大妻女子大学紀要 ― 文系 ―　No. 41,　2009年 3 月

語彙の統計量と総合評価の関係

―― 作文評価の基準特定にむけて ――

中　尾　桂　子

あらまし

作文評価の信頼性を検討するために，評価項目の妥当性を調べ，初期の文章表現上の技能指導における作文の目的と評価項目について考察する。評価の信頼性については，同一の評価項目で採点した 4 人の教師の採点結果の相関係数や信頼係数，分散分析に基づき判断する。また，評価項目の妥当性については，重回帰分析を用い，作文中のテキスト特性を画一的に示す指標として初期の文章自動採点システムで利用される項目に基づく 14 の指標で，作文に与えられた総合計点，ならびに，評価項目毎の得点を予測できるかどうかで判断する。これにより，作文の評価には，主観的で分析的な評価項目では非常に厳密な採点指標を設定するか，テキストの語彙的な特性のうち，語彙の意味内容に基づく関連性で内容の深さを測る方法を利用するかのどちらかの方向性を明確に打ち出しておくべきであることを主張する。

キーワード：日本人大学生の日本語作文，作文の評価項目，I-T 相関，分散分析，重回帰分析

1. 大学生の文章表現能力と文章表現教育

某大学の新入生に，「身近な事例を取り上げて意見を書く」という課題を与えたとき，以下のような文章が提出された。

> どこかのレストランに入った時に店の中の壁に「お昼のランチあります」って書いてあったけど，どんなものがついていて，いくらなのかは書いてなかった。
> わざわざ店員に聞かなければいけなかった。そういうことはしっかりと，何が入っていていくらなのかもしっかりと，書いておくべきだと思いました。
> それと似たような例は「日替わり定食」などだと思う。こっちの例のほうがこのようなことがよくある例だと思います。

この学生は，幼少の頃より教えられてきたように，自分の意見を述べる際には「と思います」を使って事実と意見を区別し，身近な場面で誤解を生む表現の問題点を指摘している。

しかし，文体の非統一，口語表現の混在，口語表現と文章表現の差への無配慮，指示詞を多用した省略による事実関係の不明瞭さ，平易で冗長な説明による層の薄い論理展開など，この学生の記述は，アカデミックな場にはそぐわないとして問題だとする教員が多いだろう。

昨今，大学生の文章表現能力に問題があるという指摘が教員の間で増え，その対策が必要だとする認識が広まっている。そのため，ここ数年，大学の新入生対象に，アカデミック・ライティング(1)の基礎としてレポートや文章の書き方を指導する目的の必須科目を設置する大学が増えている。

このアカデミック・ライティング関連の科目では，主に，文章記述の際に必要な基礎知識として，文の構造や文章構成，レポートの書式，原稿用紙の使い方といった，文章記述における形式的な面を指導することが多いが，学校のニーズや学生の能力に応じて，その内容は様々である。

例えば，情報検索スキル等を盛り込んだ形での情報リテラシー教育の一環として情報科学の範疇で扱う場合，また，語彙力としての漢字の読み方や助詞の使い方といった表記・文法面から記述を指導する場合，大学での学習の基盤となる知的活動の基礎スキルを指導する場合，論文作成を目的とし，思考方法，論の立て方を含めた，いわゆるコースワークまでと，幅広い内容がアカデミック・ライティングの範疇で捉えられている。諸外国の論文指導とは異なり，日本でのアカデミックライティングの指導が幅広いのは，表現教育に対する認識が未だ確立されていないことによると考えられる。

学生や諸機関の事情で，その内容や達成目標，指導教員や取得単位数が異なり，様々な取組があって，一概にはまとめられないものの，概ね，指導の目的は，大学生としての学術活動における正確な情報伝達能力の向上を基本としながら，論理の構築と展開，その際の客観的な考察の深め方を学び，さらに，その思考活動を正確に伝達するスキルを身につけることと言えそうである。

また，このアカデミック・ライティングの指導は，各大学の各講座へ進む前の基本的な指導として位置づけられることもあり，「文章表現」などの科目名で，全学共通科目として複数の教員で担当することが多い。さらに，文系理系を問わず，多様な教員が指導に参加することもある。したがって，教室経営，授業進行，同一方式の評価で，学部単位や全学単位で一斉指導体制により行われることも多くなり，教師用マニュアルを完備する学校も多く，採点基準やテストを統一しているところもある。

このような科目の問題としては，同一科目名で複数の教員が指導する場合は特に，その指導内に差が生じやすく，受講者アンケートで，教員によって内容が異なることや，指導自体が有益かどうかについて学生自身からも疑問を指摘されることがあげられる。これは，各評価者の意図する「良さ」と，評価観点の捉え方にずれが生じることにもよるようで，指導の観点，並びに，評価の観点が指導者により異なることにも原因があるのではないかと考えられる。

また，それらの科目を履修した学生が，必ずしも，専門科目でのレポートや卒業論文の提出の際に，基本的な約束事を踏まえた文章を書き，体裁を整えられているかどうかは，評価の対象にされないため，その目的が達成されているかどうか，本当の意味では不明である。

アカデミック・ライティングの指導には，独自の目的と背景に応じたそれぞれの機関独自の問題や，表現活動と作文技術に対する意識差の問題，また，評価方法の確立における問題など，社会的な位置づけ，作文指導に対する認識，評価自体の捉え方，といった複数の問題が伺える。これらの問題は，現在の高等教育において，アカデミック・ライティングというものが暫定的な指導であるという意識や，科目としてまだ確立されていないことによると考えられる。

2. 本研究の目的とその方法

本稿では，アカデミック・ライティングにおける作文指導の際の評価方法や評価基準特定にむけ，評価の観点と評価の捉え方，その方法について考察する。

まず，作文の性質を，語彙量，構成から明らかにする。次いで，それらの作文に対する教師の評価を分析する。作文の性質と評価の関係を調べるとともに，教師の評価観点，評価の妥当性を考察する。以上にもとづき，作文の性質と評価の関係を考察し，大学生のアカデミック・ライティング

の指導における作文の意義や評価観点，評価法を検討する一助としたい。

調査方法とその流れは以下の通りである。

1. 大学生の作文の特徴を，語彙統計量から明らかにする。
2. 評価項目自体の状態を概観した後，各評価結果を多重比較し，評価者間の差や，差のある部分を調べ，各評価の信頼性，妥当性について評価者間の信頼性係数を求める(2)。
3. 評価の総合得点に対して文章テキストの特性における何が影響するかという点については，テキスト変数からテキスト全体の評価点が予測可能かどうか，重回帰分析を行って判断する。
4. 初期の文章表現技能指導における作文指導の目的と評価項目の関係について考察を試みる。

本稿は，教師が認識する「良さ」や「重要性」が，測定可能な「目に見える」ものとして存在し，それが適切に評価されているのか，されていないとすれば，その原因がどこにあるのかという問題に対して，作文の文体的性質を評価と評価方法に関する意見や捉え方の違いから考察するのであるが，これは，評価の妥当性検証と，良い作文と評価される基準を明らかにすること，また，より公平な評価を目指すこと，並びに，良い作文モデルを提示することを目指すものである。指導の形式的な側面を確立するために役立つ基礎資料となることを期待している。

