英語学習者のエッセイに見られる過剰使用語
―事前準備ありと即興のエッセイにおける差異―

松田紀子*1・石井隆之*2・岩田雅彦*3・西美都子*4・濱崎佳子*5・林美登利*6

Overused Words in the Essays of English Learners:
The Difference between the Essays with and without Advance Preparation

Noriko MATSUDA, Takayuki ISHII, Masahiko IWATA, Mitsuko NISHI
Yoshiko HAMAZAKI, Midori HAYASHI

Abstract

In order to examine the degree of versatility in the data from learners' essays collected at the Faculty of Applied Sociology at Kindai University, the present study analyzes learner-specific overused words using essays with and without advance preparation. In the former case, learners were allowed to use dictionaries and other reference materials, while learners in the latter case improvised. We compared the results with findings from the large learner corpora, such as ICNALE and NICER, and found that essays without advance preparation that used similar data-collection factors as the large learner corpora contained more commonly overused words. The same results were obtained in the proficiency-based analysis. The results highlight the significance of controlling data-collection factors and demonstrate a high versatility of the data, which will contribute to creating a new learner corpus in the future.

Keywords：① Overused Words　② Learner Corpus　③ Data-collection Factors

1．はじめに

近畿大学の総合社会学部では，1年生で必修となる基幹科目の「英語演習1」（前期）と「英語演習2」（後期）において，発信型スキル（ライティングやスピーキング等）の向上を目指している．「英語演習」では週に2回ある授業のうち，1回はSocio ICT Stations（SIS）と呼ばれるコンピュータールームを使用し，パラグラフ・ライティングの基礎を学ぶ．その際，オンライン上で自動採点ができる*Criterion*®[1]という英語ライティング指導支援ツールを使用している．本学部の「英語演習1」（前期）と「英語演習2」（後期）では，中間試験及び期末試験を含めて，この*Criterion*®を年間計10回使用することになっているのだが，前期と後期の期末試験では，実施方法，モードの違いやトピック（*Criterion*®ではPromptと表記）の選択方法に大きな違いがある．

実施方法に関しては，どちらも30分という時間制限があり，前期の期末試験は辞書等の参考資料[2]を使用して事前に準備したものを覚えて画面に打ち込むのに対して，後期の期末試験はその場でトピックが与えられ，参考資料を使用せずに即興で画面に打ち込むことに

受付：令和2年11月6日　受理：令和3年1月8日
*1 近畿大学総合社会学部　講師，*2 近畿大学総合社会学部　教授
*3 近畿大学総合社会学部　非常勤講師，*4 近畿大学総合社会学部　非常勤講師
*5 近畿大学総合社会学部　非常勤講師，*6 近畿大学総合社会学部　非常勤講師

なっている．また，期末テストのエッセイのモードは，前期ではNarrative（物語型）またはDescriptive（記述型），後期はPersuasive（説得型）となっており，トピックの選択方法に関しては，前期は*Criterion*®に設定してあるPromptから各教員が選べるのに対し，後期はPromptが限定されており，各教員が選べないようになっている．

現在，教員は提出された学生のエッセイをオンラインでみることができるため，個別のデータを指導に直接生かすことができる．しかし，こうしたデータを全体で収集・分析することで，全体像を把握してクラスや学部全体，さらには学習者の言語教育に生かすという視点も重要だと考える．こうした学習者の産出言語を集めてデータベース化したものは学習者コーパス[3)]と呼ばれ，すでに国内外の大学で集積されている．学習者コーパス研究にはデータから学習者の特性（誤用，過剰使用語や過少使用語等を分析）を抽出し，教材開発や効果的な学習方法の開発等につなげる，言語教育への貢献を目的としたものがある．本研究でも最終的には英語教育に寄与するデータベースを構築したいと考えているが，対象としているのが一つの学部の1年生という限定的な学習者の集団であるため，データに大きな偏りが生じる可能性がある．しかし，大型の学習者コーパスの分析によって判明している学習者の特徴を本学部のデータの分析でも同様に示すことができれば，データの汎用性の高さを一定程度示すことが可能だと考える．

本稿の目的は，2019年度の期末テストとして課された前期（辞書等の参考資料の使用を伴う事前準備あり）のエッセイと後期（参考資料の使用なしで即興）のエッセイから学習者特有の過剰使用語を抽出・分析したものを大型の学習者コーパスで得られている知見と比較し，データの汎用性の高さを示すことで将来的な目標である学習者コーパス構築に向けての布石とすることである．

2．先行研究

学習者コーパスを使用した研究には2つの方向性があると言われている（山西，2018）．学習者のデータの観察から共通項を見出して第二言語習得のメカニズムや普遍的な原理を探るものと，学習者特有の誤用や過剰・過少使用のパターン等の言語特徴に着目して言語教育に生かそうというものである．どちらもデータ駆動型の研究となり，統計的な手法を用いて語句の出現頻度という観点から学習者と母語話者の言語使用を比較している．また，前提としてのデータ収集の方法にも，2つの方向性があると考えられる．言語運用に影響する様々な条件（変数）をできるだけ統制するものと，反対に多様で膨大なデータを収集することで各変数の影響を可能な限り抑制するものである．本研究で扱っているデータは，可能な範囲で前者のデータ収集方法をとっており，英語を外国語として学習している日本国内の大学生を対象としている．そのため，書きことば，特にエッセイを前者の方法で収集している大型の英語学習者コーパスのうち，大学生を対象としているICNALE，NICER，及びKUBECを中心に言及する．

