを目的とした。

2. 方法

2.1. 参加者と測定道具

実験参加者は，関西の私立大学の 2 回生（商学部）を中心とした 40 名の EFL 学習者であり，英語習熟度は TOEIC 300 点台前半の false-beginners（疑似初心者）が大半であった。参加者の英語習熟度を調査する目的で，TOEIC 形式の模擬試験（リスニングセクション 25 問，リーディングセクション 25 問の計 50 問の 4 択問題で構成されているもので，以下，「TOEIC 模擬テスト」と呼ぶ）を実施した。その結果の基本統計量を表 1 に，ヒストグラムを図 1 に提示する。

表 1　TOEIC 模擬テストの基本統計量（N= 40）

平均 (Max = 50)	19.60
最小値/最大値	7 / 30
標準偏差	5.67
歪度	-0.11
尖度	-0.96
測定標準誤差(SEM)	3.15
信頼性係数(α)	.69

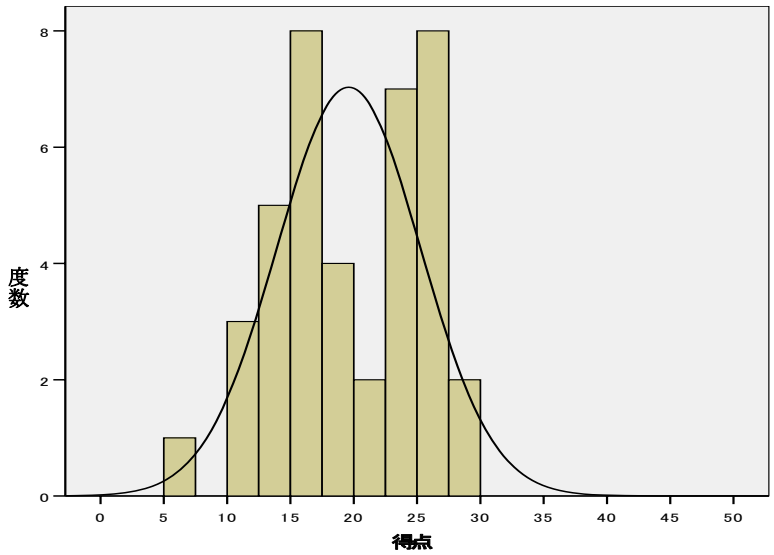

図 1　ヒストグラム

内的一貫性を表す信頼性係数の 1 つであるクロンバックのαの値が多肢選択形式であるにも関わらず，低い値（.69）であった理由は，50 点満点中のテストの最低得点が 7 点で，最高得点が 30 点という，比較的狭い範囲にテスト受験者の得点が集中したためであると考えられる。靜（2007）は「信頼性は，得点の順位の安定性の指標だと考えるのが最も誤解がないと思います。観測得点に含まれている誤差得点が小さければ，仮にもう 1 度同じテストを実施したときに，似たような得点順位が得られる可能性が大きいでしょう。また，誤差得点がそれほど小さくなくても，もともとの実力差が，誤差が問題にならないほど大きく離れていれば，何度実施しても順位の逆転は起こりにくいはずです。」（p.117）としている。ゆえに，今回の参加者のように，テストの得点（英語能力）が比較的狭い範囲にある場合には，順位の逆転が起こりやすいと考えられるため，信頼性係数も低くなってしまう。このような理由から，この信頼性係数は許容範囲であり，今回実施したテストを英語

能力の指標として利用することに問題はないと判断し，以降の分析で英語能力を示す指標として扱うこととした。

実験参加者のライティング能力を測定には，授業時間内に自由英作文を以下のトピックで書くように指示し，提出されたものを用いた。

トピック

Do you agree or disagree with the following statement?

"It is important for college students to have a part-time job."

Use reasons and specific details to support your answer.

