広島大学大学院教育学研究科紀要　第二部　第64号　2015　139－146

第二言語習得研究のための新たな学習者コーパス構築

阪　上　辰　也
（2015年10月５日受理）

Compilation of a New Learner Corpus for Second Language Acquisition Research

Tatsuya Sakaue

Abstract: The aim of this paper is to point out the issues that make the existing learner corpora difficult to use for second language acquisition and English education research, and to report on the characteristics of a new learner corpus, the Hiroshima Interlanguage Corpus (HIC), which has been compiled for conducting language acquisition research. Experimental methodologies, such as a grammatical judgment test or a picture-description task, are encouraged for use in second language acquisition research, and some kinds of language tests or questionnaires are used for English education research. In addition to these methods, learner corpus research should be adapted in order to do foreign/second language acquisition research. Although some learner corpora, such as the International Corpus of Learner English (ICLE), are available for language research, these have some difficulties as learner variables, such as particular learners' proficiencies, are not always well described. Further, it is difficult to analyze the process of language production. This paper strives to explain the characteristics needed to compile a new, useful learner corpus for second language acquisition research.

Key words: learner corpus, second language research
キーワード：学習者コーパス，第二言語習得研究

１．はじめに

本稿の目的は，既存のコーパスを第二言語習得研究および英語教育研究に利用しようとする際の課題について検討した上で，今回新たに構築した学習者コーパスの特徴について報告することである。

第二言語習得研究においては，文法的な文であるかどうかを判断するテスト（grammatical judgement test）や，絵を見てその様子を描写する課題（picture-description task）などを用いて，学習者の習得状況を観察することが一般的である。また，英語教育研究においても，何らかのテストやアンケートを用いることで，達成度や心的態度を測定・評価することが多く行われている。特に前者では，実験計画に基づいた要因統制を重視し，なおかつ，客観性を確保するために，量的研究が実施されることが多い。いずれの研究分野においても，何らかの手法を用いて，学習者の言語運用能力や言語行動を測定･評価しようと試みているが，可能な限り，学習者の言語行動そのものを観察し，その傾向や特徴を把握する手法も必要だと考えられる。そのひとつの手段として，学習者コーパス（learner corpus）の利用が挙げられる。しかしながら，コーパスの構築に際しては条件統制などが不十分なため，そ

本論文は，課程博士論文を構成する論文の一部として，以下の審査委員により審査を受けた。
審査委員：深澤清治（主任指導教員），中尾佳行，築道和明，松見法男

のまま第二言語習得研究に利用することが難しい面がある。そこで，本稿では，これまでの学習者コーパスの持つ特性や課題を概観するとともに，言語習得研究・外国語教育研究のための学習者コーパスに必要な条件について検討した上で，新たなコーパスを構築し，その特徴を説明する。

2．学習者コーパスとその研究利用

2.1 学習者コーパスの定義

学習者コーパスの定義として，Granger (2002) は，次のように述べている。

Computer learner corpora are electronic collections of authentic FL/SL textual data according to explicit design for a particular SLA/FLT purpose. They are encoded in a standardised and homogeneous way and documented as to their origin and provenance.
(Granger, 2002: 7)

つまり，Granger (2002) によれば，学習者コーパスとして満たすべき条件として，ある特定の目的を定めた上で電子的なデータとして構築すること，そして，そのコーパスデータは，統一的な方法・形式で記録され，そのデータがどのような条件等で得られたものか，その出所が明らかであることという２点が挙げられている。

次節にて，実際に研究利用されている学習者コーパスについて概観するとともに，それらが抱える課題を挙げる。

2.2 既存の学習者コーパスと研究利用上の課題

既存の学習者コーパスとして，２つの先駆的なものが存在する。１つは，CHILDES (CHIld Language Data Exchange System) であり，もう一つは，ICLE（International Corpus of Learner English）である。CHILDES に関しては，これまでに，３千件を超える研究事例が公開されており（研究事例の一覧は http://talkbank.org/info/usage/childesbib.pdf を参照），ICLE についても数百件の研究事例が報告されている。

