

第34回 研究助成

**A 研究部門・報告Ⅲ・英語能力テストに関する研究**

# 英文要約採点への自動英文解析ツール CRAT の 利用可能性の検証

研究者:茨城県／筑波大学大学院 在籍 丹藤 慧也

《研究助言者:小泉 利恵》

**概要**

本研究では、自動英文解析ツールである CRAT (Constructed Response Analysis Tool) の英文要約採点への利用可能性について 2 つの調査を行った。まず、調査 1 では CRAT を用いて、複数の協力者によって作成された英文要約の模範解答のテキスト分析を行い、テキストの特性を検証した。結果として、共通点が見られたが、協力者によって多少のばらつきが見られた。次に調査 2 では、CRAT を用いて、日本人英語学習者の英文要約のテキスト分析を行い、算出された指標の傾向とライティング力との関係を調査した。結果として、読解テキストの受験級が上がるにつれて、同義語や似た意味の言葉に言い換えることが困難になり、原文をそのまま使用してしまう傾向があることがわかった。また、受験級が上がるにつれて、語彙の洗練度が高くなる傾向があることがわかった。算出された指標とライティング力の関係に関しては、テキストのトピックや語数、さらに要約で求められる語数によって、ライティング力を予測する指標は異なる可能性が示唆された。

**1**

## はじめに

近年技能統合型タスクの需要も高まっている。タスクをより現実生活に近づけるという理由のためである。しかし、近年の大規模な大学入試では、リーディングやリスニングテストに関し

て多肢選択式問題が取り入れられ、採点の信頼性が確保されている。その一方で、英文要約などのライティングにおいては、熟練した採点者に依頼したり、採点する前に採点者間でトレーニングを実施したりしているにも関わらず、多肢選択式問題に比べ採点者の主観が介在してしまうため、採点者間でのズレが問題になることがある (Knoch, Fairbairn, & Jin, 2021)。また、英語の授業においても、学習者の英語ライティング能力を測定するために、教師がエッセイライティングや、技能統合型タスクとして英文要約を課す場合がある。教師にとって、この英作文や英文要約を採点・評価する作業というのは、多くの労力と時間を要する。そこで、近年 Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004) や Text Inspector (Bax, 2012) など英作文の自動評価が注目を浴びてきている。一方、英作文自動評価ツールには有料のものも多く、手を出しにくい場合も多い。以上の背景を踏まえ、本研究では、近年読解研究において無料の自動英文解析ツールとして活用されている CRAT (Constructed Response Analysis Tool) の英文要約自動評価への応用可能性を検証する。このツールは自然言語処理とデータマイニングを利用し、さまざまな指標を用いて、英文の言語的特徴を客観的に評価することが可能である。

## 2

# 先行研究

## 2.1 技能統合型タスク

近年、コミュニケーション能力の育成を目指して独立した4技能5領域(聞くこと・話すこと[やりとり]・話すこと[発表]・読むこと・書くこと)を総合的・統合的に学ぶことの重要性が指摘されている。第二言語(L2)教育実践において、4技能5領域を伸ばそうという指導から発展し、複数の技能を組み合わせたタスクを用いて運用力を高めようという指導が広く用いられるようになっている。

その中でも技能統合型ライティングタスク(Reading to Writeタスク)は、学術的な場面における真正性が高いことや、「要約」や「言い換え」など技能統合型タスクに特有な能力を測定できることなど、多くの利点から大規模テスト(TOEFL iBTなど)でも導入されており、その教育的価値も広く認められている(e.g., Plakans, 2015)。しかし技能統合型タスクの評価方法に関しては、研究の蓄積が少なく、どのように評価するかが確立されていない(Cumming, 2013)ことに加え、採点における困難さが指摘されている(e.g., Gebril & Plakans, 2014)。採点における困難点としては、テキスト借用が不適切に行われやすい点、記述内容が資料に基づいているか、隣接するスコアの記述子をどのように識別するか、などの判断が困難である点などが挙げられる。これらは採点の妥当性・信頼性・実用性・波及効果に影響を及ぼす可能性がある。そのため、より評価しやすく、理解しやすい評価方法が確立されることが必要不可欠である。

## 2.2 英文要約

英文要約は統合的なライティングタスクであり、リーディングからライティングへのタスクとして理解されている。具体的には、理解(comprehension)、凝縮(condensation)、産出(production)という一連の複雑なプロセスを伴い(Kintsch & van Dijk, 1978), 「圧倒的な認

知的負荷をかけることができる制約を伴う、非常に複雑で再帰的な読み書き活動」(Kirkland & Saunders, 1991, p. 105)とみなされる。

英文要約スキルは教育において必須のスキルであるため、その重要性は広く認められている(Delaney, 2008; Hidi & Anderson, 1986; Plakans, 2008)。なぜなら、特に大学生にとっては、「単独の課題として要約を書く」(Marshall, 2017, p. 71)ことや、様々な種類の原文を文章に組み込む他のタイプの課題をこなすことを求められることが多いためである。

## 2.3 英文要約の評価

英文要約は内容理解の測定方法の1つだが、その採点にはマクロ命題(Li, 2014)やアイデア・ユニット(Ushiro, et al., 2009)を用いた指標等があり、採点方法によって評価が異なり(高木, 2018)，意見の一一致を見ていない。また、英文要約はリーディングだけではなく、ライティング力の測定方法の1つでもある。Yamanishi, Ono, and Hijikata(2019)では、日本の大学で学ぶEFL学習者を対象に、分析的評価と全体的評価の両方を特徴とするL2の要約文の採点ルーブリックを開発した。特に総合的なqualityの側面と、paraphraseの2つの側面(paraphrase(量)とparaphrase(質))に焦点を当てている。Sawaki(2020)では、Bachman and Palmer(2010)の評価使用論証(Assessment Use Argument: AUA)フレームワークの観点から、低・中の影響力を持つ用途の2種類の要約内容得点法(内容得点と統合という全体的要約内容評価尺度)がどのように機能するかを検討した。結果として、AUAにおける評価記録のための得点一貫性の保証を支持することが示唆された。

## 2.4 英文解析ツールCRAT

CRATは、語彙の洗練さ、重複度、原文/要約テキストの重複に関する700以上のインデックスを含むツールである。CRATは、他の自動英作文評価ツールとは異なり、重複に関する指標を算出できるという利点があるため、要約に関連するライティングの質の探求に特に適している。CRAT