作成の際には，パソコンのワープロソフトで 150 語以上，制限時間を 60 分，辞書の使用は禁止と設定し，本論文の著者の監督の下，授業時間内に提出することを義務づけた。語数は「150 語以上」を自由英作文作成時の条件としていたが，規定語数を満たせない者もいたため，そのような場合には制限時間内に書けた語数を対象とした。その後，提出された自由英作文を JACET8000 に収録されている v8an というプログラムを使ってレマ化と頻度付与を行った。このプログラムは JACET8000 に基づいて，テキストの中で使われている語をレマ化し，レベル 1（1000 位までの 1000 語）からレベル 8（8000 位）までのランク付けとテキストカバー率を算出するものである。JACET8000 を基にしたランクの分類は表 2 のようになる。JACET8000 のリストに含まれていない語は over8 となり，その他，省略形（cont. forms）や固有名詞（proper nouns），そして数字などの語としてはカウントされないもの（non-words)がこのプログラムによって自動的に分類された。総語数（total token）は 6944 語で，実験参加者 40 名の平均語（average）が 147.74 語であった。使用語彙はトピックで提示されていたもの（例えば，important，students，job など）の他には上位 100 位以内に，money，many，friend，good，experience，social，learn，parent，earn，useful，future，meet，person，company，chance などの語があったため，「アルバイトをすることにより，友人を作り，人と出会い，社会的な経験を積むことができる。また，親からお金をもらうのではなく，自分でお金を稼げるため，役に立つと考えられるものなので重要ある。」という意見が多かったことがわかる。

表 2　v8an による自由英作文の分析結果

	L 1	L 2	L 3	L 4	L 5	L 6	L 7	L 8	over8	cont.	prop.	non.	Total
Token	6149	204	73	73	19	26	2	19	267	34	45	33	6944
%	88.55	2.94	1.05	1.05	0.27	0.37	0.03	0.27	3.85	0.49	0.65	0.48	100

2.2. 自由英作文の評価

自由英作文の評価は，本論文の著者（評定者 A）と，中学校で 3 年間の英語指導経験を有する外国語教育学専攻の大学院生（評定者 B）の 2 名で行った。評価基準は TOEFL（Test of English as a Foreign Language）の TWE（Test of Written English），および現在の TOEFL iBT（Internet-Based Testing）の実施以前に，2000 年から 2006 年にかけて利用されていた TOEFL CBT（Computer-Based Testing）の ライティング・セクションの評価基準であった 0 点から 6 点の総合的評価尺度（holistic scoring）を 10 段階にしたものを使用した（表 3)。点数の配分を 0 点から 10 点の範囲に変更したのは，対象とした学習者のレベルが低いため，点数にばらつきが出にくく，統計分析の際に必要な情報となる分散が小さくなると考えられたためである。評定者が 2 名以上の場合には，評価者の基準の個人差が採点に影響を及ぼすと考えられるため，今回の実験参加者と同レベルの学習者 5 名が書いた自由英作文を評定者 A が先に採点し，採点者 B がそれに合わせるという形の評定者トレーニングを行った上で，参加者 40 名分の自由英作文を採点した。

表 3　本研究で使用した総合的評価尺度の概略

点数	採点基準
10 点	構成，構文，文法的な側面からほぼ完璧
8-9 点	構成，構文，文法的な側面から若干の誤りが見られる
6-7 点	構成，構文，文法的な側面から多くの誤りが見られる
4-5 点	構成，構文，文法的な側面から改善すべき問題が多く見られる
2-3 点	英作文を通した意志伝達の能力だけ持つ
1 点	英作文を通した最小限の意志伝達能力を持つ
0 点	答案を提出しない，またはエッセイのタイトルと違った内容を作文する