CHILDES は，こどもの口語英語を収集したデータベースであり，Brian MacWhinney と Catherine Snown により1980年代から開始された研究プロジェクトである。元々は，第一言語獲得研究のために構築が始まったものであるが，近年では，第二言語習得用のデータも収集されており，http://talkbank.org/data/SLABank/ で公開されている。例えば，大関（2008）により，子どもによる関係節の習得過程を分析するために，日本語版の CHILDES を利用した研究が行われている（なお，CHILDES に収録されている子どもの発話データは，http://childes.talkbank.org/data/EastAsian/Japanese/ から入手可能）。

CHILDES が構築され始めた当時は，言語のデータベースが「コーパス」として呼ばれることはなかったが，データは電子的なものであり，言語習得研究という目的で構築されているという点では，本章冒頭で引用した定義に適合するものであり，事実上のコーパスであると言える。さらに，母語話者のデータだけでなく，第二言学習者のデータも収集されて公開されており，CHILDES で公開されているデータは先駆的な学習者コーパスであると言える。

次に，大規模な国際的学習者コーパスとして，International Corpus of Learner English（ICLE）が構築され有償で利用できるようになっている。この ICLE は，初版が2002年に公開され，2009年に，新たなデータを加えた第２版が公開されている。大学３・４年生の書き言葉データを収録しており，総語数は約370万語となっている。ヨーロッパ圏の学習者を中心としているが，日本人学習者や中国人学習者のデータもサブコーパスとして含まれており，様々な母語を話す学習者のコーパスデータを利用できる点が最大の特徴である。また，CHILDES も同様だが，ICLE においても，複雑な言語処理を行うことなく，言語分析を支援するためのソフトウェアが同梱されており，研究利用の促進につながっている。さらに，ICLE の場合は，国際的な学習者コーパスであることから，Contrastive Interlanguage Analysis（CIA）という母語話者と学習者の比較，そして，学習者と学習者の比較を可能にした点も特筆すべき点である。

なお，ICLE のマニュアルには，"Two fuzzy variables" として挙げられているものがあり，ESL か EFL かという学習環境（learning context）の違いと，習熟度（proficiency）がある。習熟度については，Common European Framework of Reference for Languages (CEFR) の指標に基づき，一部のデータについては評価が行われており，概ね上級レベル（C1あるいは C2レベル）にあることが確認されている。つまり，ICLE では，一部のデータが上級レベルとみなせるため，全体的にも中上級レベルの学習者による作文データであるとみなしているわけであるが，習熟度に基づくデータの分類を行うには評価の規模が不十分であり，習得研究に利用することを想定するのであれば，全データに対して評価が付与されるべきであろう。

前述の通り，CHILDES や ICLE にも日本人英語学習者のデータは含まれているのだが，近年になり，日

本人学習者を分析対象の中心に据えた学習者コーパスも構築され，一般に公開されるようになった。ここでは，主要な日本人英語学習者コーパスとして公開され利用可能となっている，The National Institute of Information and Communications Technology-Japanese Learner English（NICT-JLE）Corpus，JEFLL (Japanese EFL Learner) Corpus，Nagoya Interlanguage Corpus of English（NICE），The International Corpus Network of Asian Learners of English（ICNALE）の４種を取り上げ，それぞれの学習者コーパスが持つ特徴と，第二言語習得研究向けにこれらの学習者コーパスを利用する上での課題を述べる。

まずは，４種の学習者コーパスの特徴を表１に示す。

表１　主要な日本人英語学習者コーパスの特徴一覧

コーパス	種類	特徴
NICT-JLE	S	評価・エラータグを付与
JEFLL	W	中高生が対象
NICE	W	母語話者による添削文を付与
ICNALE	S+W	アジア圏の大学生も対象