2.1 作文の性質調査の方法 ── 語彙量，構成の調べ方 ──

本稿の調査対象の作文は，ある私立大学の日本人1年次生が1学年の終了時に記述した意見記述型作文，84人分（05年度：45人＋06年度：39人）である。大学1年次生の作文として，傾向を見るため，テーマ，記述量，時間についての記述条件を，可能な限りで統一している。

84人の日本人学生の文章能力を客観的に示す資料はないが，授業の副教材として利用した日本語文章能力検定協会の『日本語文章能力検定問題集3級』(3)の問題で，およその平均が80点程度の学生である。このことから，採集時期や履修時間が異なる学生や，入学年度の異なる学生の作文

表1 調査対象の概略

調査条件・対象	内　　容
採集時期	日本人1年生の1年次終了時（1月）
採集文	「大学生にとってアルバイトは重要か」に対する意見記述型作文
採集人数	84人分（05年度：45人＋06年度：39人）
採集条件	90分授業内の30〜40分の時間で400字程度記述可能なB5サイズの用紙1枚に記述する。用紙はこちらから配布したものを利用する。記述の際の注意として，以下3点を指示 ・文章力評価用実力テストである ・学生の能力向上の状況を分析する ・指導内容へのフィードバックに利用する
被採集者の客観的実力	日本語文章能力検定協会の『日本語文章能力検定問題集3級』[i] の問題で，およその平均が80点程度
コーパスの構成	文末に句点を1つ含む，1文1行のデータに学生ID No.と行番号，段落番号が付加
作成時利用のシステム	語彙リスト作成には日本語用形態素解析システムChaSen 2.2［ii］を利用
計量対象	実質語（名詞，動詞，形容詞，副詞，接続詞）

を利用することは，作文の性質に影響を与えないものと考える。

テーマを「大学生にとってアルバイトは重要か」として，90 分授業内の 30〜40 分の時間を設定し，300 字から 400 字の記述が可能な B 5 サイズの記入用紙を 1 枚配布した。記述の際の注意として，文章力評価用実力テストであること，学生の能力向上の状況を分析すること，指導内容へのフィードバックに利用することを説明している。

また，この日本語作文テキストから語彙リストを作成するにあたっては，日本語用形態素解析システム ChaSen 2.2[(4)] を利用した。語彙リスト作成にあたって利用した品詞情報は，全てこの ChaSen 2.2 のデフォルトの性能に起因するものである。以上の分析対象の概略を，表 1 にまとめる。

2.2　作文評価の調査方法 ── 評価者，観点の選定理由 ──

作文の評価は，5 年以上の教授経験がある日本語教師，4 人に依頼する。依頼者とする際に配慮した条件は，アカデミック・ライティングの指導と「日本留学試験」の作文問題の指導経験があること，作文の評価において，今回依頼する方法で採点することに慣れていること，作文の形式面や技術面に着目して構成や論拠を捉えることになれていることの 3 点である。通常，主観的評価の分析では，評価者間の評価観点に対する見解を統一するために事前トレーニングを行うが，今回は，上記条件を満たすことで，行わなくても良いと判断した。

また，日本語教師に依頼する理由は，評価の主旨に対する理解を得やすいことにあった。日本語教師の場合，「日本留学試験」（http://www.jasso.go.jp/eju/whats_eju.html）の指導経験があると，作文の技術的側面の客観的な評価に慣れているが，国語科の教員は，作文を質的に評価する傾向があると考えられることによる。

たとえば，以下は，長野県立上田染谷丘高等学校の入試における作文の問題である。学校 HP より引用した。

図 1 を見ると，⑵の評価の観点は，内容の論理性を問うものではなく，質的であることがわかる。国語科教育では，このような観点が含まれた評価が多いが，図 1 はその一例である。

次に，日本語教育の作文評価の 1 例をあげる。図 2 は，外国人留学生として，日本の大学（学部）等に入学を希望する者について，日本の大学等で必要とする日本語力及び基礎学力の評価を行うことを目的に実施する試験における作文記述問題の評価基準である。こちらは，語学教育をベースとした言語教育の評価観点の 1 例であるが，図 2 を見ると，作文の技能的な面に着目した評価であることがわかる。

問題：
「雪つり」は，どのような目的で行われるものですか。二種類の「木」の性質をふまえて述べなさい。また，あなたのこれまでの経験の中から，自分を「強い木，堅い枝」「竹や柳」のいずれかに置きかえられる事例を探し，それについて述べなさい。

評価の観点：
⑴　課題文を正確に読み取ることができているか。
⑵　課題文の具体例を，自分自身に置き換えて考えることができているか。
⑶　論旨の展開が自然で，説得力のある述べ方になっているか。
⑷　わかりやすい述べ方，適切な用語，表記であるか。
＊これらの観点から，総合的に受検生の読み取り能力・分析力・論理的思考能力及び表現能力を評価する。

図 1　国語教育における作文評価例（入学試験「記述問題」評価基準）

語彙の統計量と総合評価の関係

A：技術面の評価観点

3点	個々の文についても，文章全体についても，執筆者の**意図が明快に**理解可能であるもの（文法・表記上の軽微な誤りや文体上やや不自然な点は許容する。）
2点	文法・表記上**明らかに適切でない**点を含むが，文章全体から執筆者の意図は明快に理解可能であるもの
1点	文法・表記上**明らかに適切でない点がかなり目立つ**が，文章全体から執筆者の意図を想像することは可能であるもの
0点	意味不明の文が多く，文章全体から執筆者の**意図を理解することが不可能又は極めて困難**なもの

B：論理面の評価観点

3点	**主張に根拠が示されて**おり，かつ，主張と根拠との間に**十分な論理的関係が**あり，矛盾が認められないもの
2点	主張に根拠が示されており，**概ね論理的**な関係が認められる，一部に論理的矛盾や非整合性も存在するもの
1点	主張は示されているが，その根拠が示されていない，又は，根拠が示されていても，**論理性・客観性を著しく欠いている**もの
0点	筆者自身の主張が示されていない，又は，**何を主張したか曖昧**であるもの

図2 「日本留学試験」作文問題評価基準

以上，国語科教育の場合と日本語教育の場合での作文評価の違いを見たが，日本語教育での作文指導は，基本的に質的な観点を含まないことが特徴である。

国語科教育でも，日本語教育と共通する観点として，3点あげられる。1つは「形式的規則」といったもので，1字下げなどによる段落分けなどの約束事を指し，2つめは，「文脈（話の流れ）のわかりやすさ」で，テーマと記述方法と記述トピックとの関連性が高い（明確）ことや内容・説明・表現が具体的で明確であることを指す。そして，3つめが「文章構造が明確」であるという点で，具体例の記述，テーマの絞り込み方やスタンス，内容が明確であること，文法や表現が効果的に利用されていること，文の構造，ポイント，主題，感情がわかりやすくなること，修辞的な美しさがあることによって印象的，個性的であることとなる。