のインストールは NLP for the Social Sciences (<https://www.linguisticanalysistools.org/crat.html>) のウェブサイトから行うことができる。以下の解説は Crossley et al. (2016) より引用した。

CRAT で使用される指標は a) 原文と要約文の間の言語的・意味的な類似性, b) 要約文の言語的洗練度である。

類似性の指標としては、キーワードのオーバーラップ、同義語のオーバーラップ、潜在的意味解析 (LSA) の類似性を用いて算出される語彙的類似性と、主要な bigram (2語の塊) と trigram (3語の塊) のオーバーラップ、主要な品詞反応スロットグラム (例:into the \_\_\_ のようなスロットが開いている trigram) を用いて算出されるフレーズ的類似性がある。

洗練度指標には、心理言語学的な単語情報指標 (具体性や親しみやすさなど), British National Corpus (BNC) や Corpus of Contemporary American English (COCA) に基づく語彙頻度・range (より広い範囲のテキストに出現する単語) 指標、および構文カテゴリ (形容詞や名詞の数など) が含まれている。range はコーパスに収録されているテキストの中でいくつのテキストに出現しているかを示している。例えば have のような語彙は様々なテキストに出現する可能性が高いため、range は高くなる。しかし、テキストに出現する可能性が低い語彙は、範囲が限定的になるため、range が低くなる。そのため、range が低い語彙は洗練性が高いといえる。COCA については、academic texts, newspapers, fiction などの様々なジャンルについて、頻度や範囲を示す指標を提供している。これらの指標の特徴は以下の通りである。

#### (1) 機能および内容を表す単語のみの指標

CRAT の指標は、通常、テキスト内のすべての単語を考慮している。CRAT には、内容語 (名詞、動詞、形容詞、副詞など) や機能語 (決定詞、前置詞など) のみを含む指標のバリエーションもある。内容語指標と機能語指標は、より細かい分析を行うために設計されており、場合によっては、すべての単語を指標で考慮した場合よりも予測性が高いことが示されている。

#### (2) テキストおよび文の最小指標

CRAT の指標は一般的に、テキスト全体における特徴の平均スコアで構成される。また、CRAT は一文の長さしかない小さなテキストを評価するために、テキストの各文の平均最小スコアからなる指標バリエーションを計算する。

#### (3) キーワード除外指標

上記の指標に加えて、洗練度指標には、原文中に予想以上に頻繁に出現する単語 (すなわち、「キー」となる単語) を除外する変種が含まれている。このようなキーワード除外指標の変数は、原文の洗練された語彙が要約に与える影響を最小限にするために含まれている。これは CRAT 以外のツールを使ったときの語彙洗練度の結果とは異なるため、ほかのツールにはない利点である。

#### (4) 潜在的意味解析の重み付け

LSA の類似性スコアに影響を与える変量の一つに、採用された重み付けスキームがある。CRAT では、TASA コーパスから算出した LSA 変量に、Normalized weighting (正規化重み付け), Rare words dominated weighting (希少語優位の重み付け), Frequent words dominated weighting (頻出語優位の重み付け)を行っている。正規化重み付けは、参照コーパスに含まれるすべての単語を均等に考慮する。希少語優位の重み付けでは、参照コーパスの中で出現頻度の低い単語に高いスコアを与える。また、頻出語優位の重み付けは、参照コーパスに頻繁に出現する単語に高いスコアを割り当てる。

Kyle (2017) では、英語母語話者を対象とし、原文ベースのライティングの質の全体的なスコアを予測するために、語彙の重複、意味の重複、および語彙の洗練度に関する指標が使用された。CRAT 指標をステップワイズ回帰にかけたところ、10個の指標からなる有意なモデル ( $r = .587$ ,  $R^2 = .345$ ,  $p < .001$ ) が得られ、エッセイスコアの分散の34.5%を説明することができた。

主に CRAT を使った先行研究は対象が英語母語話者中心であり、L2 学習者や日本人英語学習者で行われた研究は、著者が調べた限りではない。そのため、本研究では、L2 や EFL 環境の学習

者のライティングにもCRATが利用可能かを検証する。

## 3

### 調査1

#### 3.1 目的

調査1では、1つの読解テキストに対して、複数の採点者によって作成された英文要約の模範解答を分析し、どのような特性の違いや共通点が見られるかを明らかにする。

#### 3.2 方法

##### 3.2.1 協力者

日本の大学に通う日本人大学院生と現職の英語教員、計3名が調査に参加した。協力者は全員日本語母語話者で、6年以上日本の教育機関で英語を学んでいた。協力者A、Bは英検準1級を保持しており、Cは長期留学経験があり、英検1級を保持していた。そのため模範解答を作成するレ

ベルの英語力を有していた。

##### 3.2.2 マテリアル

英検の過去問題(2015年第2回)より、準2級の長文問題を1つ選定した(表1)。

##### 3.2.3 手順

協力者3名で準2級のテキストの要約解答モデルの作成を行った。互いに相談することは許可せず、1人ずつ1モデルを作成した。30分程度で、語数の指定は行わなかった。

##### 3.2.4 分析指標

本調査では多くの指標の中でも、Kyle(2017)で要約文において重要な構成要素として挙げられたsource text use (i.e., lexical and semantic overlap)とlexical sophistication (i.e., word frequency and word information)に関連する10の指標を使用した(表2)。これらの指標をLexical overlap, Semantic overlap, Lexical sophisticationの3つに分け、次に詳しく説明する。

■表1: 使用されたテキスト

トピック	単語数	文数	FKGL
School of the Air	304	18	8.8

注) FKGL = Flesch-Kincaid Grade Level. 米国の学年に対応したテキストの読みやすさを示す。

■表2: 本研究で使用したCRATの指標

		値のとる範囲	望ましい要約がとる値
Lexical overlap	key verb overlap	0~1	低い値
	key adjective/noun bigram overlap	0~1	低い値
Semantic overlap	LSA all words normal	0~1	高い値
	noun-verb synonym overlap	0~1	高い値
Lexical sophistication	noun use percentage	0~1	低い値
	adverb use percentage	0~1	高い値
	imageability [content words, no keywords]	100~700	低い値
	age of acquisition [no keywords]	1以上	高い値
	concreteness [no keywords]	1~5	低い値
	word range [content words]	0~1	低い値