（S は話し言葉を収集して構築したデータ，W は書き言葉を収集して構築したデータであることを示す。）

それぞれの学習者コーパスには表１に示すような特徴があり，日本人英語学習者に特有の言語使用の傾向を観察することは可能となっているものの，第二言語習得研究用のデータとして利用する際に複数の課題を抱えている。具体的に，習熟度の分類と，タスク関連の条件統制という２点に焦点を当て，それぞれの問題点を指摘する。

まず，既存の学習者コーパスでは，習熟度による分類が困難なものが多い。NICT-JLE は，独自の９段階評価を入れているが，他の指標との互換性が低い。また，ICLE では一部ながら CEFR が利用され，今後も採用される可能性があるが，ランクによる区別は，細かい分類による比較が困難となるおそれがある。可能ならば，種類は異なるが，それぞれの指標に一定の互換性があり，一般的に利用されている資格試験の得点を記録しておくことが望ましい。既存の学習者コーパスの中でも比較的分類を行いやすいのは，ここ数年で構築・公開された NICE と ICNALE であり，いずれも，TOEIC® 等の英語資格試験の結果を含めている。しかしながら，ライティングやスピーキングの能力評価は難しい面が多く，習熟度による分類は困難な状況が続いている。

次に，タスク関連の統制が不十分という問題点がある。学習者コーパスとして最大規模となる ICLE では，宿題として課された作文や，テストとして実施した作文など，産出した際の条件統制が十分になされているとはいえないデータが混在している。第二言語習得研究での利用を考慮すると，学習者の能力のみで産出させる必要があり，辞書などの使用を制限するべきであるが，ICLE のデータでは，産出環境の混在により，条件の統制がとれなくなり，研究利用が難しくなる。

主要な日本人学習者コーパスについては，時間制限を設けるなど，ある一定の管理がなされた環境下で産出されたものとされている。例えば，産出させる際にどのような指示がなされたか，あるいは，辞書使用を禁じた場合に監督者による監視があったかどうかなど，詳しい産出状況・環境に関するより詳細な記述が必要とされるが，十分に公開されていない面がある。この中でも，NICE については，産出時の監督者の有無や指示の内容について情報が提供されている。

習熟度に関連してさらに考慮すべき問題は，産出行為そのものが日本人英語学習者にとっては負荷の高いものとなっている点である。つまり，一定の習熟度に達しなければ，英語で内容的にまとまりのあるものを書いたり話したりすることは困難で，コーパス構築に必要なデータ量と，分析対象となる有効な産出データが得られにくいという現状がある。例えば，中高生を対象にした JEFLL では，そもそも，既知の語彙や構文の種類が少ないために，表現できる分量が限られ，各学習者から得られるデータ量も十分なものとはならない。何を分析するかにもよるが，必要十分なデータをどの習熟度にある学習者から得るべきかについては，継続して検討すべき課題である。

総じて，CHILDES や ICLE をはじめとして，学習者コーパスが構築・整備されてきたことで，母語話者と学習者，学習者と別の学習者との比較が可能となりつつあるのだが，各コーパスの細かい部分において，習熟度の区分けが曖昧であったり，データ収集時の条件統制が不十分であったりする部分もあり，第二言語習得研究用に利用可能なコーパスが少ないのが現状である。日本国内においては，NICE や ICNALE といった日本人英語学習者コーパスが利用可能になり，条件統制の必要性も意識されつつあるが，前述のとおり，詳細な統制が行われているかどうか，行われたとして，諸条件が十分に明記されているかどうかの問題があり，第二言語習得研究の利用上の課題を抱えている状態にある。

これまで，第二言語習得研究の利用を想定した学習者コーパスの構築が十分に行われてこなかった背景に

ついて，Tono (2003) では，以下のように指摘されている。

> Many corpus-based researchers do not know enough about the theoretical background of SLA research to communicate with them [i.e. SLA researchers] effectively, while SLA researchers typically know little about what corpora can do for them.
>
> (Tono 2003: 806)