Note. 実際には TOEFL で使用されていた詳細な採点基準を使用した。

表 4 は 2 名の評定者の採点結果の基本統計量をまとめたものである。この結果からわかるように，評点者 A は比較的厳しい採点（$M = 2.98$, $SD = 1.46$）を行い，評定者 B はそれに比べて甘めの採点（$M = 4.53$, $SD = 2.62$）であったことがわかる。2 名の評定者の採点の平均値の差を検討するために，対応なしの t 検定（両側検定）を行ったところ，$t(78) = -3.27$, $p < .01$, $r = .35$（中程度の効果量）で平均値に有意差が見られた。しかし，評定者間信頼性を示すピアソン相関係数は $r = .81$ であったので，相対的には自由英作文の評価に対する観点，つまり自由英作文の上手い者には高い点数を，下手な者には低い点数を 2 人の評定者がともに与えていたということがわかる。そのため，2 名の評定者の採点の平均点（平均評定）を以降の分析では用いることとした。

表 4　2 名の評定者の採点結果の基本統計量（N= 40）

	評定者 A	評定者 B	平均評定
平均値	2.98	4.53	3.75
最小値/最大値	1 / 7	1 / 10	1 / 8
標準偏差	1.46	2.62	1.95
歪度	0.88	0.60	0.64
尖度	0.54	-0.72	-0.53

2.3. 分析手法

はじめに，TOEIC 模擬テスト（習熟度テスト），そして評定者によるエッセイの総合的評価（平均）に加えて，自由英作文から機械的に得られる以下のような語彙の指標を用い，変数間の相関分析を行った。

(1) TOEIC 模擬テスト（習熟度テスト）
(2) エッセイの述べ語数（Tokens；総語数）
(3) エッセイの TTR（Type/Token Ratio；総語数における異なり語の割合）
(4) 標準化 TTR（Standardized TTR：TTR を 50 語で標準化したもの）
(5) 平均語長（Average Word Length）
(6) D (Meara & Miralpeix, 2004 による D_Tools を使用)
(7) Flesch-Kincaid Grade Level (リーダビリティーの指標)
(8) JACET8000 における難しい語の使用割合指標（異なり語数×8 段階のレベル）
(9) 評定者によるエッセイの総合的評価（平均）

「(8) JACET8000 における 8000 語の使用割合指標」は，JACET8000 の 1000 語単位のレベル分けにおいて，難しいレベルが使われれば使われるほど難易度の高い語彙を使用しているという考えを基に計算された。例えば，ある学習者の自由英作文の中の使用語彙が JACET8000 における level 1 での使用語が 49 語，level 2 が 3 語，level 3 が 1 語，Level 4 が 1 語という場合には，(1×49) + (2×3) + (3×1) + (4×1) = 62 という値が得られる形で，レベルの高い語には重み付けがなされるようにした。加えて，上記の変数がどのような関係にあるのかを調べるために，変数間の近さをクラスター分析により検証した。

また，今回の研究のもう 1 つの目的である，「総合的指標を予測する回帰式を作成し，その回帰式がどの程度使えるものであるのかを検証する」については，自由英作文の総合的評価を目的変数とし，今後利用される可能性の低い TOEIC 模擬テスト（習熟度テスト）以外の指標を説明変数として，重回帰分析を行った。また，重回帰分析から得られた回帰式がどの程度適応できるものなのかを検討するために，算出される予測得点を，オンライン

自動採点システムの Criterion[1]の採点結果と比較することにした。そして，回帰式から算出される予測得点が実測値から大きく外れた実験参加者については，その理由を質的に分析した。

3. 結果と考察

3.1. 各変数の関係

表 5 は今回の研究で使用した各指標の相関行列を示している。まず，TOEIC 模擬テストとの相関を見てみると，自由英作文との相関が $r = .80$ とかなり高くなっている。これは，英作文能力の高いものは習熟度（TOEIC 模擬テスト）でも高い点数になる（逆の関係でも同じ）ということを表している。また，各種語彙指標の中では，Tokens が一番高い相関係数（$r = .57$），で，次いで，J8 Index（$r = .49$），D（$r = .41$），Standardized Type/Token（$r = .39$）という順番になっている。これらの結果から，学習者の習熟度の高さはこのような指標に反映されるのではないかと考えられる。しかし，対象とした学習者のレベルが低かったために，単純に「レベルが高い学習者は他の学習者に比べてたくさん書ける」ということを表しているだけかもしれないことに注意しなければならない。