国語科教育においても日本語教育においてもどちらにも共通する観点を見れば，それが評価判断時の重視点ということになるのではないかと考えるが，共通する作文評価の観点は3点あったが，いずれも，各々の立場でその見方が若干異なるようである。国語科教育では，あくまでも，深層的な質的観点が重視されており，形式的な側面は，あくまでもそれを判断する指標でしかない。

しかし，アメリカ経営大学院の入学試験（Graduate Management Admission Test; GMAT）におけるエッセイテスト[(5)]や，その他，通常の英作文評価と比較しても，その際の判断基準である技能面の評価観点には違いがなく，質的な評価観点は含まれない。表層的な観点を重視する方が，アカデミック・ライティングの作文評価の基準としては標準的なものであると考える。ここで述べた評価観点の捉え方の違いを図3にまとめる。

日本語教師の作文指導での評価経験があれば，無意識のうちに，質的な観点での評価を主観評価に含める恐れがないと考えられることから，日本語教師を評価者としたのであるが，以上のような検討に基づき，選定した評価者の概略について，表2にまとめる。

最後に，本稿での作文評価における評価観点について述べる。評価項目は，表3の7項目である。これは，ある私立大学においてアカデミック・ライティングの指導に2年以上携わる教師8人が相談して決めた項目である。これらのポイントが明確であれば，技術面でも，論理面でも最低限の質を保つ文章であるだろうというのが，設定の理由である。この項目は，性質上，形式に関する5つの項目と内容に関する2つの項目に分類できる。

一見しただけで，この7項目の評価は，形式面と論理面における項目数に偏りがあり，また，抽

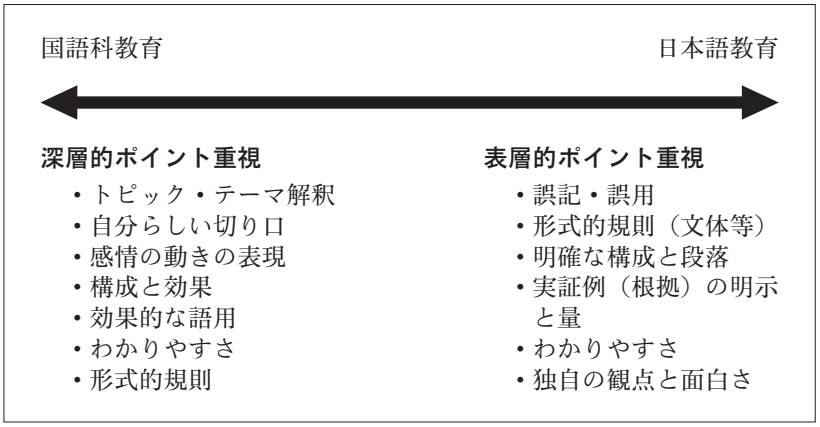

図3　国語科教育での評価観点の捉え方と日本語教育での捉え方の違い

表2　評価者の概略

調査条件・対象	内　　容
評価者の人数	4人
評価者の背景	5年以上日本語教師
評価者の条件	＊学部留学生や日本人に対するアカデミック・ライティングの指導，並びに，「日本留学試験」の作文問題の指導経験を持つ ＊作文の評価において，普段から，今回協力依頼した方法で採点することに慣れている ＊評価時，特に，形式的な部分に着目し，構成や論拠を捉える経験が多い
事前トレーニング	行わず，都度，連絡調整

象度が高いことがわかる。また，実際の評価における加点対象とその観点の設定が，評価者個人に任せられていることもわかる。

確かに，7項目が実際に何を指し，評価で何に重きがおかれているかという点は，評価者により異なると考えられるが，この評価観点は8人の指導者が文献を参照しながら考えたポイントであるため，今回は，アカデミック・ライティングの指導における作文評価のごく一般的なサンプル例と考えることにして，この項目のままで利用し，評価の違いが生じるポイントを考えることにする。

今回の評価対象の具体的な観点は，評価者G1が予備評価を行って決めた。39人の作文を概観し，平均以下だと考えられる10作文のサンプルに対して事前評価を行って問題点を挙げ，このとき挙げられた問題点が作文に見られるか見られないかで，7項目において採点を行うことにした。

採点は，7項目に対して，各々を5点満点とし，それらの集計を総合得点とする分析的評価法をとる。評価項目の詳細は表3のとおりである。

まずG1が39人分に対して評価を行い，その採点結果の信頼性係数最尤推定値（クロンバックα係数）を求めた。このときの信頼性係数は0.73であった。妥当だとされる0.8には満たないが，低いと言えるほどでもないため，この項目を用いて84人全体の評価，採点を行った。

採点・評価は，評価者G1が決めた詳細に則り，評価者G1, G2, G3, G4の4人で行う。今回の検証のために協力を得た評価者4人は，それぞれが5年以上の経験を持つ日本語教師で，学部留学生や日本人に対するアカデミック・ライティングの指導経験を持つ。評価の前にG1の基準に合わせて採点を行うようなトレーニングは行っていないが，予め，どのようなポイントを見つけ，何点ぐらいの配点を与えるかについて確認し，不明な場合は評価者G1に問い合わせ，調整しあった。

表 3　評価項目とその評価ポイント詳細

略記	評 価 項 目	評　価　観　点　詳　細
段 落	段落の有無	話題のまとまりごとの段落分け・段落開始時の 1 字下げの有無
文 体	文末の混在	文末のスタイルの統一（丁寧・非丁寧の混在の有無）
表 現	口語表現の混入	語句・慣用表現
	助動詞の過剰使用	内容に対するモダリティ表現の過剰使用「だろう」「そうだ」など
	「と思う」の多用	過剰使用，事実と意見の混同使用
	不適切な副詞	位置，過剰使用
	文法上の誤り	助詞，活用・修飾等の間違い
文	捻れ文	構文上主述関係が不一致
	長文	連体修飾節，連用形接続の多用
	体言止め	動作や状態叙述の省略，文自体の名詞化による準体助詞の省略・列挙など語句相当に扱うものを含む
表 記	誤表記，誤用	漢字・送り仮名の間違い，仲間語の利用を含む
構 成	前提欠落説明不足	修辞法上の含みを持たせた自明の省略ではなく，欠落。不十分な説明，説明と理由に同じ語を利用（A の理由は，A だから等），舌足らずな飛躍も含む
論 拠	論拠や結論不明	一貫しない意見，問題点と結論に関連性がない，意見や主張の根拠がない

3. 調査の結果

3.1　作文の性質

作文の総語彙数は，延べ 22,413 語であるが，これは，記号，数字，フィラーを抜いた語の数で，名詞，動詞などの実質語と，助詞や助動詞などの文法用機能語とを分けて計量した数である。表 4 にまとめる。実質的な異なり語数を見ると，1,000 種程度であることがわかる。