つまりは，両分野における十分な接点がなかったことで，第二言語習得研究に有用なコーパスが構築し得なかったということが言える。これらの状況を踏まえ，条件統制に関しては，1) 産出環境：自然産出か強制産出か，2) 時間制限の有無，3) 産出時のテーマ・トピック，4) 辞書使用の許可・禁止，5) 媒体：手書きかタイピングか，といった条件に加え，産出前に何らかの指示（文体や目指すべき語数など）を与えたかどうか，与えた場合はどのような指示を与えたか，などの諸条件を定める，明記することが研究利用上必要になってくるだろう。

また，習熟度の測定方法とその記録については，TOEIC® 等の理解面を測定するような標準テストのみに頼ることなく，TOEIC® S & W テストなど，ライティングやスピーキングといった産出面の能力を測定ができるようなテストを実施し，そのスコアをデータとして含めるべきである。さらには，産出能力を構成するものとして，語彙サイズや文法知識，ワーキングメモリの容量などを測定できるようなテストの結果も，各個人の属性データとして含めることで，学習者の習熟度による詳細な分類を行うことができるようになるはずである。

3．Hiroshima Interlanguage Corpus (HIC) の構築

前節で述べた既存の学習者コーパスが抱える課題を踏まえ，第二言語習得研究用の学習者コーパスとして，新規に Hiroshima Interlanguage Corpus（HIC）を構築した。HIC の総語数は執筆時点でおよそ３万語となり，既存の学習者コーパスと比較した際には，小規模な部類に入るものであるが，さまざまな条件統制や新たな属性情報などを加えて構築を続けている学習者コーパスである。本章では，その構築手順・特徴について説明する。

3.1 HIC の構築手順

既存の学習者コーパスが抱える問題点として，主に，個人の属性情報に係る変数の記述に不十分な点があること，さらに，産出させた際の過程を観察できないことという２点を述べた。今回構築した HIC では，属性情報として，すべての学習者から TOEIC® IP テストのスコアを習熟度に関する１つの情報として含めている。既存のものでは，一部のデータで記録がないものや，複数の資格試験のデータを取り入れ，分析時に換算するような形で用いることも可能だが，HIC では，TOEIC® IP テストのスコアを統一して利用している。

さらに，ライティングの習熟度を判定するための１つの観点として，現在は各データに記録されてはいないが，タイピングの平均速度を新たな属性情報をとして含めることを計画した。なお，タイピングの平均速度というのは，ある英文を見ながらそれをタイプさせるというタスクから得た時間データから算出され記録されたものである。習熟度判定のための１つの材料として平均速度を用いるのは，習熟度が高く，言語の処理が速く行われれば，タイピングにもその速さが反映される可能性があるという着想に基づくものである。加えて，産出させた際の過程を１文字単位で時間データとともに別ファイルとして記録をしており，１人の学習者に対し，最終結果の作文が記録されたデータと，過程が記録されたデータの２種類が作成されている。これら２種類のデータを組み合わせて分析することで，どのような過程を経て，最終的な作文が産出されたのかを観察することができるようになっている。

HIC の構築にあたっては，日本人大学生の学習者の中でも，習熟度が中級レベルにある学習者94名からデータを得た。対象者となった学習者の TOEIC® IP テストにおける平均スコアは，479.5点（標準偏差は69.9）であり，最低点は305点，最高点は655点であった。図１にスコア取得者のヒストグラムを示す。

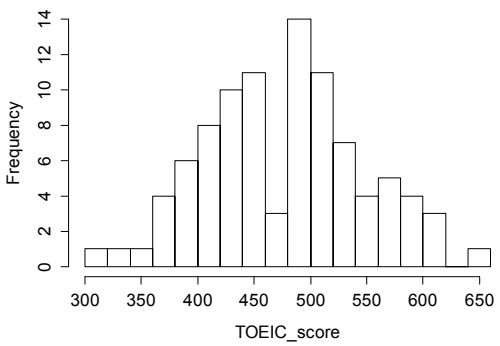

図１　スコア取得者のヒストグラム

図１から，400点台から500点台の学習者が多くを占めており，全体的に中級レベルの習熟度にあることが見て取れる。ただし，一部のスコアに偏りが見られ，最低点と最高点の差が300点を超えていることから，研究目的によっては，スコアが低い学習者のデータについて，その扱いを検討する必要があるだろう。