本研究の目的である，「自由英作文から比較的簡単に得られるさまざまな客観指標のうち，初級レベルの大学生 EFL 学習者の自由英作文（エッセイ）の総合的評価を予測するのに役立つものは何であるのかを明らかにする」について，総合的評価と他の変数との相関係数を確認してみると，Tokens が一番高い相関（$r = .72$）を示し，J8 Index（$r = .62$），D（$r = .48$），Standardized Type/Token（$r = .40$）という順番で続き，Flesch-Kincaid Grade Level（$r = .32$）などに中程度以上の相関が見られた。Type/Token Ratio と 総合的評価の相関係数は $r = .01$ で，無相関であったが，標準化した TTR である Standardized Type/Token と総合的評価では $r = .40$ であり，総語数（Tokens）と総合的評価では $r = .72$ であった。Type/Token Ratio は語数が多いと低くなることが問題点として指摘されており，それのみを指標として，総合的評価との関係を予測するのは危険であることが以上の結果よりわかる。また，Ave. Word Length と総合的評価の相関係数も $r = .01$ で非常に低い値であった。これは，1 語の文字数が多い語をたくさん使ったとしても，総合的評価にはそれほど影響しない可能性があることを示していると考えられる。

[1] Criterion についての説明は http://www.cieej.or.jp/toefl/criterion/index.html を参照。

表 5　変数間の相関行列（N= 40）

	1	2	3	4	5	6	7	8	9
1. TOEIC 模擬テスト	—								
2. Tokens	.57	—							
3. Type/Token Ratio	.00	-.21	—						
4. Standardized Type/Token	.39	.45	.57	—					
5. Ave. Word Length	-.05	.03	.18	.31	—				
6. D	.41	.54	.56	.80	.27	—			
7. Flesch-Kincaid Grade Level	.24	.11	.39	.34	.66	.37	—		
8. J8 Index	.49	.79	.32	.73	.09	.77	.23	—	
9. 総合的評価	.80	.72	.01	.40	.01	.48	.32	.62	—

Note. それぞれの変数の略称の説明は次のとおり。2. Tokens = エッセイの述べ語数，3. Type/Token Ratio = エッセイの TTR（総語数における異なり語の割合），4. Standardized Type/Token = 標準化した TTR，5. Ave. Word Length = 平均語長，6. D (Meara & Miralpeix, 2004 を参照)，7. Flesch-Kincaid Grade Level = リーダビリティーの指標，8. J8 Index = JACET8000 における 8000 語の使用割合指標（異なり語数×8 段階のレベル），9.総合的評価 = 参加者の書いた自由英作文の総合的評価（2 名の評定者の平均値）

相関行列で相関が高いもの同士が何であるかを確認するだけでも，傾向を確かめることができるが，全体的にどの指標とどの指標が近い関係にあるのかを図示するために，それぞれの指標における数値を標準化した後に，クラスター分析（ワード法，平方ユークリッド距離）を行った結果を樹形図（デンドログラム）で表したものが図 2 である。磯田（2004）は，「傾向の似ていないもの同士がつなげられる場合，結合距離が遠くなります。したがって，結合距離が大きく跳ね上がる，つまり，横の線が長くなるところを探すことが方策のひとつです」（p.118）としている。今回の例では，結合距離が 5 を超えた辺りで横の線が長くなっているため，図 2 にあるように，縦線を入れてみて，関係が近いものが何であるのかを検討した。結果は表 6 の相関行列と同じものであるが，標準化した TTR である Standardized Type/Token と D はとても近い指標であり，元となっている Type/Token Ratio と少し離れた位置ではあるが近いものであることがわかる。また，平均語長（Ave. Word Length）と リーダビリティーの指標である Flesch-Kincaid Grade Level は近い指標であることがわかる。Flesch-Kincaid Grade Level の計算には，以下のような式（Microsoft Word の説明による）が使用されている。