### Lexical overlap(語彙の重複度)

Lexical overlap指標は、タスクの回答に含まれる単語のうち、原文から直接コピーされた単語の数を測定する。これらの指標には、キーワードや、Corpus of Contemporary American English (COCA; Davies, 2009) のアカデミックセクションよりも原文に有意に多く出現する単語のみがカウントされる。Lexical overlap指標は、動詞、形容詞、名詞のオーバーラップを表す指標があり、品詞が考慮できる。また、主要な bigram, trigram, quadgram のオーバーラップ指標も含まれている。今回は key verb overlap (主要動詞の重複度) と key adjective/noun bigram overlap (原文からの主要形容詞 / 名詞bigram重複度) を使用する。

### Semantic overlap(意味の重複度)

英文要約を成功させるためのもう一つの重要な要素は、パラフレーズ（言い換え）である。パラフレーズの一つの側面は、原文とは異なるが、意味的に関連している単語を使ってアイデアを再提示することである。Semantic overlap指標は、要約内の単語が原文テキスト内の単語と意味的に関連している度合いを測定する。意味的重複の測定には、同義語の使用と潜在的意味分析 (LSA) のコサイン類似度という2種類の指標が使用される。同義語指標は、Wordnet (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990) の同義語セットに基づいている（形容詞、副詞、名詞、動詞に別々の指標がある）。LSAのコサイン類似度指標は、Touchstone Applied Science Associates (TASA) のベクトル空間 (Landauer, Foltz, & Laham, 1998) から抽出されており、3種類の重み付け（正規化重み付け、希少語重み付け、頻出語重み付け）に基づく類似度スコアが含まれている。今回は LSA all words normal (全ての語との意味的関連度) と noun-verb synonym overlap (名詞 - 動詞の同義語重複度) を使用する。

### Lexical sophistication(語彙の洗練度)

要約の成功には、洗練された言語、特に洗練された語彙項目を使用することが重要となる。単語の頻度や範囲、単語の心理言語学的特性（例：具体性）、単語の品詞（例：形容詞）などに関する指標が含まれる。頻度と範囲の指標

は、アメリカの映画やテレビの字幕のコーパス (SUBTLEXus) (Brysbaert & New, 2009) と、COCAのアカデミックセクション (Davies, 2009) から得られたものである。日常会話での頻度が低く、テキストに出現する範囲が限定的な単語は洗練されているとみなされる傾向がある (Kyle & Crossley, 2015; Laufer & Nation, 1995)。心理言語学的な単語の特性指標には、具体性、意味性、イメージ性、習得年齢などの情報が含まれる (Brysbaert, Warriner, & Kuperman, 2014; Coltheart, 1981; Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012)。品詞の指標は、テキストの中で特定の品詞（形容詞、副詞、名詞、代名詞、動詞）で構成されている割合を測定するものである。品詞の指標以外の語彙洗練度指標は、タスクの回答に含まれるすべての単語、機能語のみ、内容語のみを考慮して計算される。また、タスクの回答に含まれる原文のキーワードを無視した指標も計算される。今回は、concreteness [no keywords] (具体性 [キーワードなし]), noun use (名詞の使用), adverb use (副詞の使用), word range [content words] (単語の範囲 [内容語]), age of acquisition [no keywords] (習得年齢 [キーワードなし]), imageability [content words, no keywords] (心像性: イメージのしやすさ [内容語, キーワードなし]) の指標を使用する。

## 3.3 手順

テキスト分析ツールのCRAT (<https://www.linguisticanalysistools.org/crat.html>) を用いて採点者が作成した英文要約の模範解答例を分析した。

- ①事前に Java Development Kit (JDK) をインストールする (<https://www.oracle.com/java/technologies/downloads/>)。
- ②NLP for the Social Science のウェブサイトから、Mac, Windows どちらかの OS のファイルをダウンロードする (Linux は2022年8月現在開発中のため使用不可)。
- ③すべての入力ファイルは、いかなる種類のマーカアップ（例: XML, HTML など）を含まないテキストファイル (.txt) で行う。
- ④2種類の入力を行う。まず、原文テキストをアッ

プロードする。次に、協力者の要約文を入力する(この際、要約テキストファイルは、1つのフォルダに配置しなければならない)。その後選択された入力フォルダ内のすべての.txtファイルが自動で処理される(フォルダ内の.txtファイルは複数可)。

- ⑤CRAT は、表計算ソフトで開くことができるカンマ区切り(.csv)ファイルの形で出力される。しかし、ファイルの保存はExcel ブック(.xlsx)でも行うことが勧められる(データの保存が正常にされない可能性があるため)。

### 3.4 結果と考察

表3はCRATで算出した指標をテキストごとに集計した記述統計である。以下では指標ごとに考察をしていく。

まず、Lexical overlap(語彙の重複度)はkey verb overlap(主要動詞の重複度)とkey adjective/noun bigram overlap(原文からの主要形容詞/名詞bigram重複度)の指標があり、0～1の値をとる。これらの指標は、英文要約において原文から直接語彙を借用することは有益ではないため、値が低いことが望ましい。表を見ると、Cの模範解答は両方の指標の値が低く、原文のコピーが少ないため、模範解答にはふさわしいと判断できる。AもBもCほどではないが低くなっている。多少のばらつきは見られたが、どの解答も直接的な語彙の借用は0ではないが、あまり見られなかった。

次に、Semantic overlap(意味の重複度)はLSA all words normal(全ての語との意味的関連度)とnoun-verb synonym overlap(名詞・動詞の同義語重複度)の指標があり、0～1の値をとる。原文の単語と意味的に関連する単語を含んでいたり、原文の名詞や動詞と同義語の名詞や動詞を含んでいたりすることが望ましいとされる。原文と似た単語を使うこと、つまりパラフレーズ(言い換え)が、要約のストラテジーとして有効であるためである。3つの間にはどちらの指標もあまり差が見られず、3つともどちらの指標も高く、同義語や似た意味の言葉にパラフレーズできている可能性がある。

最後にLexical sophistication(語彙の洗練度)