神戸大学の石川慎一郎氏が中心となり作文または発話データを収集したICNALE（International Corpus Network of Asian Learners of English）では，アジア圏10か国の大学生及び英語母語話者の計2,800人（そのうち，英語母語話者は200名で同一条件での比較用データ）が対象となっている．参加者は「大学生のアルバイトの是非」と「レストラン全面禁煙の是非」の2つのトピックについて作文もしくは発話をした．データ収集要因等が厳しく統制されている点が特徴で，トピック，プランニングを含んだ執筆時間（20～40分）と作文の長さ（200～300語）が決まっている．その際，ワードプロセッサーとスペルチェッカーを使用し，辞書は使用禁止となっている．2007～2012年度に収集された5,600本からなる作文のコーパスは130万語の規模となっている（石川，2019）．Ishikawa（2013）では，英語母語話者の

エッセイと比較し，日本人英語学習者のエッセイで特徴的に頻度の高い語を対数尤度比[4]（以下，英語の log likelihood ratio を略した LL を使用）という指標を用いて抽出している．トピックに関連する語を除いた上位 10 語は we, agree, people, but, must, n't, so, reason, think 及び example であった．特に思考動詞の think，1 人称複数代名詞である we や people，接続表現である but や so 等を過剰に使用していることを指摘している．習熟度別の分析では，CEFR の B2 +レベルでは but, so, reason, think そして example は過剰使用されなくなること，各過剰使用語における LL の数値は習熟度が上がるほど数値が低くなることが指摘されている．なお，ICNALE の前身である CEEJUS の過剰使用語の分析では，I think と we can の使用頻度の高さを，日本語エッセイにおける「思う」「できる」の頻度の高さを鑑み，母語干渉の例とみなしている（石川 2008, p. 221）．

2018 年に NICER（Nagoya Interlanguage Corpus of English Reborn）1.1 として公開されている名古屋大学の杉浦正利氏が構築した学習コーパスの対象者は，非英語母語話者の大学生と大学院生及び英語母語話者で，2019 年 4 月 4 日時点で計 420 ファイルとなっている．2015 年に公開されている前身の NICE（Nagoya Interlanguage Corpus of English）では 11 トピックあったのだが, NICER では「education」「money」「sports」の 3 つのトピック（ここから 1 つを選び，自由に記述）に絞り，プランニングを含む執筆時間を 1 時間としている．また，ICNALE 同様にワードプロセッサーとスペルチェッカーを使用し，辞書を使用禁止としている．NICER ではさらに，*Criterion*® の自動評価による点数が記載されている点が特徴的である．投野・金子・杉浦・和泉（2013）では，NICER の前身の NICE の分析により，*TOEIC*® 600 点以下の下位群と 650-760 点のグループの中位群で強意副詞のうち，very や so といった booster の過剰使用が見られること，その使用頻度は習熟度が上がるにつれて下がっていくことが指摘されている．また，同じく投野他（2013）で NICE を使用して中級者（*TOEIC*® 500-650）と上級者（*TOEIC*® 700-940）の 2 群を対象に語彙の豊かさを示す指標を語彙の多様性と広範さに分けて調査した．その結果，語彙の広範さが習熟度の違いを表す指標となることを指摘したことは重要だと考える．

関西大学において山西博之氏が中心となって構築した学習者コーパスである KUBEC（Kansai University Bilingual Essay Corpus）は，2012 年以降，大学生 3・4 年生の「英語ライティング 2」の受講生を主な対象に，同じトピックで書いた英語と日本語のエッセイを収集したものである（山西，2018）．現在エッセイ数は 2031 となっている．トピックは NICE の 11 トピックと ICNALE の 2 トピックであり，他の大規模学習者コーパスと比較できるようにしている．また，インストラクションの内容がかなり統制されており，プランニングと執筆時間合わせて約 40 分，英作文の長さは 300 語以上を目標にしている．上記 2 つの大規模学習者コーパスと同様に，ワードプロセッサーとスペルチェッカーを使用し，辞書の使用は禁止している[5]．学習者は英文を書いた後に，同じトピックで英文和訳ではない日本語のエッセイを約 40 分で 800 字以上を目標に書いているのが特徴である．今尾（2019）では KUBEC の日英のエッセイを比較し，学習者の接続表現の過剰使用について検証している．単純な比較は難しいとしながらも，同じ書き手の学習者の場合，一定文数当たりの接続表現の使用頻度は，日本語より英語の方が高いが，英語母語話者よりは低く，多様性も低いとしている．なお，KUBEC は現在（2020 年 11 月）公開データではないため，比較検証はできない．

石川（2012）では，学習者コーパス研究の研究分野としての有望性を示しながらも，データ分析や解釈手法についての制約や課題について言及している．例えば，母語話者の英語を目標言語モデルと設定することの妥当性，学習者コーパスでは測れない項目（小規模コーパスにおける未出現の項目等）の存在，国際英語（International English）・共通英語（Lingua

franca）・世界英語（World Englishes）の観点から分析で得られた差異等を教授項目として安易に採用することへの批判についてとりあげている．さらに，本稿に関わる指摘として，母語話者と学習者の言語使用の差異が何に起因するものか，母語の違い以外にもデータ収集における要因，学習者属性要因，学習者の定義等に関連する多数の変数が存在することに注意を促している．その実例として，International Corpus of Learner English（ICLE）[6] を分析したAltenberg氏　が，Practical Applications in Language Corpora（PALC）'97の学会において学習者による1人称の過剰使用を，書きことばと話しことばの違いを認識できていないことによるとした解釈が，のちにÄdel（2008）の再分析によって時間制限と辞書使用の有無というデータ収集における要因の影響だと判明したことをあげている．時間制限や辞書使用の有無は学習者の言語使用に大きな影響を与えることは想像に難くない．なお，ICLEのデータを使用したGranger（1998）では，早くから学習者が基本名詞，基本動詞，助動詞や量化詞を過剰使用すること，さらにpeopleやthings等の緩和表現を使用することが指摘されており，その後の学習者コーパス研究の基礎的な知見が幾つも報告されている．また，先述の投野他（2013）ではICLEから派生した話しことばを集めたLINSEI（Louvain International Database of Spoken English Interlanguage）の国別の学習者サブコーパスを使用した分析で，学習者は否定的・肯定的な感情を示す語彙の使用頻度が母語話者よりも高く，その中で日本人英語学習者は使用する語彙が偏っている（例えば前者ではangry，後者ではhappy）ことを指摘している．