Flesch-Kincaid Grade Level = (.39 × ASL) + (11.8 × ASW) − 15.59
ASL = 文章の平均の長さ（文章数で割って得られた単語数）
ASW = 1 単語あたりの平均音節数（単語数で割って得られた音節数）

「1 単語あたりの平均音節数」は平均語長（Ave. Word Length）とも関係があると考えら

れるので，この 2 つが近い指標であるという結果は直感的にも納得できるものである。

TOEIC 模擬テスト（習熟度テスト）と評定者によるエッセイの総合的評価とデンドログラムで近い位置にあるのは，Tokens（総語数）と J8 Index（JACET8000 における難しい語の使用割合指標）であり，今回の実験参加者（N= 40）の場合には，「総合評価で高い点数がつけられる自由英作文は，総語数が多く，使用語が難しいものであるもの」という傾向があったことがわかる。

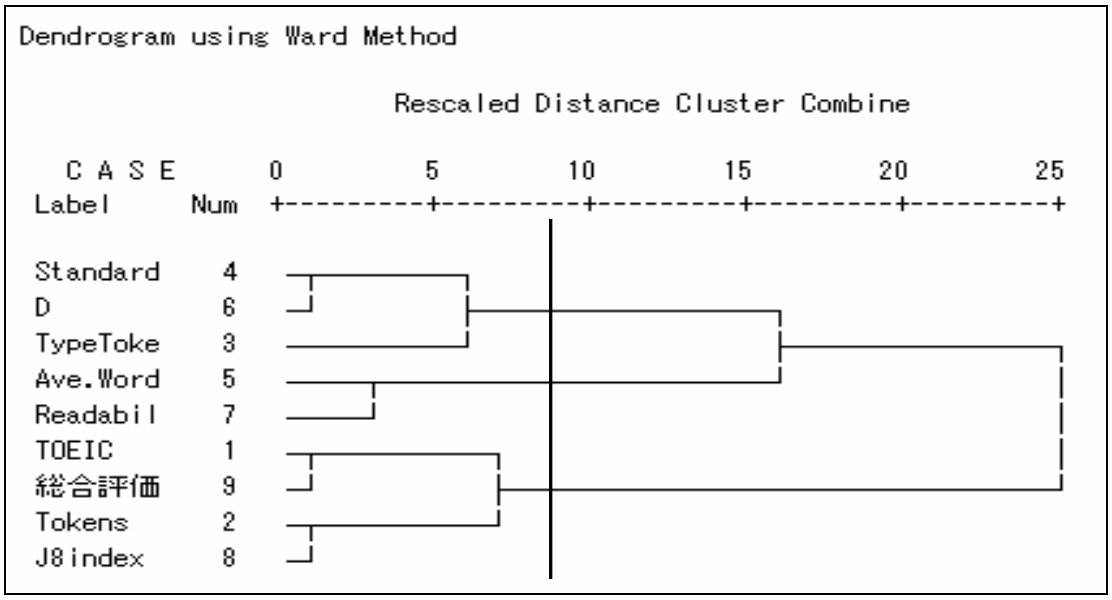

図 2　クラスター分析の結果（ワード法，平方ユークリッド距離）

3.2. 重回帰分析による回帰式の作成

自由英作文の総合的評価を目的変数とし，TOEIC 模擬テスト（習熟度テスト）以外の指標を説明変数として，重回帰分析（ステップワイズ法）を行った。川本（2004）は，重回帰分析を行う際の変数選択の一般的な前提（p.132）として，(1) 目的変数と相関の高い変数を選ぶ，(2) 説明変数相互で相関の高い変数がある場合は，いずれかの変数を除外する，という 2 点を挙げている。