今回の学習者については，いわゆる４技能（リーディング，リスニング，ライティング，スピーキング）の能力を10段階で自己評価させており，その結果も属性情報に含めることで，データの分類に活かすことができるようになる。例えば，英語を書くことについて自己評価が高い学習者と低く評価する学習者で比較し，書く内容や文体に差が見られるかといった新たな観点での調査が可能になると考えられる。

分析データとして利用したのは，同一の学習者が第１回または第２回の授業および第15回の授業で書いた２種類のエッセイである。トピックは，「学校教育」で統一し，初期と後期での比較ができるようになっている。なお，最終回の授業において，研究用データとしてエッセイを使用することについての同意を得ている。なお，授業では，アカデミック・ライティングの基本を学ぶことを目的とし，表２に示すような条件で収集された。

表２　英作文の提出時期および条件

授業	活動内容	時間制限	参照物	トピック
Week 1/2	１回目の作文収集	50 分間	なし	学校教育
Week 2/3-14	各種演習	なし	あり	Part-time job 等
Week 15	２回目の作文収集	50 分間	なし	学校教育

Week 2/3から Week 14の間には，あるトピックに応じて英文エッセイを書く活動を中心とし，エッセイを書くにあたり，授業を通して，パラグラフ内にトピック文やトピック支持文を含めること，既知情報と新情報の流れに沿って文を配置することなど，情報構造や文章構成法に関する知識を得ていることに加え，Google を使った英語表現の検索方法（フレーズ検索）などについて指導を受けている。

さらに，教員がエッセイの内容面・形式面について助言するだけでなく，学習者自身も，相互批評の機会を通じて，他の受講生からの内容面・形式面についての助言を受けて文章の修正を行っている。なお，授業時間外の課題としてエッセイを加筆・修正することもあり，辞書や機械翻訳を参考にしたものと思われる英文も含まれているため，必ずしも産出時点で学習者が有している知識のみでエッセイが仕上がっているわけではない。そこで，今回は，学習者が参照物を使わず，時間制限がある中で，その時点で有している知識のみで書かれた作文データを利用することにした。なお，どのような指導を受けたかによって，学習者が産出する英文には何らかの影響が出るものと予測されることから，指導方法や内容についても変数化し記録するべきであるが，今回のデータでは変数化および記録はなされていない。この点については今後の課題となるだろう。

データの収集にあたっては，杉浦正利氏（名古屋大学）が作成したキーボード入力の記録システムを一部改変して利用している。このプログラムは，Hot Soup Processor（HSP）という言語によって作成されており，文字がタイプされた際の文字やその時の時間をミリ秒単位で計測することにも対応している。入力画面は，図２に示すように，極力簡素なものとし，英文を入力することに集中できるインターフェイスとなっている。

なお，コーパス構築にあたっては，その効率を考えて，収集時に PC を利用することも少なくない。例えば，NICE では，Microsoft Word のようなアプリケーションを利用しているため，学習者が気づく範囲での単語の修正がなされることになる。今回利用したプログラムには，Microsoft Word のような自動校正機能がついていないため，スペリングの間違いなども，学習者本人がその誤りに気づかなければ，そのまま記録されることになる。したがって，コーパス化することで，日本人英語学習者が，どのようなタイプミスをしやすいのかという新たな分析も可能になる。無論，単語数を数える際には，別の未知語としてカウントされることになるため，分析の方針によっては，適宜タイプミスされた単語の修正が必要になる。

図２　英文収集用プログラムの画面

この収集プログラムの最大の特徴は，タイピングの過程として入力された文字列や入力に要した時間を記録している点にある。まず，入力された文字列が記録

されることで，どのような過程を経て最終的にある表現・文が産出されたかが観察できるようになる。これにより，ある表現を途中まで書きながら，推敲の結果，最終的には削除されたり，別の新たな表現に書き換えられたりしている事例も観察できる。この点については，次節で事例を挙げて説明する。