ここまで，非英語母語話者の大学生を主な対象とし，データ収集要因などを統制した大規模学習者コーパスから得られた知見を概観してきたが，実際にエッセイを集め，同様の傾向が見られるかを検証した松田・石井・岩田・西・濱崎（2020）についても述べておく．松田他（2020）では，本学部の英語学習者のエッセイをICNALE及びNICERの英語母語話者のデータと比較・分析して過剰使用語の実態を探った．その結果，ICNALEやNICERで見られたような学習者の過剰使用語の実態がおおむね見られた．おおむね，というのはICNALEの分析結果ではweの過剰使用が見られたが，松田他（2020）ではweだけではなく，人称代名詞全般の過剰使用が見られたからである．また，先行研究では思考動詞thinkの過剰使用が見られたが，松田他（2020）ではthink（表記形はthought）のみではなく，基本動詞（表記形でcame, happened, said, told, wanted, was, went, were）が過剰使用されていた．さらに身近な存在であるmotherとfriendの過剰使用も特徴としてあげられた．習熟度別の分析では，先行研究と同様に先述の基本動詞とmotherやfriendを含めた過剰使用語のLLは習熟度が高いグループでは低くなることがわかった．ただしweは例外で，習熟度が上がるとLLが高くなることが示された．

松田他（2020）では，ICNALEやNICERの分析で示された学習者の過剰使用語に類似した結果が認められたが，差異も存在しており，それがどのような要因に基づいているのかは不明だった．まず，エッセイのトピックの違いは大きな差異につながる可能性が高い要因と考えられる．次に，データ収集要因の違いも大きな差異につながる可能性が高い．松田他（2020）において分析対象となったエッセイは先行研究とは異なり期末試験として実施されたもので，辞書等の参考資料を使用して事前に準備したものを覚えて30分内に打ち込んだものであった．しかし，先述のÄdel（2008）の例で示されているように，辞書使用等の有無はデータ収集における要因の中で大きな影響を与える変数となる可能性がある．また，事前に準備したものを覚えて打ち込んだ点も，ICNALEやNICERで収集されている即興のエッセイとは大きく異なる．さらに，松田他（2020）ではICNALEとNICERのように比較用に同じトピックで英語母語話者のデータを収集したものが存在しないため，試験的にICNALEとNICERの母語話者のデータを使用し，両者に共通して抽出される

過剰使用語を分析対象としたが，方法の妥当性については検証できていない．本稿でも同じ方法をとるが，将来的には比較用の英語母語話者のデータが必要となる．

2020年度後期に実施された期末テストは，辞書等の参考資料を使用せずに即興で書いたエッセイであり，データ収集要因がよりICNALEやNICERに類似している[7]．そのため，松田他（2020）と同じ方法を用いた場合，2020年度後期のエッセイを分析した場合の方がICNALEやNICERの分析で得られた知見に近い結果がでる可能性が高いと考え，次の2つのリサーチクエスチョンをたてた．

(1) ICNALEとNICERのデータ収集要因に類似するほど，これらのコーパスの分析で見られたような学習者の過剰使用語の実態に近づくか．
(2) それは学習者の習熟度別の分析についてもいえるか．

仮説は以下の通りである．

(1) ICNALEとNICERのデータ収集要因に類似するほど，つまり辞書等の参考資料を使用せずに即興で書くエッセイの方が，思考動詞のthink，1人称複数代名詞であるweやpeople，接続表現であるbutやso，そしてveryやsoといった強意副詞を過剰使用する．
(2) 各過剰使用語におけるLLは習熟度が上がるほど数値が低くなる．

3．方　法

期末テストとして課された前期（辞書等の参考資料の使用を伴う事前準備あり）のエッセイと後期（参考資料の使用なしで即興）のエッセイを収集し，松田他（2020）と同じようにICNALE及びNICERの英語母語話者のデータを利用して両方に共通してみられる過剰使用語を各々抽出し，比較した．

3.1　参加者

筆者らが担当する1年生の「英語演習」クラスの学生に，調査の内容と参加の如何が成績に反映されることはないことを説明し，任意で同意書に記名してもらった．本研究では，同意書に記名した学生のうち，与えられたトピックと無関係と思われる内容を記述した人数分のエッセイを除外し，前期は156人分のうちの153人分，後期は176人分のうちの173人分のエッセイを使用した．表1は参加者の習熟度を示すものである．対象となる「英語演習」は習熟度別のクラス構成で，学部独自のレベル分けの基準に沿って大きくI～IIIに分類される．本学部ではプレイスメントテストの点数を使用して習熟度別のクラス分けを行うが，1年生の場合は通常習熟度が低い方から7クラスがI（本研究への参加は前期1クラスで後期2クラス），6クラスがII（本研究への参加は4クラス），5クラスがIII（本研究への参加は3クラス）となっている．学生はプレイスメントテストとして4月に*TOEIC Bridge*®を受験し，12月に*TOEIC Bridge*®もしくは*TOEIC*®を受験する