は concreteness [no keywords](具体性[キーワードなし]), noun use(名詞の使用), adverb use(副詞の使用), word range [content words](単語の範囲[内容語]), age of acquisition [no keywords](習得年齢[キーワードなし]), imageability [content words, no keywords](心像性:イメージのしやすさ[内容語, キーワードなし])の指標がある。具体性に関して、英文要約では、具体的な内容や詳細情報はまとめられて上位語におかれるため、具体的な単語が少ない要約が望まれる。値は1に近いほど抽象性が高くなり、5に近いほど具体性が高くなる。表ではBの具体性が他よりやや高い。具体的な情報をまとめたりする一般化は質の高い要約には必要不可欠であるため、上位語への変換や命題のまとめが必要になる可能性がある。さらに、イメージのしやすい語彙は簡単なものになりがちなので、心像性は低いことが望ましいとされている。全体的に、今回の3つの模範解答の心像性は低い傾向にあった。また、3つの中で多少のばらつきがみられたため、AとBの解答はより詳細な指標の検証が必要かもしれない。また、名詞の使用的割合が少なく、副詞の多使用が質の高い要約文の予測因子になることが報告されている(Kyle, 2017)。しかしどの解答でも副詞の使用割合は少なく、名詞の割合はほぼ一緒であった。これはテキストの難易度やトピックによって異なりが出る指標だと考えられる。age of acquisition(その語を母国者が習得する年齢のこと)は高い年齢で習得するほど洗練された語彙であるため、値が高いことが望まれる。今回の結果からは、5～6歳くらいの母語話者が習得する語彙を平均的に使用していることがわかった。英検のモデル英作文のテキスト分析を行った小室(2021)では、age of acquisitionについて、英検1級の平均が7.39、英検準1級の平均が6.69、2級の平均が5.60であることが示されている。小室(2021)では TAALES を用いており、本研究で用いたCRATでの洗練語彙の指標と同じではないが、ある程度比較をすることは可能である。この値と比較すると、今回の結果は、5.85～6.11であり、英検準1級と2級のモデル英作文の中間ほどの値であることがわかった。また、今回の3つの模範解答の間にはほとんど差は見られなかった。異なるテ

キストにおいては、級を分ける重要な指標になるが、同じ内容のテキストの比較では差が出ない指標であることが示唆された。語彙の範囲も語彙の洗練度には重要であり、より範囲の狭い語彙が使用されている要約文は質が高いとされている。例えば、haveやmakeのような語彙は頻出なので値は高くなる。そのため、値が低い語彙は範囲が限定的である。

そのため、洗練性が高い。AとBはCよりも値が低いため、Cは洗練性が高いことがわかる。

また、以上の指標を考慮すると、全体的に協力者CはA、Bに比べてCRATの指標の値が優れていた。Cは英語熟達度が他の2人よりも高かったため、より質の高い要約例を書くことができた可能性が示唆された。

表3: 英文要約例をCRATで算出した指標の記述統計

		解答A	解答B	解答C	$M(SD)$
Lexical overlap	key verb overlap	.13	.11	.05	.10 (.04)
	key adjective/noun bigram overlap	.04	.13	.09	.09 (.05)
Semantic overlap	LSA all words normal	.90	.92	.89	.90 (.07)
	noun-verb synonym overlap	.91	.93	.92	.93 (.07)
Lexical sophistication	noun use percentage	.23	.28	.26	.26 (.03)
	adverb use percentage	.07	.04	.06	.06 (.02)
	imageability [content words, no keywords]	207.83	174.33	152.08	178.08 (28.06)
	age of acquisition [no keywords]	5.85	5.97	6.11	5.98 (.13)
	concreteness [no keywords]	2.74	3.01	2.75	2.83 (.15)
	word range [content words]	.30	.33	.27	.30 (.07)

### 3.5 調査1のまとめ

調査1では、1つのテキストに対して、複数の採点者によって作成された英文要約の模範解答例にどのような特性の違いが見られるかを明らかにするために、CRATを用いて分析を行った。結果より、差がないところもあったが、模範解答によって多少のばらつきが見られたところもあった。そのため教師は主觀で模範解答を作成するのではなく、CRATを使いながら模範解答を修正するなど、客観的な視点も取り入れ、より質の高い解答モデルを作成することが求められる。また、結果より、要約の模範解答の質に英語熟達度が関係することが示唆された。調査2では異なる級の読解テキストを用いた英文要約の分析を、CRATを用いて分析する。また、英文要約と学習者のライティング能力の関係を調査した。

## 4 調査2

### 4.1 目的

調査2では、日本人英語学習者の英作文をCRATを用いて分析し、算出された指標の特徴やその指標と学習者の英語力にどのような関係があるのかを調査する。

検証課題(RQs)は以下のとおりである：

RQ2-1

学習者の英文要約におけるCRATで算出された指標の平均値は、テキストの受験級ごとに違いが見られるか。

RQ2-2

学習者の英文要約におけるCRATで算出された指標の中で、学習者のライティング力と関係がある指標はどれか。また、テキストの受験級ごとに違いが見られるか。

#### 4.1.1 協力者

公立の大学に通う日本人英語学習者51名が調査に参加した。協力者は大学1・2年生であり、専攻は経営・経済系であった。協力者は全員日本語母語話者で、少なくとも6年以上日本の教育機関で英語を学んでいた。自己申告アンケートプロフィールによると、協力者は英検3級から準2級程度であり、CEFRのA1からA2レベルであると推察される(日本英語検定協会、2012)。実験は3週間に分けて行われたため、全ての実験を完遂した51名を分析対象とした。

■表4: 使用されたテキスト

トピック	単語数	文数	FKGL
3級 Rick Hansen	258	21	7.8
準2級 School of the Air	304	18	8.8
2級 Asteroid Mining	371	15	11.7

#### (2) 英作文タスク

英検の過去問題より、準2級(2017年第1回)、2級(2016年第1回)から1つずつトピックを選定した。選定したトピックは次の表5のとおりである。指定語数はそれぞれの級にならって、準2級は50

#### 4.1.2 マテリアル

##### (1) 英文要約タスク

英検の過去問題より、3級(2013年第3回)、準2級(2015年第2回)、2級(2013年第2回)の長文問題から3つのトピックを選定した(表4)。受験級に関しては、大学1年生を高校卒業程度とし、2級までとした。準2級のテキストは調査1で用いたのと同じものとした。

語から60語、2級は80語から100語とした。指示は実際の英検の出題形式に則り、「以下の質問について、あなたの考え方とその理由を2つ英文で書きなさい」とし、2級のトピックではポイントを3つ与えた。解答は紙ベースで行われた。

■表5: 英作文の受験級とトピック

受験級	トピック
準2級	Do you think it is better for people to eat at restaurants or at home?
2級	Today, some companies allow their employees to wear casual clothes like jeans or T-shirts. Do you think the number of such companies will increase in the future?