また，手書きをする場合はその産出速度を記録することは困難であるが，コンピュータを介した入力では，速度の記録が可能となる。タイピングには個人差が生じやすく，タイピングが速い学習者もいれば，遅い学習者もいる。同じ時間で十分な速度で入力できない学習者がいれば，分析対象から外す，あるいは，別の観点から分析するといったことが可能になる。全体の時間制限だけではなく，入力媒体に関する時間データを含めた学習者コーパスはほとんどなく，習熟度のみにとどまらず，媒体に対する慣れという観点から，過程データに記録された時間データをもとにデータの選別できる点も，HIC の特徴といえる。

3.2 HIC の概要

50分間の英文作成後に生成されるデータの記録ファイルから，英文データのみを抽出し，１名ごとにデータファイルを作成した上で，英文不要な文字や記号の削除・置換処理，Perl および UNIX コマンドによるテキスト処理を行って，産出データを集計した。

まず，Week 1/2と Week 15に英作文データをもとに構築した各コーパスの語数などの基本的な数値を求めた結果を表３に示す。なお，語数の集計には，UNIX コマンドの wc（word count）コマンドを使用している。

表３　各エッセイの総語数・最大語数・最小語数・平均語数（N = 94）

	Week 1/2	Week 15
総語数	15728	18263
最大語数	276	328
最小語数	70	94
平均語数	167.3 (*SD* =49.3)	194.3 (*SD* =48.3)

表３から，Week 1あるいは Week 2の授業初期に収集したデータよりも，授業終盤で収集したデータの方が初期のデータよりも規模が増しており，一人あたりの平均語数が増加していることが分かる。

次に，各学習者の Week 1/2および Week 15に書いた作文の総語数の散布図を図３に示す。点線の補助線よりも左側にあるプロット（丸点）は，ある学習者の Week 15の総語数が Week 1/2の総語数を上回っていたことを意味する。

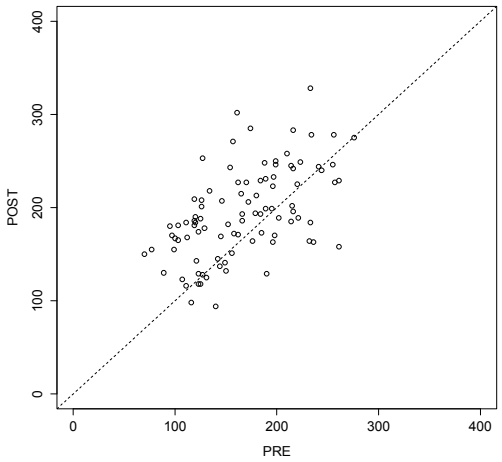

図３　Week 1/2（PRE）および Week 15（POST）で書かれた作文の総語数の散布図

表２で示したように，作文のトピックは統一されていることから，トピックの違いによって語数が増減したわけでなく，学習者が教室内外での指導や学習を通じて，より多くの語を産出できるようになったことが分かる。

これまで述べてきたように，HIC は，同一の学習者が同一のトピックで50分という制限時間を設けて産出したデータから構成され，加えて，母語話者による添削も行われており，既存の学習者コーパスの利点を踏襲しつつ，課題の克服を試みた新規の学習者コーパスとなっている。次の図４は，HIC のデータの一部である。

```
@Begin
@PID:    NNS002
@Proficiency: TOEIC_430
*NNS002: It seems that students should use iPad in the class.
*NTV: I think students should be allowed to use iPads in class.
%COM: It seems describes unclear situations, not opinions. Plural subjects (students) need plural objects (iPads).
*NNS002: Everyone knows that iPad is very useful item.
*NTV: Everyone knowns that iPads are very useful items.
%COM: Use plural forms when speaking generally.
（以下省略）
```