目的変数である総合的評価と，説明変数である他の指標との相関係数（表 5）を確認してみると, Tokens，J8 Index，D，Standardized Type/Token，Flesch-Kincaid Grade Level などにある程度の相関が見られ，Type/Token Ratio と Ave. Word Length が無相関という結果であったが，「重回帰分析を行うときには，相関の低い説明変数でも加えてみると，示唆に富んだ結果が得られる可能性がある」（前田, personal communication, December 10, 2007）ので，すべての指標を説明変数として，重回帰分析を行うことにした。

また，重回帰分析において，説明変数相互で強い相関がある場合には多重共線性

（multicollinearity）が存在している可能性があるが，表 5 を確認してみると，一番相関の強いものとして，D と Standardized Type/Token が r = .80 であり，それ以外は問題となるほど強い相関が存在していないと判断したので，すべての指標を利用することとした。また，本研究の重回帰分析で使用した統計パッケージの SPSS では，多重共線性の判断となる情報である，許容量（tolerance）と VIF（Variance Inflation Factor; 分散増大要因）を出力することができるので，それらの情報も参照した。

表 6　自由英作文の総合的評価を目標変数とした重回帰分析（ステップワイズ法）の結果

	非標準化係数		標準化係数			共線性の統計量	
	B	標準誤差	*B*	*t*	*p*	許容量	VIF
(定数)	5.99	3.93	—	1.52	.14		
Tokens	0.028	0.04	0.68	6.68	.00	.98	1.02
Flesch-Kincaid Grade Level	0.66	0.20	0.44	3.25	.02	.55	1.80
Ave. Word Length	-2.50	1.10	-0.31	-2.28	.03	.56	1.78
調整済み $R2$ 乗	.60						
F 値	20.50 (p < .01)						

表 6 は自由英作文の総合的評価を目的変数とした重回帰分析（ステップワイズ法）の結果である。自由英作文の予測得点を求める回帰式の回帰係数としては，Tokens，Flesch-Kincaid Grade Level，Ave. Word Length の 3 つが使用された。つまり，今回の実験参加者のレベルの学習者が書く自由英作文の予測得点は，以下の式で求めることができる。

自由英作文の予測得点

= 5.99 + 0.028×(Tokens) + 0.66×(Flesch-Kincaid) - 2.50×(Ave. Word Length)

回帰モデルの適合度を調べた分散分析では，$F(3, 36) = 20.50$, $p < .01$ であり，回帰モデルの適合度は有意であった。ゆえに，この重回帰式は予測に役立つといえる。重回帰分析では，モデルの説明力の指標として，決定係数と呼ばれる $R2$ 乗が求められる。この $R2$ 乗値は，説明変数の個数を増やせば，値が大きくなってしまう傾向があるので，調整済み $R2$ 乗が使用される。今回の重回帰分析で得られたモデルの説明力は，調整済み $R2$ 乗で .60 だったので，これらの指標を使って自由英作文の得点の 6 割は説明できるということがわかった。多重共線性については，VIF が 10 以上，許容量が 0.20 以下の場合には，多重共線性が生じていると判断し，その変数を除去したほうが良いという基準があるが（Field,

2005)，表 6 の結果ではそのような変数は見当たらなかった。以上の結果から，Tokens, Flesch-Kincaid Grade Level，Ave. Word Length の 3 つを利用することによって，自由英作文の得点の 6 割が予測できる回帰式を作成することができたといえる。

ちなみに，重回帰分析の結果では，説明変数が目的変数に及ぼす影響の大きさの解釈に注意しなければならない。具体的には，重回帰分析における，「ある説明変数の偏回帰係数（標準偏回帰係数）の大きさは，他の説明変数の影響を一定にした場合に，目的変数にどれだけ影響を与えるかの指標である」（Field, 2005, p. 192）ため，それぞれの説明変数がどれだけ目的変数に影響を及ぼすかは，偏回帰係数（標準偏回帰係数）の大きさでは判断できないという事実を覚えておかなければならないということである。