#### 4.1.3 手順

調査は4週間に分けて実施された。1～3週目に英文要約タスクを行った。事前に要約の手順を説明したのち、各タスクとも30分間で行った。協力者の英語熟達度を考慮し、辞書の使用を許可した。4週目にはライティングタスクを行った。それぞれ30分間行われ、協力者は実際の試験を想定し、辞書等で単語を調べることはできなかった。

者が取り組んだ各級の英文要約を分析した。ライティングタスクについては、2つのトピックをそれぞれ8点(内容・構成・語彙・文法の配点を各2点)とし、その合計を最終スコアにした。採点は日本人大学生と2名で行った。評価者間信頼性は87.66%と十分な一致率が見られた。

#### 4.1.4 採点・分析

英文要約タスクに関しては、調査1と同様、テキスト分析ツールであるCRATを用いて、学習

#### 4.2 結果と考察

##### (1) 英文要約タスク

英文要約タスクにおいて、CRATで算出した指標を受験級ごとに集計した記述統計を表6に示す。

それぞれの指標ごとに考察をしていく。

まず, Lexical overlap(語彙の重複度)は key verb overlap(主要動詞の重複度)と key adjective/noun bigram overlap(原文からの主要形容詞 / 名詞bigram重複度)の指標がある。表を見ると、読解テキストの受験級が上がるにつれてどちらの値も増加し、英文要約の質としては低くなる傾向が見られた。この結果により、級が上がるにつれて内容が難しく、馴染みがないトピックになり原文をそのまま使用してしまう可能性があることが示唆された。

次に、Semantic overlap(意味の重複度)は LSA all words normal(全ての語との意味的関連度)と noun-verb synonym overlap(名詞-動詞の同義語重複度)の指標がある。準2級と2級の間にはどちらの指標もあり差が見られないが、3級はどちらの指標も高く、同義語や似た意味の言葉にパラフレーズをほかの受験級よりも行うことができている可能性がある。

最後に Lexical sophistication(語彙の洗練度)は concreteness [no keywords](具体性 [キーワードなし]), noun use(名詞の使用), adverb use(副詞の使用), word range [content words](単語の範囲 [内容語]), age of acquisition [no keywords](習得年齢 [キーワードなし]), imageability [content words, no keywords](心像性: イメージのしやすさ [内容語, キーワード

なし])の指標がある。表では受験級が上がるにつれて、具体性がやや上がっているのが見られる。キーワードは除外されてはいるが、これは受験級が上になればなるほど文章が長く難しくなり、具体的な情報をまとめるなどの一般化が困難になっていることが示唆される。さらに、級が上がるにつれて、心像性が下がっているのが見られる。これは受験級が上がると語彙が難しくなるためであり、3級ではテキストにある語彙の心像性がそもそも高いと考えられる。また、名詞と副詞の割合に関して、どの級でも副詞の使用割合は少なく、3級が名詞の割合が一番高かった。これは他の級よりもテキストが易しいため、内容語の中でもより基本的な品詞が中心になっているからだと考えられる。age of acquisition(その語を母国者が習得する年齢のこと)は級が上がるにつれて、上昇する傾向にある。語彙の範囲に関しては、級が上がるにつれて洗練度の値は下がっているため、洗練度が上昇し、2級が一番洗練度が高くなっている。受験級が上がった結果、語彙が難しくなり、範囲が限定的な語彙が現れた可能性がある。しかし、[no keywords] の指標に関しては、洗練度が高い語が使われたといっても、原文にあるキーワード以外の語、もしくは受験者が自分で使った語の可能性がある。そのためより詳しくみるには、指標の値に加え、原文と受験者の要約を比較しながらの吟味が必要になる。

■表6: 英文要約例をCRATで算出した指標の記述統計

		3級		準2級		2級	
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Lexical overlap	key verb overlap	.10	.05	.15	.05	.17	.06
	key adjective/noun bigram overlap	.13	.05	.15	.04	.19	.06
Semantic overlap	LSA all words normal	.94	.02	.88	.04	.89	.04
	noun-verb synonym overlap	.89	.05	.85	.06	.84	.06
Lexical sophistication	noun use percentage	.31	.03	.29	.03	.24	.04
	adverb use percentage	.07	.02	.07	.02	.08	.03
	imageability [content words, no keywords]	270.34	31.48	268.18	25.99	261.95	31.48
	age of acquisition [no keywords]	4.84	.23	5.40	.27	6.02	.28
	concreteness [no keywords]	2.90	.17	2.93	.14	2.98	.15
	word range [content words]	.42	.04	.39	.09	.35	.05

また、調査1で行われた英検準2級のモデル英作文の指標と比較すると、調査1で作成されたモデル要約の結果は、調査2で得られた結果よりもより優れている傾向にあることがわかった。例えば語彙の重複度では、調査1の方が調査2よりも原文からの借用が少なかった。意味の重複度では、調査1の方が、値が高いことから、より同義語や似た意味の語に言い換えることができている。語彙の洗練度においても、調査1の方がage of acquisitionが高く、具体性や心象性が低いことから、より洗練された語を使用していることが

わかる。調査1の協力者の中でも熟達度の高い協力者が他の協力者よりも指標の値がすぐれています。今回調査2の協力者と調査1の協力者を比べた際に、調査1の協力者の方が、熟達度が優れているため、やはり英語熟達度が高い協力者は質の高い要約を作成することができる事がわかった。この結果から、CRATの指標が熟達度の差をある程度弁別できる可能性が示された。

## (2) ライティング能力測定タスク

表7はライティングテストの記述統計である。

■表7: ライティングテストの記述統計

	<i>M</i>	95%CI	<i>SD</i>	<i>Min</i>	<i>Max</i>	評価者間信頼性
ライティング	10.04	[9.23, 10.85]	2.88	4.00	15.00	.87