表 4　作文コーパスの語彙数

	全体（機能語）	実質語（名動形副）
延べ語数（記数フ抜）	22,413 語	7,077 語
異なり語数（記数フ抜）	1,699 語	1,035 語
TTR	7.58（%） 75.8 （‰）	14.62（%） 146.2 （‰）

また，1 人当たりの作文の平均実質語数を見ると，総語数 224 語，段落数 4 つ，文数 10 文，一文中の平均語数が 23 語となっている。これを表 5 にまとめ，1 文中の平均単語数を図 4 に示す。

表 5　1 人当たりの作文の語彙，段落，文数の平均語数

総 語 数	段 落 数	文　　数	1 文中の語数
224	4	10	23

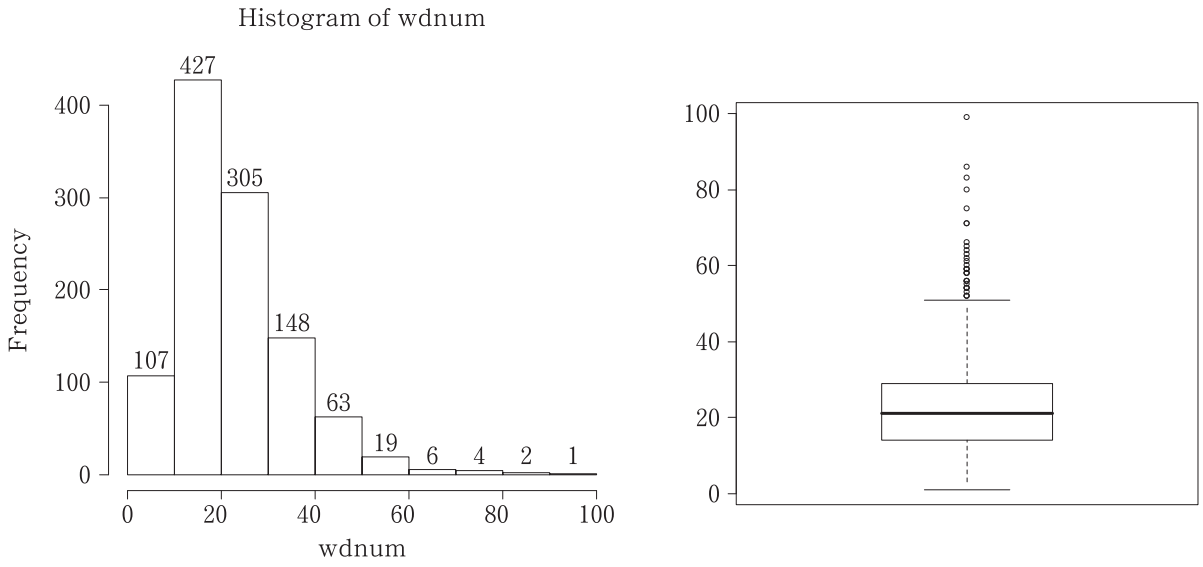

図 4　1 文当たりの平均単語数

図 4 からも，平均単語数が 20〜30 語に集中し，それ以上のものやそれ以下のものが少ないことが伺える。B 5 サイズの用紙 1 枚に記述していることもあるが，各段落が平均して 2 文ずつで記述されていること，また，4 段落で，平均 2 文ずつ，1 文の単語数が 20〜30 語であることから，既習内容の 1 つである「意見文の型」に当てはめ，4 分割して記述していることがわかる。過去に，土居光知が，ある限られた語で日常生活の諸事情を言い表すことを目的に汎用性の高い語を基礎語彙として選定したが，その『基礎日本語』での語彙数が 1,100 語である。そして，1939 年の日常生活での年齢別使用語彙調査による小学校 1 年生の男女平均が約 5,000 語強である。今回のデータは実質語が 1,000 語程度であることから，形式的には均質なデータが集まっていると言える。

もちろん，大学生のアルバイトに関する話題と限定されているため，その語彙量の多少を一概には結論づけられないが，このことは，限られた語を何度も利用していることを示唆するのではないか。

3.2　評価者 4 人の採点

評価項目の妥当性については，G 1, G 2, G 3, G 4 の 4 人の評価結果を基に，評価者間の①採点の平均値，②標準偏差，③総合点評価の評価者間信頼性を示すピアソン相関係数，④分析的評価の信頼性係数の数値ばらつきと評価項目に関する評定者間信頼性から判断する。

まず，評価者間のばらつきを見るために，各評価者の計上した 7 項目の得点の平均を比較し，それぞれの標準偏差を求めた（表 6）。若干の差があるもののG 1, G 2, G 3 は概ね同様であるが，G 4 の評価は異なる点が多く見られる。

各項目の得点により，総合得点に違いがあるのか，両者の間の交互作用を分析するために，二元配置による分散分析を行った。結果を表 7 に示す。

評価者間，評価者内の自由度，分散と，観測された分散比，すなわち F 値を見ると，(1)[P 値 ＞ 棄却域の確率] となり，P 値が 5%より大きい。また，(2)[F 値境界値 ＜ F 値]である。これらから，「評価者の総合評価結果は同等に行われている」とする帰無仮説を棄却し，差があるとわかった。

評価項目は，評価者同士で相談した結果，妥当だとして設定されたものであるにもかかわらず，差が生じるため，項目毎の得点の相関（Item-Total 相関）を調べた。

表 8 は，各評価者間の評価項目とその合計点との相関をとり，合計点と各項目の結果のみを比較

表 6　評価者の採点平均と各標準偏差

評価項目	G 1		G 2		G 3		G 4		4 人の平均
	平均	STDEV	平均	STDEV	平均	STDEV	平均	STDEV	
形　段　落	3.00	1.91	3.20	1.51	3.30	1.68	**2.50**	2.47	3.0
形　文　体	4.60	1.21	4.80	0.72	4.60	0.85	4.70	1.19	4.7
形　表　現	3.20	1.57	3.50	0.73	3.60	1.15	**1.70**	1.35	3.0
形　構　文	4.20	0.95	3.90	0.91	3.60	1.26	**2.10**	1.57	3.4
形　表　記	4.30	0.81	3.90	0.82	3.90	0.92	**1.70**	1.56	3.5
構　　成	4.10	0.89	4.20	0.86	4.00	0.90	**1.20**	1.58	3.4
論　　拠	4.30	0.96	4.00	0.85	4.20	1.05	**1.90**	1.93	3.6
合 計 平 均	27.80	4.29	27.50	3.98	27.20	4.18	**15.80**	1.94	
合計標準偏差	5.63	5.63	3.79	3.79	5.73	5.73	**6.87**	6.87	

表 7　評価項目毎の 4 人の評価の分散分析（繰り返しのある二元配列）結果

分散分析表

変動要因	変　動	自 由 度	分　散	観測された分散比	*P*－値	*F* 境界値
評 価 者 間	641.9932	6	106.9989	64.02183	8.77 E－74	2.102482
評 価 者 内	1226.724	3	408.9082	244.6666	5.8 E－138	2.608729
交 互 作 用	426.966	18	23.72033	14.19285	1.96 E－41	1.608292
繰り返し誤差	3884.071	2324	1.671287			
合　　計	6179.755	2351				