前田（2004a）は，「相関分析をした場合の相関係数と，重回帰分析における独立変数から従属変数への影響の強さを表す標準偏回帰係数とが大きく異なってしまう場合があります。これは，複数の独立変数に共通して従属変数を予測できる割合が存在するために，最も予測力が強い独立変数の偏回帰係数以外の偏回帰係数は小さめに算出されたり負の値になってしまったりするからです。」（p.78）としており，重回帰分析は目的変数と説明変数の相関の強さの証明に用いるのではなく，目的変数の予測のためだけに使うべきであると主張している。

実際に，今回のデータでも，相関分析では，総合的評価と，Tokens （r = . 72）を示し，J8 Index（r = .62），D（r = .48），Standardized Type/Token（r = .40）などに，総合的評価とある程度の相関が見られたが，重回帰分析では Tokens （β = . 68）以外は重回帰モデルには含まれなかった。しかし，重回帰モデルに含まれなかったからといって，実質科学的に意味のある関係がないというわけではなく，特に，ステップワイズ法を用いた重回帰分析では，最もよく目的変数を説明する変数を選択して重回帰式を作成するために，十分に説明ができるモデルのできた場合には，それ以外の説明変数は相関が強いものでもモデルに組み込まれないということが起こるので，その他の説明変数がなぜ含まれなかったかという情報は，重回帰分析では知ることはできないということに注意する必要がある（前田, 2004b)。

3.3. 予測得点の妥当性検証

3.3.1. 自動採点システムとの比較

重回帰分析から得られた回帰式がどの程度適応できるものなのかを検証するために，実験参加者 40 名から 9 名を Excel の乱数関数を使ってランダムに選び, オンライン自動採点システムの Criterion の採点結果と比較することにした。表 7 はその結果であり，太字下線を付した 3 件以外は，予測得点と，Criterion の採点の結果が一致しているということがわかる。これにより，今回の研究で作成された回帰式による予測得点には，ある程度の妥当性があることがわかる。

表 7　自由英作文の予測得点と Criterion の採点結果の比較

参加者	予測得点	Criterion
1	2	2
2	2	2
3	2	**3**
4	2	**3**
5	2	**3**
6	3	3
7	3	3
8	3	3
9	3	3

Note. 予測得点は 10 点満点を 6 点満点に換算したものを使用

3.3.2. 予測得点と総合的評価に差がある学習者の自由英作文

次に，回帰式から算出される予測得点が実測値から大きく外れた実験参加者について，その理由を質的な観点から分析した。そのために，予測得点と総合的評価（実測値）の残差を求め，残差が±1.5 以上となった実験参加者 8 名（表 8）を対象に，これらの学習者の自由英作文の特徴を調べた。

表 8　予測得点と実際の評価の残差が±1.5 以上だった参加者

参加者	Tokens	Flesch -Kincaid	Ave. Word Length	予測得点	総合的評価	残差
1	184	7.2	4.38	4.94	7.50	-2.56
2	155	5.6	4.13	3.70	6.00	-2.30
3	161	5.4	4.11	3.79	6.00	-2.21
4	160	7.1	4.1	4.91	7.00	-2.09
5	182	7.2	4.23	5.26	7.00	-1.74
6	163	3.9	3.68	3.93	2.00	1.93
7	162	7.1	4.25	4.59	2.50	2.09
8	158	5.9	4.46	3.16	1.00	2.16

まず，残差がマイナスになっている，予測得点が実際の総合的評価よりも低いものは，文法・構文や論理・構成がまとまっているもので，Tokens，リーダビリティー，平均文字数などに関係がある語数，文の数，音節数では違いがわからないため，作成された回帰式では，このような特徴は予測得点の算出に反映できなかったと考えられる。反対に，残差がプラスになっている，予測得点が実際の総合的評価よりも高い実験参加者の書いた自由英作文は，文字数がある程度あっても，語法・文法・構文に間違いが多く，言いたいことが伝わらないもので，こちらも，語数，文の数，音節数などでは違いがわからない特徴であった。これらの結果から，今回の研究で作成された回帰式は，かなり表面的な特徴のみを利用しているため，これらの特徴では測定できない自由英作文を書く学習者の自動採点には向いておらず，自動採点として使用するには改良が必要であるといえる。