### (3) CRATで算出された指標と学習者の

#### ライティング力の関係

それぞれの級において、協力者のライティング力を従属変数に、有意な相関関係が見られた変数を独立変数にステップワイズ法による重回帰分析を行った。ステップワイズ法は統計的に最も予測率が高いと考えられる変数から順に自動的に投入される方法である。まず、最も従属変数と相関の高い独立変数が投入され、その後、偏回帰係数の有意性が次に最も高くなる独立変数が選ばれ、順に投入されていく。ステップワイズ法は統計的に最も予測に寄与する独立変数が順に投

入される。よって、最終的に適合度が最良の重回帰式を調べる際に使用するが、あくまでも統計的な根拠に基づいて投入されるため、投入された独立変数が理論にかなっているかは別途判断する必要がある(平井・岡・草薙, 2022, pp. 244-245)。

分析を行う前に、多重共線性の確認を行った。VIF(variance inflation factor)値は1.09であり、相関係数が.80を上回る変数もなかった。重回帰分析の結果3級～2級のそれぞれの級に関して、以下の結果が得られた。

$$3\text{級} \quad F(1, 50) = 2.15, p = .15, R^2 = .181, \text{adjusted } R^2 = .176$$

$$\text{準2級} \quad F(2, 50) = 2.47, p = .09, R^2 = .231, \text{adjusted } R^2 = .223$$

$$2\text{級} \quad F(2, 50) = 5.33, p = .008, R^2 = .154, \text{adjusted } R^2 = .145$$

それぞれの級の変数選択によって選択された指標が表8から表10に示されている。また、Appendix 1と2に、それぞれの受験級ごとの要約の指標とライティングの相関行列が示されている。

結果より、3級においては、CRATにおける指

標が学習者のライティング力を18%予測することができ、準2級においては、ライティング力を23%、2級においては、15%予測することができた。英語母語話者のライティングを分析したKyle(2017)では34.5%であり、それと比較すると、予測力は低かった。3級においては、noun-

verb synonym overlap(名詞一動詞の同義語重複度)が高い協力者はライティング力が高いことがわかる。準2級においては, noun use と key adjective/noun bigram overlap の  $\beta$  の値が負であるため, 名詞の使用が少ない協力者ほど, また原文からの主要形容詞 / 名詞bigram 重複度が低い, つまり原文からの主要形容詞 / 名詞のコピーが少ない協力者ほど, ライティング力が高いことがわかる。

2級においては, adverb use percentage が高い, つまり副詞の使用が多い協力者はライティング力が高いことがわかる。対して, imageability [content words, no keywords] の  $\beta$  の値が負であるため, そのスコアが小さい協力者, つまりイメージのしにくく難しい語彙を使う協力者ほど, ライティング力が高いことがわかる。まとめると, それぞれの受験級において, 固有の予測変数は存在するが共通した予測変数は見つからなかった。しかし, すべての  $\beta$  の符号から, 要約として質が高い方向に指標の値が高くなっていた学習者が, 一般的な英語ライティング力が高い傾向があるという結果になった。また, トピックや, 求められる制限語数によって, ライティング力を予測す

る指標は異なる可能性あるが, 3級では Semantic overlap, 準2級では Lexical overlap, Lexical sophistication, 2級では Lexical sophistication が残っており, 一般的な英語ライティング力を予測するのは, 要約の中のこれらのカテゴリーという可能性が示された。

### 4.3 調査2のまとめ

調査2では, CRAT で算出された指標の英文要約評価への利用可能性を検証するために, 日本人英語学習者の英文要約に対する CRAT の指標を算出し, その指標のライティング力の予測率を調査した。結果として, RQ2-1については, 受験級が上がるにつれて原文からの借用が多くなることがわかった。また, 英文要約はテキストに依存するため, 使用される語彙の洗練度も高くなった。また, RQ2-2については, それぞれの受験級において, 固有の予測変数は存在するが共通した予測変数は見つからなかった。これにより, トピックや, 求められる制限語数によって, ライティング力を予測する指標は異なる可能性が示唆された。

■表8: 3級英文要約のCRAT指標を用いたライティング力の予測:ステップワイズ法による重回帰分析の結果

	B	SE B	95%CI	$\beta$	t	p
noun-verb synonym overlap	28.30	19.29	[-10.46, 67.05]	.21	1.47	.15

■表9: 準2級英文要約のCRAT指標を用いたライティング力の予測:ステップワイズ法による重回帰分析の結果

	B	SE B	95%CI	$\beta$	t	p
noun use	-18.52	11.08	[-40.81, 3.76]	-.23	-1.67	.10
key adjective/noun bigram overlap	-9.31	6.30	[-21.99, 3.37]	-.20	-1.48	.15

■表10: 2級英文要約のCRAT指標を用いたライティング力の予測:ステップワイズ法による重回帰分析の結果

	B	SE B	95%CI	$\beta$	t	p
adverb use percentage	29.48	10.88	[7.60, 51.37]	.35	2.71	.01
imageability [content words, no keywords]	-.03	.02	[-.06, 0.00]	-.24	-1.87	.07

## 5

## 結論と今後の課題

本研究では、無料で使用可能な自動英文解析ツールであるCRATの妥当性に関して2つの調査を行ってきた。まず、調査1では複数の熟達度が高い協力者によって作成された英文要約の解答モデルのテキスト分析を行い、モデル英文要約の特徴を検証した。結果として、モデル間に差がないところもあったが、模範解答によって多少のばらつきが見られたところもあった。

調査2では、受験級ごとの日本人英語学習者の英文要約の分析を行い、算出された指標の傾向とライティング能力との関係を調査した。結果として、読解テキストの受験級が上がるにつれて原文からの語彙的、意味的な借用が多くなることがわかった。また、使用される語彙の洗練度も高くなる結果となった(ともにRQ2-1の結果)。さらに、算出された指標と学習者のライティング力の関係に関しては、受験級によって、ライティング力と関係のある指標が異なることがわかった(RQ2-2の結果)。求められるトピックや語数が異なることで、ライティング力を予測する指標が変わる可能性が示唆された。また、調査1と2から副詞の使用が少ないことが確認できた。これは日本人英語学習者が苦手とする指標である可能性がある。そのため、今後の研究でさらなる検証の余地がある。

以上の本研究の結果より得られた教育的示唆としては、以下の点が挙げられる。まず、調査1の結果より、同じ受験級のテキストであっても人によって指標に多少のばらつきがあることが示唆された。そのため、CRATが、自分で模範解答の作成をする英語教師に客観的な視点を提供し、より精巧な模範解答作成に貢献するだろう。CRATで算出された数値を見ることによって、より客観的な模範解答を作成することや、客観的な指標を用いて、生徒へフィードバックすることが可能となる。