表1　参加者の習熟度のデータ

習熟度レベル	2019年4月実施 TOEIC Bridge IP【TOEIC換算点】	2019年12月実施 TOEIC IP（TOEIC BridgeはTOEIC換算点を使用）	人数（クラス数） 事前準備あり	即興
I & II	138.58 (*SD* = 4.64) 【345-395】	407.96 (*SD* = 76.94)	96 (5)	120 (6)
III	156.78 (*SD* = 6.16) 【470-570】	494.83 (*SD* = 75.87)	57 (3)	53 (3)

ことになっている．4月の *TOEIC Bridge*® では，レベルIとIIのクラスに属する5クラスの平均点は138.58（*SD* = 4.64），レベルIIIのクラスに属する3クラスの平均点は156.78（*SD* = 6.16）[8] であった．なお，12月は，参加者が *TOEIC Bridge*® もしくは *TOEIC*® を受験しているため，*TOEIC Bridge*® を受けている場合は *TOEIC*® 換算点を使用している．参加者の多くがCEFRのA2のレベルに属している．また，個人の学習の様子は把握できないが，1年生の必修科目は「英語演習」（週2回）と「Oral English」（週1回）となっており，英語の学習環境は類似している可能性が高い．

3.2　手続き

前期（2019年8月1日と8月6日に実施）と後期（2020年1月28日と1月30日に実施）の期末テストのうち，同意書に記名した学生から提出されたエッセイを *Criterion*® からダウンロードしてテキストファイルとした．分析前の下処理として第一筆者とカナダ出身の英語母語話者（英語講師及び翻訳者として約10年日本に在住．人類学で修士号取得．）がWordのスペルチェック機能等を使用しながらエッセイの綴りのミス（前期は483語で全体の1.19%，後期は966語で全体の2.74%）を修正した．両者が綴りのミスとして抽出した語の一致率は前期のエッセイでは98.96%，後期は89.03%であり，信頼性は十分高いと考える．なお，単語の間に不要なスペース[9] やパンクチュエーションを入れてしまっている場合やその反対で必要なのに入れてない場合は，単語の認識に関わってくるため，ミスとしてカウントしている．

前期のエッセイのトピックについてはDescriptiveのモードでは5th Grade（米国の小学生高学年レベル）の「Feeling Happy」，Narrativeのモードでは6th Grade（米国の中学生レベル）の「Alien Encounter」，「Desert Island」，「Lesson Learned」であった．後期のエッセイのトピックは期末テストの実施方法の関係で公開できないが，いずれもPersuasiveのモード内の4トピック（便宜的に①～④と表記）から選ばれている．9つのクラスの内訳だが，トピック①～③は各2クラス，トピック④は3クラスである．綴りのミス以外は原文のまま対照コーパスとして使用している．

松田他（2020）と同じく，分析にはコンコーダンス・ソフトウェアとして早稲田大学のLaurence Anthony氏が開発したAntConc3.5.8（Windows）2019を使用し，参照コーパスとしてICNALE及びNICERにおける英語母語話者のデータを使用した．AntConcで過剰使用語を抽出するには，Keyword Listを作成して特徴度（keyness）を測るが，その際，対象コーパスと参照コーパスに出現する語の頻度差の著しさをLL（対数尤度比）という指標で表す．ICNALEとNICERを使用したときでは，抽出される過剰使用語が異なるため，LLが100を超えるものを選び，トピックと課題文に関連する語を分析から除外したのち，共通して見られた過剰使用語を抽出した．

4．結　果

表2はデータの概要を示している．前期のデータの総語数（token）は40,631語，異なり語数（type）は3,023語だった．また，語彙の多様性を示す指標であるType-Token Ratio（TTR）　は7.44%でGuiraud Index（GI）[10] は15.00であった．後期のデータの総語数（token）は35,303語，異なり語数（type）1,747語，TTRは4.95%であり，GIは9.30であった．参加人数が異なるため，一人当たりの産出量

表2　データの概要

エッセイ	エッセイ数	総語数	異なり語数	TTR (%)	GI	平均総語数
事前準備あり	153	40631	3023	7.44	15.00	265.56
即興	173	35303	1747	4.95	9.30	204.06

を示す平均総語数で比較すると，前期の 265.56 語に対して後期は 204.06 語であり，辞書等の参考資料なしで即興のエッセイを書く場合，一人当たりの産出量はかなり減ることが分かる．さらに TTR や GI の数値が示すように，語彙の多様性は明らかに低くなっている．

表 3 は ICNALE と NICER を使用して抽出した過剰使用語（表記形）を示したものである．後期のエッセイで抽出された過剰使用語は前期のエッセイよりも数が少ない．これは後期のエッセイが短く，語彙の多様性が低いことに関係していると思われる．また，ICNALE のデータを使用した先行研究では学習者の 1 人称複数代名詞 we や people の過剰使用が指摘されていた．前期のエッセイでは we を含む人称代名詞（I, my, me, he, she, her）が過剰使用されているが，後期のエッセイでは 3 人称が過剰使用語としては抽出されなくなる．また，特徴的であった mother と friend については，後期のエッセイは friend のみが抽出された．ICNALE を使用した研究で指摘された思考動詞 think は両方に残っている．前期のエッセイでは，思考動詞（表記形では thought）だけではなく，基本動詞（表記形では came, said, told, wanted, was, went, were）を過剰使用していたが，後期のエッセイでは think と want のみでこれらの結果は先行研究とある程度一致する内容である．また，助動詞 can があるが，先述したように，ICNALE の前身である CEEJUS の分析から，I think と we can の使用頻度の高さは母語干渉による可能性が高く（石川，2008），can の存在はそれを反映したものと考える．また，後期のエッセイでは ICNALE の分析で指摘されていた論理関係や接続関係を明示する語である so が過剰使用されている．so のコンコーダンス検索結果を図 1 に示した．明らかに接続詞の so の数が多いのが見て取れる．実際に数えてみると，接続詞の so の頻度は 271 回，副