表 8　評価者別 Item-Total 相関結果の比較

	形　段	形文体	形表現	形　文	形表記	構　成	論　拠	合計点
G 1	0.727198	0.528216	0.747543	0.55844	0.724011	0.733953	0.680762	1
G 2	0.609154	0.533807	0.54489	0.57307	0.50231	0.68823	0.645193	1
G 3	0.693804	0.64447	0.749571	**0.815503**	0.692669	0.764109	0.785364	1
G 4	0.501542	**0.273438**	0.651892	**0.465983**	0.624997	0.732879	**0.811453**	1

したものである。これを見ると，それぞれは，概ね，緩やかな相関を保っているように見えるが，中には，重要視されているものと，逆に軽視されていると考えられる評価項目が見られる。

例えば，特に，「論拠」に関する項目を非常に重要視する立場と，「形文（形式における構文上の捩れ，主題の立て方）」に関する項目を重要視する立場があり，また，「論拠」重要視者は，「形文体（形式上で見た丁寧，である体等の文のスタイル）」や「形文（構文）」を軽視しているということなどである。

評価者の能力に差はないという前提で考えた場合，これは，評価基準設定に詰めの甘いところがあることを表していると考えられる。

論拠や構成といった，人により異なる観点を示す可能性が推測される項目だけではなく，割と具体的で明確な形式における評価項目においてもゆらぎが見受けられる。

そこで，評価項目に対する妥当性を判断するために，クロンバックの α 係数を計算した。G 4 の

み四捨五入して，信頼性があるとされる 0.8 以上とならなかったが，0.6 以上は信頼性に問題はないとされることが多い（表 9）。

表 9　評価者の信頼係数（クロンバック α）

	G 1	G 2	G 3	G 4
クロンバック α	0.765837	0.845029	0.843764	**0.661732**

係数値の低い G 4 の評価者が評価した採点結果の内容を見ると，確かに，厳しい評価であるという印象を持つが，それでも，G 4 の評価が特に並外れて不当なものであるという印象は生じない。逆に，各項目に対する「良さ」の基準や評価項目に対する重要性の程度差が他 3 名よりも明確だと言える程である。項目の設定理由から項目の妥当性を検証する必要があるだろう。

以上，評価者間のばらつきや，信頼性を比較したが，その結果，評価項目は，教師の内省で妥当だと考えられる項目が挙げられたものであるとはいえ，基本的には分析的な印象評価として，一応，機能しているように見える。ただし，評価者によって，ゆれが生じていることは事実として確認できた。

4. テキスト特性と評価合計の関係

評価項目として挙げられている観点は，4 人の評価者が評価対象とすることに違和感を感じるものではなく，教師の印象として問題はないというものであった。

しかし，各評価者で，評価項目の位置づけが異なる場合があり，評価したい観点がどの項目に相当するかについては評価者から何度も質問が出された。このことから，項目の詳細に対する，各評価者の位置づけの差が，項目の妥当性に対する印象と実際の評価の差が生じる原因につながっているのではないかと考えられる。

ゆれが生じない評価基準の設定には，より具体的な評価項目を挙げ，その項目がどの評価項目の範疇にカテゴライズされるものかという点を明確にしておく必要があるだろう。

教師の着目している点を具体的に作文の中から拾い出し，評価基準の具体的なポイントとして明示するために，実際の作文テキストの特性を調べ，評価との関係を見ておきたい。

本節では，テキストの語彙的指標を検討し，ついで，選出したテキスト特性の指標と各評価項目との間で重回帰分析を行う。これにより，項目毎の評価結果で総合点が予測できるか調べる。

4.1　欧米の文章自動採点システムの動向と評価指標の歴史的変遷

現在，欧米の高等教育における英語の文章評価では，文章自動採点システムが利用されている。このシステムの利用の最大の目的は，教員の業務軽減，文章評価の一元性維持の 2 つである。この文章自動採点に関する研究は，教育支援の目的で始められた Page（1966）による PEG（Project Essay Grade）というシステムの開発に単を発し，現在では，文章自動採点システムは人間の評定値と比較しても信頼性に差がない精度にまで徐々に向上してきている（石岡，2004）。そして，一元評価の信頼性（評定時，エッセイの評価順番による系列的効果の影響を受けないことや課題選択の際の等化）などの良さが考慮され，十分実用的なものとして教育会で機能しており，この分野は，様々な研究が展開される場としてその地位が確立されている。

本節では，この文章自動採点システムにおいて利用されてきた採点指標の変遷をたどり，採点時

の一般的で具体的な指標があり得るかについて確認する。本稿で検討するテキスト特性の指標決定の参考にするのが目的である。

4.1.1　欧米の文章自動採点システムの動向と評価指標の歴史的変遷

60 年代にアメリカで開発された PEG では，当初，文章の特徴を量的に計る指標として，「平均ワード長」，文章の長さとして文章中の「全ワード数」，「コンマの数」，「前置詞の数」，「一般的でないワードの数」を，文章評価のための特徴量としていた。

しかし，この指標では，単語さえ関連性が深ければ，内容が支離滅裂でも高得点を取りやすく，作文の質を決定する，内容（content），組織化（organization），文体（style）を捉えていないため，教育的なフィードバックができないという問題が，欧米の教育界で指摘されていたという（石岡，2004）。

これを受けて，80 年代初期の WWB（Writers Workbench）システムでは，「スペリング」や「語法」，文章に含まれるワード，文節，文数に基づく「可読性（readability）」が指標として付加された。

60 年代から 80 年代に至るまでのシステム世代では，文章の表層に現れる形態的特徴を指標にしている。しかし，実は，こういった特徴量は，本来測定しようとする作文要素の代用として位置づけられていたものであり，質的な分析を行う指標としては十分とはいえないものであったとされていた（石岡，2004）。

90 年代に入ると，Dumais（1997）により，情報検索手法を発展させた潜在意味分析（Latent Semantic Analysis; LSA）が応用され始めた頃から，自動採点システムへの自然言語処理技術の適応が進み，文章の質的分析が行われるようになった。

アメリカ経営大学院の入学試験（Graduate Management Admission Test; GMAT）におけるエッセイテスト（Analytical Writing Assessment; AWA）では，評価の観点として文法の多様性（syntax variety），内容（topic content），組織化（organization of idea）が挙げられているが，これらの評価により，言語上の特徴を示す量的な数値に基づいて，かつ，人間の専門家評価と同程度の評価が得られるところまで精度が保障されるようになっている。