図４　HIC（NNS002）のデータの一部

HIC のデータは，CHILDES の記録形式と同じ CHAT フォーマットを採用し，１行に１文が記録された状態で記録されている。なお，ヘッダー情報として，@で始まる行には学習者の属性情報，＊ NNS の行には学習者により産出された英文，＊ NTV の行には母語話者による添削文，＊ COM の行には母語話者による補足情報が付与されている。なお，入力された文字列やその時間データは，別のファイルとして記録されている。

3.3 HIC 内の高頻度表現

本節では，HIC において高頻度で出現した表現について報告する。具体的には，N-gram 表現を抽出する。N-gram 表現とは，機械的に連続した単語を抽出して得られた表現のことであり，N には任意の数値が入る。例えば，"I lived in Hiroshima three years ago." という文から，2-gram 表現を抽出すると，"I lived", "lived in", "in Hiroshima", "Hiroshima three","three years","years ago. " という６種類の表現が抽出されることになる。機械的な抽出を行うことから，例えば，“Hiroshima three”のように，必ずしも意味的にまとまりのある表現が得られるとは限らない。しかしながら，データの規模が増すと，一定の傾向が見られることから，頻出表現の抽出に用いられる手法としては一般的なものとなっている。今回は，データの規模が約３万語ということを考慮し，2-gram 表現（表４）および3-gram 表現（表５）を報告する。

まず，表４から，*it is* や *I think* のような表現が，PRE および POST において，高頻度で出現している

表４　HIC から得られた2-gram の比較（上位20表現）

順位	PRE	頻度	POST	頻度
1	i think	161	it is	140
2	think that	85	in the	122
3	it is	85	i think	116
4	there are	59	there are	74
5	have to	53	school education	65
6	go to	53	go to	64
7	in the	52	high school	61
8	school is	50	think that	60
9	is very	45	of the	59
10	a lot	44	a lot	57
11	want to	43	lot of	56
12	high school	43	we can	53
13	lot of	39	have to	51
14	of the	37	want to	50
15	is not	37	to study	49
16	to study	33	is the	47
17	they can	32	is not	45
18	school education	31	education is	43
19	for example,	30	is very	42
20	to be	29	school is	39

表５　HIC から得られた3-gram の比較（上位20表現）

順位	PRE	頻度	POST	頻度
1	i think that	54	a lot of	56
2	a lot of	39	i think that	53
3	there are many	17	school education is	26
4	i want to	17	the number of	25
5	and so on.	16	there are many	22
6	when i was	15	junior high school	22
7	is very important	15	high school students	22
8	go to school	14	and so on.	22
9	education of school	14	it is important	19
10	to go to	13	go to school	19
11	the most important	13	the most important	18
12	so i think	13	is very important	18
13	there are some	12	that it is	15
14	school. i think	12	we have to	14
15	go to school.	12	we can learn	14
16	we have to	11	it seems that	13
17	they have to	11	it is not	13
18	the number of	11	to go to	12
19	school education is	11	the students who	12
20	junior high school	11	in the school.	12

ことが分かる。これらの表現は，文頭で使用され，特に *I think* については「～だと思う」という母語から英語への変換処理を行った結果として多く用いられたものと推測される。一方で，*in the* という前置詞句が見られるが，PRE よりも POST で頻度が大きく増加しているが，これは，日本人学習者が前置詞句の産出が十分にできないことが先行研究で明らかとなっており，教室内での指導を通じて，産出が増えたものと思われる。

次に，表５から，*I think that* が PRE・POST のどちらでも高頻度で出現していることが分かる。しかし，頻度そのものには大きな差は見られない。さらに，POST の16位の表現として *it seems that* が見られるが，これは，学生が使用していた教科書に掲載されていた表現であり，授業内で *I think that* の多用を避けるように指導を受けたことによる影響と考えられる。

表４と表５の結果を総合すると，主語として用いる単語としては，*I*，*it*，*we*，そして，トピックとなっている *school education* が多く，多種多様な主語・無生物主語構文を用いて表現するほどの習熟度に達してはいない様子が伺える。これらの結果を踏まえ，母語話者による添削結果も加えることで，どのような表現・文法項目の習得が困難かについて，より詳細な分析を今後行うことができるだろう。