表 3　ICNALE と NICER を使用して抽出した過剰使用語（表記形）

ICNALEとの比較						NICERとの比較					
事前準備あり			即興			事前準備あり			即興		
過剰使用語	頻度	LL	過剰使用語	頻度	LL	過剰使用語	頻度	LL	過剰使用語	頻度	LL
i	2619	1491.01	i	1554	482.50	i	1594	1135.45	i	1554	1429.80
was	973	1344.26	you	595	479.90	was	588	1030.27	think	464	732.10
said	324	633.80	we	437	316.00	said	195	497.20	you	595	614.53
my	663	436.41	think	464	219.60	my	469	452.26	we	437	360.42
we	535	390.42	life	213	203.37	went	98	254.37	want	221	319.35
went	160	317.23	happy	119	192.24	mother	76	220.64	can	465	288.53
me	357	278.29	things	190	181.87	me	219	215.46	so	337	277.70
were	269	276.47	can	465	174.67	we	264	190.00	things	190	223.98
day	181	207.19	want	221	170.18	day	118	177.65	happy	119	218.13
mother	93	200.32	friend	78	159.21	he	99	158.35	friend	78	184.23
came	101	185.75	first	143	116.11	were	138	157.04	my	341	165.29
she	112	171.60	so	337	111.05	she	76	154.90	second	113	152.03
told	92	150.38	second	113	103.71	told	67	148.37	life	213	132.94
thought	107	146.78	my	341	103.53	her	65	142.76	first	143	108.05
he	128	141.40				came	53	122.48			
her	85	135.72				thought	65	113.43			
friend	70	126.70				wanted	52	112.90			
wanted	76	123.90				friend	45	106.92			

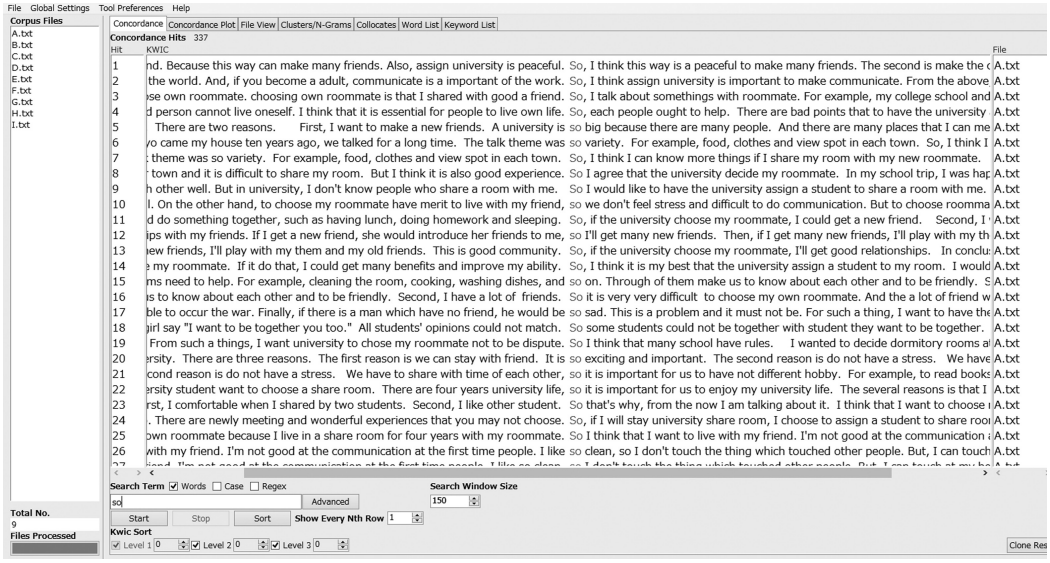

図１　so のコンコーダンス検索結果

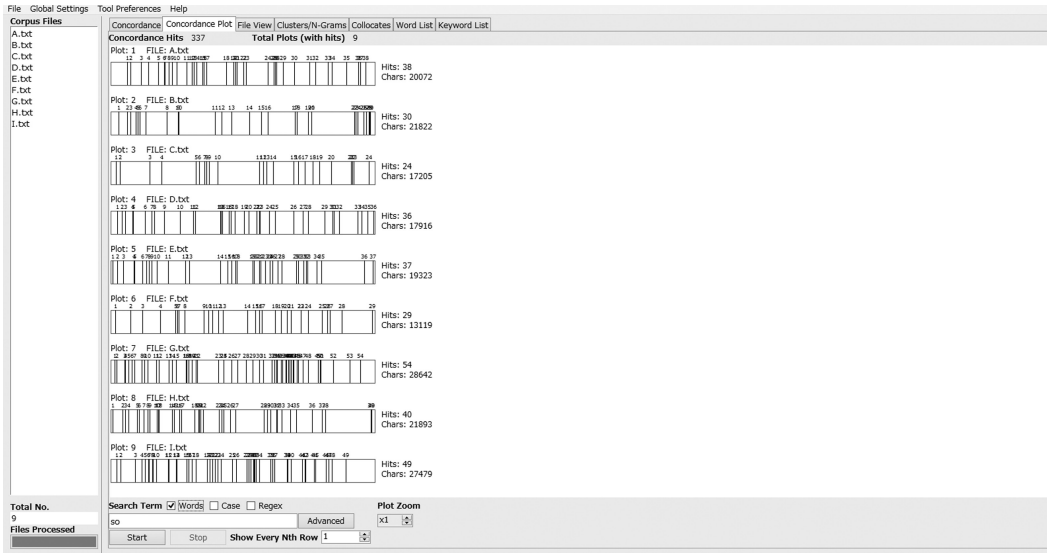

図２　so のコンコーダンスプロット

詞の so の頻度は 66 回の内訳になっている．クラス別（A~I と表記）のファイルでコンコーダンスプロットを見ていくと，接続詞 so（縦線）はトピックに関係なく過剰に使用されていることがわかる（図 2）．さらに後期（辞書使用なし，事前準備なし）のエッセイに特徴的な語をみていくと，first, second といった順序を示す frame markers（Hyland, 2005）が出てきている．