90 年代以降の文章自動採点システムで指標とされている言語上の特徴量とは，自然言語処理技術における構文解析や，当時の情報検索の分野で主流の単語共起頻度に基づいたベクトル空間モデル利用による統計的数値が代表的である[(6)]が，2000 年以降，他に，語彙を意味内容における一致の度合いを測定する方法や，ベイズ理論やルール発見アルゴリズムを搭載したものもある。その上，80 年代までに開発されたシステムが改良され，シソーラスリストや，作文の質的測定のための複雑な変数による語彙の重み付け等を利用して内容，文体，構成（メカニズム）の 3 観点の評価ができるようになっている。このことから，現在は，その指標のバリエーションはかなり多くなっており，何を採用するかは，評価結果として何を見たいかという点に合わせて，重点をおく指標が変えられるように，選択肢が増えている。逆に言えば，利用者の意図がどこにあるかでどの指標を利用するシステムを使うか，よく考えなければならないということであろう。

以上から，現在は，深層的な特徴を指標として，適切な語彙選択を評価するといった方向で進められていることがわかる。これらは多数の日本語文章評価システムにも応用されている（石岡，2004）。もちろん，言語上の特性に応じた違いはあるが，概ね同等のことが可能である。

4.1.2　作文評価の合計点を左右するテキスト特性として考えられる文章評価指標

前節では文章自動採点システムで評価時の指標を概観したが，取り上げられていた指標を以下の表 10 にまとめ，本稿で利用するものについて考える。

表 10 の 1 から 8 までの指標は，90 年代以前の文章自動採点システムで評価指標とされていたものに相当するが，これらは，教師の評価観点とその印象への影響が考慮されていると考えられるもの，すなわち，仮想特徴量指標とでも言うようなもので，教師の「良さ」判断のためのテキストの特徴を示す代用物である。

例えば，1 のワード長は，長い単語が多いと，抽象的で高度な内容だという読み手の印象を反映するとされる。また，4. 前置詞の数や 8. 可読性を左右する要因は，単純さや平易さに対する心象を反映し，6. スペリングや 5. 非一般的な語は，その種類により，幼稚さや軽さに対する心象を反映すると見られ，概観的心象との一致を推測する要素として利用が可能である。

もちろん，表 10 にあげた文章採点システムの指標は英語での場合であり，1〜4 のように，日本語に応用する場合はその方法が異なるものがある。平均ワード長を見ても，日本語では 2 字の語の方が 4 字の語よりも平易だとは言えない。それは，ひらがなで書かれた文字の方が長くなるが，大和ことばであれば漢字より抽象度は低くなることなど，言語の差による。

また，本稿の検証においては十分生かしきれないものもある。今回の対象テキスト数は 84 件であり，一般化のための十分なサンプルが得られていないことや，そもそもアカデミック・ライティングの修作として記述が促される作文にはまだこれと言った正解の基準が設定されておらず，比較対象のモデルがないことから，10〜14 のような指標は採用しにくい。

ただし，正確に言うと，13 は比較対象のモデルが存在しないという問題ではない。今回，評価の対象となった日本人大学生の文章は，文字数制限によるものか，平易で単純であり，繰り返し同じ語を多用していることからシソーラスを使った関連性を図るような条件が整いにくいため，13 に限っては，本稿で対象としたテキストにそぐわないということである。

表 10 の 5.「一般的ではないワード数」，すなわち，頻度 1 などの低頻度語，6.「スペリング」すなわち，表記ミス，7. の「語法」の適切性，について見ると，今回の評価対象がネイティブの作文で，明らかな間違いが少ないことから，ほとんどの作文において評価が同じものとなる。

また，「8. 可読性（Readability）」にしても，今回のデータは文章が平板で，複文，長文が，いわゆる重文の範囲を出ず，文節数も少ない。よって，埋め込みの深層化を計る計算や変数を使うほどでもない。これは今回調査対象で用いる 1 作文の分量が 300 字程度で内容が平易なことによるた

表 10　明確な評価指標

採用判断	表層的な指標	採用判断	深層的な指標
△	1. 平均ワード長	○△	9. 構文情報
○	2. 全ワード数	×	10. 単語共起頻度によるベクトル空間モデル
△	3. コンマの数	×	11. ベイズ理論
○△	4. 前置詞の数	×	12. ルール発見アルゴリズム
×	5. 一般的でないワードの数	△	13. シソーラスリスト
×	6. スペリング	○△	14. 語彙の重み付け（北，1999） （N グラムモデル，隠れマルコフモデル，確率文法，最大エントロピーモデルなど）
×	7. 語　法		
×	8. 可読性（語数・文節数・文数）		

表 11　テキスト特性指標

採用判断	表層的な指標	採用判断	深層的な指標
○	1. 文字種の違い（漢字含有率）	○△	9. 構文情報（係助詞・文種・文体）
○	2. 形態素数（全体と作文毎）	×	10. 単語共起頻度によるベクトル空間モデル
△	3. 句点数	×	11. ベイズ理論
○	4. 助詞・助動詞の数	×	12. ルール発見アルゴリズム
×	5. 頻度 1 の単語語彙リスト	×	13. シソーラスリスト
×	6. 表記ミスの数	×	14. 語彙の重み付け （N グラムモデル，隠れマルコフモデル， 確率文法，最大エントロピーモデルなど）
×	7. 語法の適切性		
△	8. 文数（文節数は今回未備）		

めで，他の条件のデータでも同様だというわけではないだろう。

語彙の量的な差がないものは，統計的な計算であることを考慮して指標とせず，検証対象外とするほうがよいと考え，表 10 の「採用判断」に○または△と記号を記した。その判断と日本語特有の条件，評価観点が計れるかどうかを考え合わせて，本稿では，日本人作文のテキスト特性を計るための指標を表 10 から表 11 のように改変する。

4.1.3　各評価項目別評価との関係

本節では，指標相互の関係を見，その後，指標から各評価者の総合得点を推測することにより，関係をより明確にしたい。

まず，テキスト特性同士の相関である。

一般的な文章自動採点システム等に倣って表 11 にまとめた「テキスト特性」とは，「漢字含有率（他文字種と比較した漢字の使用割合）」，「品詞数（名詞，動詞，形容詞，接続詞，副詞，助詞，助動詞，係助詞）」，「総語数」，「段落数」，「文数」，「句読数」，「字数」である。

品詞数はすなわち語数であるが，各品詞をそれぞれ独立させたことで，具体的には，表 12 の 1 行目の項目に上がっているこの 14 指標に分けられる。

各テキスト毎に，14 の指標の相関係数を求めたところ，文数と単語数の 2 つ，各品詞毎の語数の間で相関が高かった。1 文中の語が多いと文が長くなるのだから，それは当然のことであるが，そのほかの項目の間には，特に高い相関が見られる項目はなかった。そこで，評価者 4 人の合計得点を合わせ，テキスト特性としてあげた指標と各評価者別合計点の相関を求めた。

相関が顕著に見られないという原因は，一つに，対象データの文章量が少なく，かつ，構成に配慮して書く学生が多かったために，内容が浅いことが考えられる。

ただし，動詞が多い文章では，名詞と助詞の使用が他より増え，さらに，句読点が多くなる。動詞が多いテキストで語数が多くなる場合，文の種類のベースが名詞文で，動詞の利用により，その分だけ，更に名詞と助詞が増えること，さらに，動詞利用で語数や文章が増えることが，また，別途，句読点の多さに関連すると推測できる。