3.4 作文の過程データ

本節では，ICLE や日本国内の既存の学習者コーパスには含まれていない，何がどのように書かれたか，あるいは，削除されたかを示す過程データについて説明する。図５に示すのは，その過程データの一部である。

```
[SP]
s[BS]
[BS][BS]
ne[BS][BS][BS][BS]
[BS]
a[BS][SF]all[SP]over[SP]the[SP]world,[SP]th
ere[SP]is[SP]som
e
[BS][BS][BS][BS][BS][BS][BS]are[SP]some[SP]
people[SP]
that[SP]
not[SP]to[SP]go[SP]o[BS]to[SP]scholl[SP][BS
][BS][BS]ol.
[SP]
[SF][SF][SF][SF][SF][SF][SF][SF][SF][SF][SF
]and[SP]
some[SP]of[SP]them[SP]
can[SP]
not[SP]
write
[BS]e[SP]a
[BS]or[SP]
read[SP]book
.[SP]
[SF]so[SP]they[SP]cant[SP][BS][BS][SP]not[S
P]work
.
[SP]
[BS][BS][BS][BS]who
[BS]
 [SP]
[ET]
[ET]
[ET]
[ET]
[ET]
[BS]
[ET]
[BS]theyt[BS][SP]helo[BS]p[SP]the
[BS]
eir[SP]family[SP]
[BS]
.
```

図５　HIC NNS002（PRE）による産出過程の一部
注　[BS] はバックスペース，[SF] はシフト，[SP] はスペース，[ET] はエンターの入力を示す。

作文の過程データは，個人ごとの入力速度から得た閾値を超えると，その部分がポーズとみなされ，改行されて記録されるようになっている。つまり，１行に記録されているのは，ポーズを含まない連続した産出ということになる。

今回の過程データに含まれる関係詞に着目すると，1）最初は，関係詞 *that* を産出し，2）次の文を書き終え，前の文の修正を行ってから，3）修正として，*that* を削除して，*who* に置き換える，という一連の編集過程が見て取れる。このように，最初に書かれたものが削除され，別の表現に差し替えられたという現象は，既存のコーパスでは観察が難しく，今回の過程のデータを観察することで初めて解明できるものであり，今後新たな観点を加えた習得研究につながる可能性がある。

４．おわりに

本稿では，既存の学習者コーパスが第二言語習得研究での利用時に生じる問題点を指摘した上で，新規に HIC を構築し，その概要と特徴を説明した。特に，産出時の過程を記録することで，作文結果のみに依存しない言語分析が可能となっている。HIC の構築により，実験データとして得られる正答率や反応時間といった従来のデータに加え，第二言語習得研究用の新たなデータとして学習者コーパスの有用性が高まったと言える。

【参考文献】

Granger, S. (2002). A bird's-eye view of computer learner corpus research. In S. Granger, J. Hung, S. Petch-Tyson, & J. Hulstijn (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (Vol. 6, pp. 3–33). Amsterdam & Philadelphia: John Benjamins.

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. In Ishikawa, S. (Ed.), *Learner Corpus Studies in Asia and the World Vol.1* (Kobe University), 91-118.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. Third Edition.* Mahwah, N.J.: Lawrence Erlbaum Associates.

大関浩美 (2008).『第一・第二言語における日本語名詞修飾節の習得過程』東京：くろしお出版.

Tono, Y. (2003). Learner corpora: design, development and applications. In D. Archer, P. Rayson, A. Wilson, & T. McEnery (Eds.), (Vol. 16, pp. 800–809). Presented at the Corpus Linguistics 2003 Conference (CL 2003), Lancaster (UK): Lancaster University: University Centre for Computer Corpus Research on Language.

投野由紀夫・金子朝子・杉浦正利・和泉絵美（編著）(2013).『英語学習者コーパス活用ハンドブック』東京：大修館書店.