表 4 は習熟度別のデータの概要を示している．松田他（2020）同様に，レベル I と II を 1 つのグループ（レベル I & II），レベル III をもう 1 つのグループにした．まず，前期のエッセイと後期のエッセイ両方において，レベル I & II と比較するために平均総語数を見ると，レベ

表 4　習熟度別のデータの概要

習熟度	エッセイ	エッセイ数	総語数	異なり語数	TTR (%)	GI	平均総語数	Word Level
I&II	事前準備あり	96	23908	2252	9.42	14.56	249.04	—
	即興	120	22792	1390	6.10	9.21	189.93	1.35
III	事前準備あり	57	16723	1872	11.19	14.48	293.39	—
	即興	53	12511	1037	8.29	9.27	236.06	1.41

ル III は産出量が多いことがわかる．また，語彙の多様性を示す TTR の数値を見ると習熟度が上がるほど高くなっているが，同様の指標である GI の数値を見る限り差がないことがわかる．制限時間の短さと語数制限を考えると，語の難易度を考慮していないこれらの数値は表面的な指標ではないかという疑問がわく．また，先述したように先行研究の投野他（2013）で語彙の多様性と語彙の広範さでは，語彙の広範さが習熟度の違いを示すことが指摘されているため，各習熟度別に染谷（2009）の Word Level Checker を使用して後期のエッセイの語の Word Level も併せて調べている．習熟度が上がれば産出語彙の難易度が少し上がることが分かる．

表 5 は ICNALE の英語母語話者のデータと比較したもので，表 6 は NICER の英語母語話者のデータと比較したものである．抽出された過剰使用語の LL の数値を習熟度別にみていくと，習熟度が上がると過剰使用語の LL が下がることが分かる．しかし，we（前期はそれに共起する were も）に関しては例外的に LL の数値が高くなっている．つまり習熟度が高くなると，1 人称単数代名詞よりも 1 人称複数代名詞に重きが置かれることがわかる．また，後期のエッセイでは，my は特徴語として抽出できず，happy と things の LL が例外的に上がっている．NICER の英語母語話者のデータを使用した場合もほぼ同じことが言える．

表 5　ICNALE を使用して抽出した習熟度別の過剰使用語（表記形）

習熟度レベル I&II						レベルIII					
事前準備あり			即興			事前準備あり			即興		
過剰使用語	頻度	LL	過剰使用語	頻度	LL	過剰使用語	頻度	LL	過剰使用語	頻度	LL
i	1594	1135.45	you	422	432.83	was	385	736.21	we	187	220.04
was	588	1030.27	i	1059	419.48	i	1025	708.66	happy	64	175.70
said	195	497.20	friend	69	182.47	said	129	377.26	things	92	152.85
my	469	452.26	we	250	180.83	we	271	322.52	i	495	150.44
went	98	254.37	want	169	177.90	were	131	206.90	you	173	148.50
mother	76	220.64	can	327	163.50	went	62	183.68	life	83	116.91
me	219	215.46	think	294	154.38	came	48	135.37	think	170	110.34
we	264	190.00	my	270	138.34	me	138	134.66	so	127	60.88
day	118	177.65	life	130	138.12	my	194	103.05	can	138	44.85
he	99	158.35	first	99	99.88	day	63	88.81	first	44	40.97
were	138	157.04	second	81	94.69	thought	42	79.56	second	32	31.68
she	76	154.90	happy	55	92.30	she	36	71.27	want	52	30.63
told	67	148.37	things	98	88.84	friend	25	64.56	friend	9	19.33
her	65	142.76	so	210	72.84	wanted	24	51.46			
came	53	122.48				mother	17	51.12			
thought	65	113.43				told	25	50.42			
wanted	52	112.90				her	20	36.11			
friend	45	106.92				he	29	26.05			

表６　NICER を使用して抽出した習熟度別の過剰使用語（表記形）

習熟度レベル I&II						レベルIII					
事前準備あり			即興			事前準備あり			即興		
過剰使用語	頻度	LL	過剰使用語	頻度	LL	過剰使用語	頻度	LL	過剰使用語	頻度	LL
i	1594	2292.24	i	1059	1203.91	i	1025	1574.62	i	495	574.75
was	588	795.95	think	294	562.08	was	385	561.94	think	170	409.07
my	469	551.11	you	422	553.03	we	271	362.55	we	187	252.86
said	195	394.85	want	169	316.50	said	129	295.78	you	173	214.55
me	219	266.94	can	327	261.83	were	131	185.37	happy	64	200.12
went	98	230.47	we	250	215.61	me	138	174.41	things	92	186.20
we	264	225.76	friend	69	207.50	went	62	166.75	so	127	153.60
day	118	171.75	my	270	199.96	my	194	152.55	can	138	90.58
were	138	137.22	so	210	197.18	came	48	102.32	want	52	89.48
friend	45	131.96	second	81	138.01	friend	25	87.58	life	83	78.77
told	67	123.53	things	98	119.51	day	63	87.22	second	32	55.63
he	99	121.65	happy	55	114.59	thought	42	71.45	first	44	39.00
mother	76	110.37	first	99	94.09	told	25	38.83	friend	9	35.94
thought	65	101.59	life	130	88.30	wanted	24	34.91			
came	53	89.09				she	36	28.24			
wanted	52	84.36									
she	76	79.87									
her	65	61.72									