以上から，単語数の多いテキストでは，助詞を構文情報を計る要素とできる可能性があるのではないかと考えられるが，今回のように文章量が少ない場合には有意な差が出るのか不明であるため，利用しない方がよいのではないか。

次に，各テキスト特性の総合点への重回帰分析を行う。

テキスト特性とした指標が，4 人それぞれの合計得点を推測するものとなるか見るため，テキスト特性指標を説明変数として，各評価者が出した得点を推測した。表 12 は，その結果のうち，各指標から見た評価得点の行を抜粋し，評価者間の比較をしやすいようにまとめたものである。

表 12 を見ると，強い相関を示す指標は見られない。0.3 以上の相関係数が見られるものが 3 人の評価者において「語数」と「句読点」となっていることからすると，評価者 3 人は，文章が長いこと，句読点に対してよい印象を持っていることが共通しており，それが，若干，緩やかに評価にも影響を与えているのかと考えられる。しかし，全体的に数値が高くないことを見る限りでは，その評価自体において，いずれの評価者もこれら 14 の指標にウエイトを置いて採点しているわけではないようである。評価観点と点数を決めていても，結局は，テキスト特性は二次的な参照情報でしかなく，評価者の内部の何らかの別基準に基づく印象で評価しているのではないか。

ここまでは，テキスト指標から 7 つの評価項目に対して評価者の出した総合得点を推測したが，以下，視点を変え，14 のテキスト特性の指標（名詞，動詞，形容詞，接続詞，副詞，助詞，助動詞，係助詞，総語数，段落数，文数，句読点数，字数，漢字含有率）で，各評価項目となる，形式段落，形式文体，形式表現，形式構文，形式表記，構成，論拠といった 7 項目毎の得点を，それぞ

表 12　テキスト特性と 4 人の評価との重回帰分析の結果

	名詞	動詞	形容詞	接続詞	副詞	助詞	助動詞	係助詞	総語数	段落数	文数	句読	字計	漢字率
G 1	−0.162	−0.147	0.038	0.046	−0.134	−0.243	−0.202	−0.137	−0.077	0.050	−0.019	0.060	−0.115	0.101
G 2	0.295	**0.317**	0.124	−0.033	0.157	0.244	0.052	0.076	**0.389**	0.250	0.259	**0.417**	0.292	0.156
G 3	0.152	0.242	0.041	0.000	0.150	0.135	−0.044	−0.028	**0.270**	0.223	0.191	**0.315**	0.194	0.104
G 4	0.166	0.218	−0.007	0.010	0.019	0.181	−0.066	0.012	**0.309**	**0.351**	0.171	**0.330**	0.175	0.031

表 13　評価項目別に見た *P* 値 0.6 以上の指標（全体出現 10 回以上太字）

	形段落	形文体	形表現	形構文	形表記	構　成	論　拠	.6＞の回数
名　詞	1	3	2	1	4	0	2	**13**
動　詞	2	2	3	2	3	1	2	**15**
形容詞	2	3	1	2	3	2	4	**17**
接続詞	1	2	2	2	4	2	4	**17**
副　詞	4	2	1	1	2	3	2	**15**
助　詞	2	0	1	0	0	0	2	5
助動詞	4	3	1	0	1	0	2	**11**
係助詞	3	1	2	3	1	1	2	**13**
総語数	2	2	1	2	2	0	0	9
段落数	1	3	1	1	1	0	2	9
文　数	2	2	3	1	3	0	0	**11**
句　読	1	0	3	1	0	1	1	7
字　計	0	0	0	0	0	0	0	0
漢字率	0	1	2	0	0	0	2	5

れ重回帰分析で予測する。表 13 は，重回帰分析の結果，P 値 0.6 以上の指標を品詞別にまとめたものである。

7 項目のいずれにおいても，補正 R 2 は 0.4 に満たないため，特徴的な傾向があるとは言えないが，それぞれの P 値が 0.5 以上で有意差が見られたテキスト特性は，名詞，動詞，形容詞，接続詞，副詞，助動詞，係助詞，文の数であった。有意差の見られた指標として，4 人全ての評価項目に共通する「形式段落」,「形式表記」,「論拠」の 3 項目で見られたテキスト特性を示す指標は，形式段落で「助動詞」，形式表記では「名詞」と「接続詞」，論拠で「形容詞」「接続詞」であった。これらの共通性には，言語的にも，関連性においても，特に顕著な意味はないと考えられ，2 章でみた相関や分散値のとおり，ばらつきがあることを意味しているのみである。これは，つまり，評価者で明確な共通観点があるわけではなく，それぞれの評価時の着目点もまた異なっていることを示唆すると考えられる。

一方，評価者別に，表 14 にまとめなおすと，それぞれの着眼点の違いが見られた。常に，意識していると考えられるような指標を確認する目的であったが，観点により異なることと，人により，若干，よく影響を受けるものがあるということが，緩やかに指摘できる程度である。

以上から，評価者により，分析的な評価が異なることが追認でき，また，結局，実際の評価では，指定評価項目の有無や程度にあまり関係なく，不明瞭ではあるが，評価者によっての内的基準と言えるようなものの存在があり，それが改めて確認できたと考えられる。

これは，推測でしかないが，それぞれ異なる内部基準のようなものにより，分析的な評価を下しても，全体の総合計点の結果は，それなりに「妥当」に見えるのだということが言えそうである。教師として評価してきた経験に培われた職人業なのかもしれないが，この点をより明確にしたいものである。

ここまでの分析の結果からは，今後，形式上のテキストの表層的な指標ではないものを指標とし

表 14　評価者別に見た P 値 0.6 以上の指標（4 回以上太字）

G 1		G 2		G 3		G 4	
助動詞	**5**	**形容詞**	**6**	**名詞**	**4**	**形容詞**	**5**
係助詞	**5**	**接続詞**	**5**	**動詞**	**4**	**係助詞**	**5**
形容詞	**4**	**係助詞**	**5**	**接続詞**	**4**	**名詞**	**4**
文数	3	**動詞**	**4**	**副詞**	**4**	**動詞**	**4**
動詞	3	**総語数**	**4**	形容詞	2	**接続詞**	**4**
総語数	3	**文数**	**4**	助詞	2	**字計**	**4**
名詞	2	名詞	3	助動詞	2	副詞	3
副詞	2	助動詞	3	係助詞	2	助動詞	3
接続詞	2	段落数	3	段落数	2	総語数	3
字計	2	副詞	2	総語数	1	段落数	3
句読	2	助詞	2	文数	1	文数	3
漢字割合	2	漢字割合	2	句読	1	句読	3
段落数	1	句読	1	字計	1	漢字割合	1
助詞	1	字計	1	漢字割合	0	助詞	0

て検討する必要性が指摘できたが，本稿では，それを一般化できる文章数も評価者数も備えていないため，深層的なテキスト特徴を利用した主観的評価に関する詳細な分析は，今後の課題としたい。