また、調査2の結果から、英文解析ツールを用いた英文要約の自動評価には、重複度に関する指標、具体的にはnoun-verb synonym overlapとkey adjective/noun bigram overlapが有効で

ある可能性が示唆された。CRATは他の自動英文評価ツールとは異なり、オーバーラップの指標を算出できるという利点がある。したがって、生徒の英文要約を評価する際には、語彙的や意味的な重複度に関する指標を評価やフィードバックに有効利用できると考えられる。

本研究は、ダウンロードすればだれでも利用可能で、読解研究に利用されているCRATを用いた英文要約自動評価の利用可能性を検証することを目的に行われた。十分に利用可能性を検証することはできなかったが、CRATを用いた英文要約の自動評価の可能性を中心に議論を進めた。一方、本研究には以下のような限界点がある。

まず、各級ごとのテキストのトピックが英文要約に影響を与えた可能性がある。先述したように、書き手はテキストに含まれている語彙や表現を使用する傾向が高い。したがって、今後の研究では、同じトピックで制限語数が異なる課題や幅広いトピックで同じ語数の要約課題を使用したり、結果を一般化する手立てを考えたりする必要がある。また、英語ライティング力の測定についても、過去の英検問題2問を使用したが、その測定が適切だったかについてはさらに検討が必要である。さらに今回のライティング力の測定は、技能統合型ではなく独立型を採用していたため、統合型のライティングでの検証も必要である。また本研究ではCRATで算出可能な指標の中で、Kyle(2017)で重要と示された10個の指標を用いたが、他の指標を用いることで結果が変わることも考えられる。以上のような限界点を考慮し、研究を進めていくことで、英文要約自動評価の利用可能性についてより説得力のある示唆を出すことができると考えられる。次に調査サンプルに関する点である。本調査の協力者は限定的であったため、この結果を日本の英語学習者の傾向として一般化することは難しい。最後にCRATの問題点について述べる。700を超える膨大な数の指標が算出されるため、かなり詳細に分析を行うことができるが、それぞれの指標が何を指しているのか理解に時間がかかり、意味をなかなか捉えることができない。そのため、ユーザーにとってわかりやすいチュートリアルやツールの開発を期待する。

## 謝 辞

本研究の実施、発表にあたりまして、公益財団法人 日本英語検定協会と関係者の皆さん、並びに選考委員の先生方からのご支援をいただき、心より御礼申し上げます。特に、助言者である小泉利恵先生には、研究の実施および報告書の執筆にあたりご指導いただきましたこと、深く感謝申し

上げます。また、筑波大学の卯城祐司先生をはじめ、研究室の先輩・同輩・後輩の皆さんには、本研究の立案から実施、報告書の執筆にあたりまして、ご助言いただきました。心より感謝申し上げます。最後に、本調査に協力してくださった協力者の皆さんに厚く御礼申し上げます。

## 参考文献(\*は引用文献)

- \* Bachman, L. F. & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.
- \* Bax, S. (2012). Text Inspector, online text analysis tool. <https://textinspector.com/>
- \* Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4), 977-990. <http://doi.org/10.3758/BRM.41.4.977>
- \* Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3), 904-911. <https://doi.org/10.3758/s13428-013-0403-5>
- \* Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, 33(4), 497-505. <https://doi.org/10.1080/14640748108400805>
- \* Crossley, S. A., Kyle, K., Davenport, J., & McNamara, D. S. (2016). Automatic assessment of constructed response data in a chemistry tutor. In T. Barnes, M. Chi, & M. Feng (Eds.), *Proceedings of the 9th International Educational Data Mining (EDM) Society Conference* (pp. 336-340). <https://files.eric.ed.gov/fulltext/ED592609.pdf>
- \* Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. *Language Assessment Quarterly*, 10(1), 1-8. <https://doi.org/10.1080/15434303.2011.622016>
- \* Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990-2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14(2), 159-190. <http://doi.org/10.1075/ijcl.14.2.02dav>
- \* Delaney, Y. A. (2008). Investigating the reading-to-write construct. *Journal of English for academic purposes*, 7(3), 140-150. <https://doi.org/10.1016/j.jeap.2008.04.001>
- \* Gebril, A., & Plakans, L. (2014). Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks. *Assessing writing*, 21, 56-73. <https://doi.org/10.1016/j.asw.2014.03.002>
- \* Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments*, & Computers, 36, 193-202. <https://doi.org/10.3758/BF03195564>
- \* Hidi, S., & Anderson, V. (1986). Producing written summaries: Task demands, cognitive operations, and implications for instruction. *Review of Educational Research*, 56(4), 473-493. <https://doi.org/10.2307/1170342>
- \* 平井明代・岡秀亮・草薙邦広 (2022).『教育・心理系研究のためのRによるデータ分析—論文作成への理論と実践集』東京図書
- \* Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological review*, 85(5), 363-394. <https://doi.org/10.1037/0033-295X.85.5.363>
- \* Kirkland, M. R., & Saunders, M. A. P. (1991). Maximizing student performance in summary writing: Managing cognitive load. *TESOL Quarterly*, 25(1), 105-121. <https://doi.org/10.2307/3587030>
- \* Knoch, U., Fairbairn, J., & Jin, Y. (2021). *Scoring second language spoken and written performance: Issues, options and directions*. Equinox.
- \* 小室竜也 (2021).「ライティングタスク(技能独立型 vs. 統合型)が発表語彙とその測定に与える影響: TAALESによる語彙の洗練性分析を基に」*EIKEN BULLETIN*, 32, 13-30. [https://www.eiken.or.jp/center\\_for\\_research/list\\_1X/32/](https://www.eiken.or.jp/center_for_research/list_1X/32/)
- \* Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44(4), 978-990. <https://doi.org/10.3758/s13428-012-0210-4>
- \* Kyle, K. (2017). Modelling Quality in Source-Based Writing Tasks. Retrieved from [https://a4li.sri.com/archive/papers/Kyle\\_2017\\_Writing\\_Quality.pdf](https://a4li.sri.com/archive/papers/Kyle_2017_Writing_Quality.pdf)
- \* Kyle, K. (2020). The relationship between features of source text use and integrated writing quality. *Assessing Writing*, 45, 100467. <https://doi.org/10.1016/j.asw.2020.100467>
- \* Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, 49(4), 757-786. <http://doi.org/10.1002/tesq.194>
- \* Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2-3), 259-284. <https://doi.org/10.1080/01638539809545028>
- \* Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied*