５．考　察

大型の学習者コーパスの分析で得られている知見と比較するため，英語学習者のエッセイを収集・分析して過剰使用語の実態を探り，その結果をまとめてきた．先述した仮説に沿って結果を振り返る．まず，「(1) ICNALE と NICER のデータ収集要因に類似するほど，つまり辞書等の参考資料を使用せずに即興で書くエッセイの方が，思考動詞の think，１人称複数代名詞である we や people，接続表現である but や so，そして very や so といった強意副詞を過剰使用する.」という仮説を立てた．他の大型学習者コーパスと収集要因が類似している後期のエッセイでは，think や we の過剰使用に加えて，前期のエッセイでは抽出されていない接続表現である so の過剰使用が見られた．また，他にも緩和表現とされる things, 肯定的な感情を示す語彙の happy が使用されており，ICLE や LINSEI の研究で得られている結果にも大まかに沿っている．母語干渉により，助動詞 can が過剰使用されていることも，先行研究に沿った結果である．トピックの違いという影響力の大きい変数が存在しながらも，後期のエッセイに大型の学習者コーパスの分析で見られたのと同様の過剰使用語の実態が認められることは，データの収集要因が与える影響の大きさを物語ると考えられる．また，前期のように辞書等の参考資料を使用すれば，語彙の多様性が増すが，後期の即興のエッセイでの数値の落ち込みを見ると，授業や辞書等の参考資料の使用を通して得た語彙知識は必ずしも定着していない，もしくは少なくともテストという場面で使用できていない可能性がある．学習者は時間も参考資料もない状態であれば，産出可能な表現を繰り返し使用していると考えられる．

次に「(2) 各過剰使用語における LL は習熟度が上がるほど数値が低くなる.」という仮説に関しても，おおむね見られたということができる．先行研究と同様に，例外はありながらも，習熟度が高いグループでは過剰使用語の LL が低くなることがわかった．例外としてはまず we があげられ，習熟度があがると１人

称複数代名詞に重きが置かれることが示された．また，同じく例外の things と happy だが，英語学習者全般に見られる特徴が，習熟度の高いグループにおいて顕著になるのは興味深い．特に happy は話しことばのコーパスである LINSEI の日本人学習者のサブコーパスにおいて肯定的な感情を示す語彙では使用頻度が最も高い語であり，スピーキングとライティングというアウトプット型のスキルで使用する語彙の結びつきを伺わせるもので面白い．その他，習熟度が上がると語彙の多様性ではなく，語彙の広範さが上がることもデータによって示されたが，分類したグループでは大きな差とは言えなかった．*TOEIC*® のスコアで習熟度を分けた場合，リスニングやリーディングスキルに関わる受容語彙（見て意味が分かる語彙）の知識に関しては差があっても，スピーキングやライティングに関わる発表語彙（実際に使用できる語彙）の知識に大きな差がない可能性が高いと考える．*TOEIC*® スコアから見ると，本学部の学習者の大半が CEFR の A2 に属するため，そもそも語彙知識にあまり差がない可能性も高い．さらに言えば，エッセイの内容や構成の違いに習熟度の差が表れている可能性がある．

今回，データ収集要因が類似しているほど英語学習者全般に見られる特徴的な過剰使用語をより多く抽出できており，習熟度別の分析でも英語学習者全般に見られる特徴的な過剰使用語の在り方が認められた．そのため，今回収集したデータの汎用性の高さは，ある程度示すことができたと考える．そしてデータ収集要因を統制することは，汎用性が高いデータから成る学習者コーパスの構築を目標にする場合，非常に重要であることが分かる．データ収集要因には辞書等の参考資料の使用の有無が含まれているため，近年の英語教育に大きなパラダイムシフトを引き起こしている，非常に精度の高い機械翻訳 [11] について最後に触れたい．

前期のエッセイでは機械翻訳を使用しているケースが散見された．学習という観点から言えば，機械翻訳に頼りすぎるのは問題であり，これは語彙知識の定着率の低さの一因となる可能性もある．しかし，学習者に機械翻訳の使用を禁止するのはすでに現実的ではない状況にある．特に 2020 年度は COVID-19 の影響により，本学部の「英語演習」の授業のエッセイの課題は自宅でオンライン上の参考資料を使用しながら作成することができた．教員が個々の学習者をモニターできない状態となったため，機械翻訳の使用率は高くなっていると予想できる．これは予見できない事態であったが，エッセイ・ライティングの授業で何を学ぶことを重視するかを再考する機会と捉えることもできる．*TOEFL*® 等のテスト対策に有効なエッセイの書き方の型を覚え，語彙知識を増やすことを重視するという考え方がある一方，機械翻訳を含む様々なツールを駆使して学習者が表現したいことを伝えることをより重視するという考え方もある．後者の考え方をするのであれば，英語教員は機械翻訳等の使用方法を教えることも求められるであろうし，どのように使用すれば語彙知識が定着するのか等，教授法についての検討も求められるだろう．いずれにしても，非常に精度の高い機械翻訳の誕生により，ライティングの授業については学ぶ側も教える側も発想の転換を迫られている．この状況下にあってデータ収集要因を考える場合，汎用性が高いデータというのは学習者が機械翻訳等の参考資料を使用しないで産出したデータだけであると考えて良いものか疑問を覚える．

6．結　論

2019 年度の前期と後期に期末テストとして課されたエッセイと，ICNALE 及び NICER の英語母語話者のデータを比較して過剰使用語を抽出し，データ収集要因の違いがいかに大きな影響を与えるかを見てきた．データ収集要因が大規模な学習コーパスに類似するほど英語学習者全般に見られる特徴的な過剰使用語をより多く抽出できているため，本研究で収集したデータの汎用性の高さをある程度示すことができたと考える．しかし，本研究にはいくつかのリミテーションが存在している．まず，過剰使用語の抽出に ICNALE 及び NICER の英語母語話者