5. 考　　察

以上，相関，分散，重回帰分析により，テキスト特性の指標同士の関係や，これらが作文の合計得点に与える影響の有無を調べたが，その結果，今回，テキスト特性を表す指標とした語彙的な要素は，本稿で行ったような評価に影響を与えるものではないということが，明らかになった。

今回取り上げた指標で，評価得点に直接結びつく指標が明確にならなかった理由として考えられることは，指標が抽象的であること，ならびに，語彙の形式的な面と本稿の評価に関連がないことだと考えられる。

したがって，今回のような形で一般に行われているであろう主観的評価は，よく利用されるにもかかわらず，実は，何も評価できないものではないかということが指摘できる。

つまり，評価観点の設定が極めてアバウトに行われることが多いにもかかわらず，「なんとなく」採点結果に妥当な心象を抱く点に対して，やはり関係がなかったのであるから，採点の方法やその内容，如いては，指導の内容とその目的まで考え直さなければならないということが本稿の調査結果から追認できたと言えるのではないだろうか。

このことから，「よい」作文評価，すなわち，信頼性と妥当性の高い作文評価のためには，当然のことながら，まず，評価項目の具体性を明確にしなければならないということを改めて主張しておきたい。

また，実際にテキストから論理性を判断して論拠を読み取るには，テキストの形式的な語彙的指標の設定をより詳細，かつ，具体的にすることか，反対に，大きく鳥瞰するために深層的な面からも抽象化を行うかどちらかを行う必要があるのではないか。もちろん，語彙数や品詞が大して影響していないというだけで，他の要因による影響がある可能性もある。今後も，このような主観評価やその対象に対する分析は，評価の方法の確立を目指すために，テキストジャンルを変えてより詳細な分析を継続的に行っていくべきであるだろう。

さらに，自動文章評価システムの歴史的変遷に見られるように，評価と結びつくテキスト指標としては，テキストの数量的な語彙特性よりも，語彙の意味内容からの影響面が大きいと推測される。このことから，テーマやトピックごとに語彙の意味ネットワークのモデルとそのバリエーションサンプルを多く用意すること，同時に，作文コーパスを構築し，教員が利用できるように共有化されることが望まれる。

当然のことではあるが，アカデミック・ライティングといった新しい指導科目を設定する場合，作文指導の目的と評価項目に関して，詳細に検討を重ねる必要があり，同時に，前提となる目的や文の種類の違いにおける明確な基準を設ける必要がある。自戒を含め，高等教育機関における学術活動を支える基礎教育のための今後の課題と考えたい。

6. ま と め

アカデミック・ライティングの初期指導において書かれた作文の評価の信頼性や評価項目の妥当性に関して，重回帰分析や相関をとることにより，初期の文章表現上の技能指導における作文の目的と評価項目について考察した。それにより，文章表現の経験を持つ教師が認識する「良さ」や

「重要性」が計れているのか，それが適切に評価されているのかについて考察し，初期の文章表現技能指導における作文指導の目的と評価項目の関係について，また，注意点について述べた。

本研究は，統計数理研究所言語系共同研究グループ合同研究発表会，第3回「言語研究と統計」ワークショップの第3部：神戸グループ「英作文評価の自動化と統計」において発表した内容と，その予稿集である統計数理研究所共同研究リポート215にまとめた内容とを合わせて加筆修正したものである。

〈注〉

（1） 本稿で言う「アカデミック・ライティング」とは，狭義では，論文，レポート，プレゼンテーション原稿を指し，広義では学術的な場面や公的な場面で記述される標識や文章全般を指す代名詞として利用される。

これは，元々，欧米の大学のライティングセンターなどで行われるネイティブの学生を対象とした論文指導を指す概念であるが，日本では，ノンネイティブのリテラシー教育で主に利用されてきたことばであるため，アメリカのライティングセンターの論文指導とは異なり，平均的なネイティブの文章相当の記述能力を伴わない場合の学生に対する補修的な意味合いも含み，若干，補修としての意味合いが含まれることもある。

（2） 今回の分析対象である評価は分析的評価尺度により採点された結果であるが，評価者は，予め，評価者間で評価項目に関するコンセンサスをとっている。

（3） ㈶日本漢字能力検定協会監修の日本語文章能力検定のための過去問題集と事前対策のための問題集のうち，事前対策用を利用した。この協会の設定によると，3級は，高等学校在学程度と設定されていて，「社会活動に参加し，積極的に理解・表現活動を行って学問・教養を吸収し，自己を確立することを可能にする文章能力」があると見なされるレベルである。

（4） 奈良先端科学技術大学院大学自然言語処理研究科松本研究室で開発されたフリーの形態素解析システム，茶筌のことで，本稿のテキスト処理にはバージョン2.2を利用している。京都大学で開発されたJUMANをベースにしてはいるが，統計的な機械学習の手法を用い，内部に構文解析処理が組み込まれているため，処理の性能と速度が向上している。ただし，ChaSen 2.2はIPA品詞体系を使用しており，その辞書や精度は毎日新聞の文章に基づいているため，エッセイ等の記述に対する解析精度はそれほど高くなく，解析ミスも多くなる。本稿で行った形態素解析の結果における品詞特定上のミスは未修正のままで，品詞別の語彙の計量には，解析ミスの結果もそのまま含まれている。つまり，品詞や語彙数として示す数値はだいたいの程度を表すもので，現時点では100%の正確さで単語認定を行ったデータを用いているわけではない。

（5） （Analytical Writing Assessment; AWA）で，評価の観点として文法の多様性（syntax variety），内容（topic content），組織化（organization of idea）が挙げられている

（6） 例えば，椿本（2005）の日本語の文章自動採点システムを例に挙げると，まず，内容語を検出し，それを単語数（t）として，文章数（d）と掛け合わせ，単語出現頻度行列（X）を得る。次に，これに基づき，単語ベクトル，文書ベクトルとして，ベクトル空間表現を行うことで，行列に対して，特異値分解し，ベクトル展開の次元を縮退させて文書の特徴ベクトルを得る。最後に，各特長ベクトルのコサイン類似度を用いて単語同士，単語と文書，文書同士の類似度を測定するという流れである。

参考文献

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』第31巻第2号（通巻61号），pp. 67–87.

北研二・津田和彦・獅々堀正幹，2002,『情報検索アルゴリズム』共立出版，p. 35.

椿本弥生・赤堀侃司，2007,「主観的レポート評価の系列効果を軽減するツールの開発と評価」,『日本教育工学会論文誌』30(4)，pp. 275–282.

Landauer, T. K. and Dumais, S. T., 1997, Asolution of Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review,* 104: 211–240.

Page, E. B., 1966, The imminence of Grading Essay by Computer. Phi Delta Kappan, 47: 238–243.
東京学芸大学岸研究室 HP（http://www.u-gakugei.ac.jp/~kishilab/I-Tcorrelation.htm）2008. 3. 1 訪問