4. おわりに

本研究では，自由英作文から比較的簡単に得られるさまざまな語彙指標のうち，自由英作文の総合的評価と相関の強いものは何であるのかを調査し，重回帰分析によって，総合的評価を予測する回帰式を作成し，その回帰式がどの程度使えるものであるのかを検証した。自由英作文と相関があった指標は，総語数（Token）以外では，使用語彙の難易度を示す指標（J8 Index）や語彙の豊富さを示す指標（D）などに中程度以上の相関が見られた。重回帰分析では，総合的評価の 60%を予測できる回帰式が作られたが，J8 Index や D などの指標は重回帰モデルに使用されなかった。これは，対象とした学習者の英語習熟度や自由英作文の総合的評価が，総語数とかなり相関が高いという結果からもわかるように，かなり能力が限定的な学習者集団を対象としたためであると考えられるが，研究の結果から，各種指標を組み合わせて予測を行うことはある程度可能であることが明らかになった。今後，さらに自由英作文の採点自動化を目指す場合には，対象とする学習者の母集団を正確に代表するようなサンプルをできるだけたくさん集め，表面的な特徴以外の指標を組み込むことができるようにしていくなど，更なる調査が必要である。

文　献

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment*, *4*. Retrieved June 22, 2007, from http://www.jtla.org

Field, A. (2005). *Discovering statistics using SPSS* (2nd ed.). London: Sage Publications.

石岡恒憲 (2004).「記述式テストにおける自動採点システムの最新動向」『行動計量学』*31*, 67-87.

石川慎一郎 (2005).「大学生英語学習者の受容語彙力と発表語彙力の関係－語彙サイズテストおよびエッセイ・コーパス分析に基づくアプローチー」『中部地区英語教育学会紀要』*34*, 337-344.

磯田貴道 (2004). 「生徒のプロファイリング」 前田啓朗・山森光陽 (編著) 磯田貴道・廣森友人 (著)『英語教師のための教育データ分析入門：授業が変わるテスト・評価・研究』(pp. 112-124). 東京：大修館.

川本竜史 (2004).『SPSS と Excel による[統計力]トレーニング』東京：東京図書.

小泉利恵 (2007).「スピーキング・テストにおける語彙的複雑さの指標とその特徴」第 46 回大学英語教育学会全国大会 『JACET 英語語彙研究会企画シンポジウム 語彙の測定とその問題点：さまざまな測定方法によって引き出される語彙の側面』発表資料.

前田啓朗 (2004a).「テスト欠席者の見込み点の予測－回帰分析－」 前田啓朗・山森光陽 (編著) 磯田貴道・廣森友人 (著)『英語教師のための教育データ分析入門：授業が変わるテスト・評価・研究』(pp. 73-81). 東京：大修館書店.

前田啓朗 (2004b).「因果分析の妥当性の検証－日本の英語教育学研究における傾向と展望－」*JLTA Journal*, *6*, 140-147.

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

Meara, P., & Miralpeix, I. (2004). *D_Tools*. University of Wales Swansea. Retrieved June 14, 2007, from http://www.swan.ac.uk/cals/calsres/lognostics.htm

靜 哲人 (2007).『基礎から深く理解するラッシュモデリング －項目応答理論とは似て非なる測定のパラダイムー』大阪：関西大学出版部.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

山西博之 (2004).『高校生の自由英作文はどのように評価されているのかー分析的評価尺度と総合的評価尺度の比較を通しての検討－』*JALT Journal*, *26*, 189-205.