のデータを使用していることである．本来，同じトピックで英語母語話者のデータを収集して分析する必要があるため，今後の課題としたい．また，後期のエッセイのトピックについて，期末テストの実施方法の関係で公開できていない点もリミテーションとしてあげられる．さらに，本研究は語の分析に終始しているため，学習者の言語使用を多角的に捉えられていない可能性が高い．今後はコロケーションの分析等，より大きな言語単位での分析を加える必要があると考える．最後に，本研究では各エッセイの内容や構成の違いを見ていないため，習熟度の差を示す重要な指標を見逃している可能性がある．今後は *Criterion*® の自動採点機能の利用も含めてこの件を精査する必要がある．

学習者コーパスを構築するのであれば，それをどのような形で教育に生かすのかについて検討していく必要があるだろう．今回抽出した特徴的な過剰使用語を意識化させることで学習者の語彙使用に変化が生じるか否かを調査すること等，検討課題は多い．学習者コーパスを誰でもアクセスできる教材として学習者に提供し，分析方法を教授すれば，様々な気づきが生まれる可能性もある．本研究ではデータの収集要因が与える影響について調査したが，今後は機械翻訳の存在を念頭においた収集要因を加えていく必要があるのかもしれない．いずれにしても，より良い指導や研究に役立つような学習者コーパスの構築に向けて知見を集積していきたいと考えている．

註

1）*Criterion*® はライティング専用の LMS（Learning Management System）である．アメリカの ETS（Educational Testing Service）が開発した *TOEFL iBT*® の e-rater® という自動採点エンジンによる採点が行われる．
2）参考資料には機械翻訳等，オンラインで利用できるものを含む場合がある．
3）コーパスとは，「電子化された大量の言語データベースのこと」（石川，2008，p. 4）である．
4）多数尤度比は正規分布を前提としないこと，まれな現象を過大評価しないこと，サンプルの分量の差を考慮しなくても良いという点で偏りの大きい語の頻度データには適した指標とされている（Leech at al., 2001）．
5）ただし 2012 年を除く．また，「改訂版を提出する際は辞書使用を許可した.」とある．
6）ベルギーのルーヴァンカトリック大学の Sylviane Granger 氏が中心となって 1990 年に着想，構築してきた，のちの英語学習者コーパス研究の礎となる世界最大の英語学習者コーパス．
7）ただし，執筆時間については ICNALE に類似している．
8）習熟度が高い参加者が *TOEIC Bridge*® を受験した場合，天井効果により実際に受験した場合より低い点数が出る可能性があることは考慮に入れる必要がある．
9）ただし，スペースを半角 1 文字のところを 2 文字空ける等，多くいれてしまっている場合は単語の認識に関わらないのでカウントしない．
10）TTR はサンプルサイズ依存性が高い（語数が多くなると値が低くなる）ことが問題点として指摘されているため（石川 2012, p. 143），GI も示しておく．
11）例えば DeepL などがあげられる．https://www.deepl.com/ja/translator

参考文献

Ädel, A.（2008）. Involvement features in writing: Do time and interaction trump register awareness? In G. Gilquin, S. Papp, & M. Díez-Bedmar（Eds.）, *Linking up contrastive and learner corpus research*（pp. 35–53）. Amsterdam, The Netherlands: Rodopi.

Granger, S.（Ed.）,（1998）. *Learner English on computer*, Longman, London.

Hyland, K.（2005）. *Metadiscourse: Exploring interaction in writing.* London: Bloomsbury

Publishing.
今尾康裕（2019）.「日本の大学生英語学習者によるエッセイでの接続表現を探る：日本語エッセイ・英語母語話者によるエッセイと比較して」『大阪大学大学院言語文化共同研究プロジェクト 2018』, 5–23.
石川慎一郎（2008）.『英語コーパスと言語教育―データとしてのテクスト』大修館書店, 東京.
石川慎一郎（2012）.『ベーシックコーパス言語学』ひつじ書房, 東京 .
Ishikawa, S.（2013）. The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. *Learner corpus studies in Asia and the world, 1*, 91–118.
石川慎一郎（2019）.「英語学習者コーパス研究の現状と課題」『電子情報通信学会 基礎・境界ソサイエティ *Fundamentals Review*』*12*, 280–289.
Leech, G., Rayson, P., & Wilson, A.（2001）. *Word frequencies in written and spoken English: Based on the British National Corpus*. Harlow, UK: Pearson Education Limited.
松田紀子・石井隆之・岩田雅彦・西美都子・濱崎佳子（2020）.「英語学習者のエッセイに見られる過剰使用語―学習者コーパスの構築を視野に入れて」『近畿大学総合社会学部紀要』*8*(2), 19–27.
染谷泰正（2009）.「オンライン版「英文語彙難易度解析プログラム」（Word Level Checker）の概要とその応用可能性について」『青山学院大学文学部紀要』*51*, 97–120. Retrieved from http://someya-net.com/wlc/readability.pdf
投野由紀夫・金子朝子・杉浦正利・和泉絵美（2013）.『英語学習者コーパス活用ハンドブック』大修館書店, 東京.
山西博之（編）(2018).『大規模バイリンガルエッセイコーパスの構築とデータ分析のための各種システムの開発』渓水社, 広島.

謝辞

本稿は学生の皆さんのご協力によって成り立っています．心より感謝申し上げます．また，多数の有益なコメントをしてくださった査読者の方にもお礼申し上げます．