## 参考文献 (\*は引用文献)

- Linguistics*, 16(3), 307–322. <http://doi.org/10.1093/applin/16.3.307>
- \* Li, J. (2014). Examining genre effects on test takers' summary writing performance. *Assessing Writing*, 22, 75–90. <https://doi.org/10.1016/j.asw.2014.08.003>
- \* Marshall, S. (2017). *Advance in Academic Writing: Integrating Research, Critical Thinking, Academic Reading and Writing*. Pearson ELT, Canada
- \* Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International journal of lexicography*, 3(4), 235–244. <https://doi.org/10.1093/IJL%2F3.4.235>
- \* 日本英語検定協会 (2012). 「英検とCEFRとの関連性について 研究プロジェクト報告」Retrieved from [https://www.eiken.or.jp/eiken/group/result/pdf/report\\_02.pdf](https://www.eiken.or.jp/eiken/group/result/pdf/report_02.pdf)
- \* Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. *Assessing Writing*, 13(2), 111–129. <https://doi.org/10.1016/j.asw.2008.07.001>
- \* Plakans, L. (2015). Integrated second language writing assessment: Why? What? How?. *Language and Linguistics Compass*, 9(4), 159–167. <https://doi.org/10.1111/LINC3.12124>
- \* Sawaki, Y. (2020). Developing summary content scoring criteria for university L2 writing instruction in Japan. In G. Ockey & B. Green (Eds.), *Another generation of fundamental considerations in language assessment: A festschrift in honor of Lyle F. Bachman* (pp. 153–171). Springer. [https://doi.org/10.1007/978-981-15-8952-2\\_10](https://doi.org/10.1007/978-981-15-8952-2_10)
- Stein, B. L., & Kirby, J. R. (1992). The effects of text absent and text present conditions on summarization and recall of text. *Journal of Reading Behavior*, 24(2), 217–232. <https://doi.org/10.1080/10862969209547773>
- \* 高木修一 (2018). 「英文読解に対する日本語要約課題における採点方法の比較」*TELES Journal*, 38, 27–38.
- Ushiro, Y., Hamada, A., Hasegawa, Y., Nahatame, S., Kimura, Y., Shimizu H., Takaki, S., Kobayashi, M., & Tanaka, N. (2013). Discourse-based lexical inferencing in EFL reading: Focusing on depth of vocabulary knowledge and cue availability. *ARELE: annual review of English language education in Japan*, 24, 77–92. [https://doi.org/10.20581/arele.24.0\\_77](https://doi.org/10.20581/arele.24.0_77)
- \* Ushiro, Y., Kai, A., Nakagawa, C., Watanabe, F., Hoshino, Y., & Shimizu, H. (2009). Effects of Reading Perspective on EFL Learners' Summary Writing and Importance Rating. *ARELE: Annual Review of English Language Education in Japan*, 20, 11–20. [https://doi.org/10.20581/arele.20.0\\_11](https://doi.org/10.20581/arele.20.0_11)
- \* Yamanishi, H., Ono, M., & Hijikata, Y. (2019). Developing a scoring rubric for L2 summary writing: a hybrid approach combining analytic and holistic assessment. *Language Testing in Asia*, 9(13), 1–22. <https://doi.org/10.1186/s40468-019-0087-6>
- Yu, G. (2008). Reading to summarize in English and Chinese: A tale of two languages? *Language Testing*, 25(4), 521–551. <https://doi.org/10.1177/0265532208094275>

**資料****Appendix 1****調査2: 3級(左)と準2級(右)の英文要約の指標とライティングの相関行列**

	1	2	3	4	5	6	7	8	9	10	11
1. Key verb overlap	—	-.03	.18	-.02	-.22	.09	-.01	.04	.16	.07	-.12
2. Key adjective/noun bigram overlap	.06	—	-.32	.12	-.01	.22	.08	.00	.08	-.36	-.20
3. LSA all words normal	.21	-.39	—	.46	-.07	.25	-.22	.16	-.11	.08	-.10
4. Noun-verb synonym overlap	-.01	.05	.20	—	-.35	-.01	.30	-.05	-.07	-.27	-.02
5. Noun use percentage	.38	-.40	.05	-.36	—	-.08	.37	-.23	.32	-.24	-.23
6. Adverb use percentage	-.29	-.33	.00	.23	-.23	—	-.27	.33	-.17	-.41	.02
7. Imageability [content words, no keywords]	.17	.14	.15	-.34	.35	-.59	—	-.69	.49	.21	-.17
8. Age of acquisition [no keywords]	.18	-.14	-.19	-.02	-.28	.22	-.39	—	-.35	-.10	-.01
9. Concreteness [no keywords]	.24	-.21	.35	-.28	.24	-.42	.67	-.38	—	-.06	-.08
10. Word range [content words]	-.30	.02	.10	.04	-.21	.21	.19	-.43	-.23	—	-.02
11. Writing	-.10	.09	-.10	.21	-.02	.09	-.12	.17	-.18	-.05	—

**Appendix 2****調査2: 2級英文要約の指標とライティングの相関行列**

	1	2	3	4	5	6	7	8	9	10	11
1. Key verb overlap	—										
2. Key adjective/noun bigram overlap	.14	—									
3. LSA all words normal	-.12	.08	—								
4. Noun-verb synonym overlap	.29	.14	-.04	—							
5. Noun use percentage	.01	-.05	-.17	-.17	—						
6. Adverb use percentage	-.02	-.01	-.03	.08	-.01	—					
7. Imageability [content words, no keywords]	.12	.14	-.09	.42	-.10	.02	—				
8. Age of acquisition [no keywords]	.05	.05	-.19	.04	.05	.27	.17	—			
9. Concreteness [no keywords]	-.07	.08	.27	-.06	-.06	-.08	.02	.05	—		
10. Word range [content words]	-.08	-.21	.02	-.01	.08	-.18	-.07	-.46	-.25	—	
11. Writing	-.02	-.10	-.04	-.01	.09	.35	-.24	.06	-.13	-.04	—