縦断的学習者コーパスを用いた英語表現の経時変化の分析

阪　上　辰　也
広島大学外国語教育研究センター

１．はじめに

本稿の目的は，中級レベルの日本人英語学習者が，およそ４ヶ月間にわたるライティングの授業を通じて，使用する英語表現の種類がどのように変化するのかを明らかにすることである。

国内外においては，すでに複数の学習者コーパスが構築され，その分析結果から，習熟度によって学習者が使用する英語表現の種類やその頻度に違いが見られることが分かっている。

しかしながら，阪上（2013）において指摘されているように，既存の学習者コーパスは，異なる学習者がある習熟度に達している状態で産出されたものを中心にデータが構成されており，横断的な分析を行うことが専らとなっていた。つまり，既存の学習者コーパスでは，縦断的な分析を行うことが困難であるため，そこで阪上（2013）では，同一の学習者が複数回書いた作文データをもとにコーパスを構築し，使用された表現の変化を観察している。ただし，時間制限のない作文データを使用しており、産出条件が統一されていない点が課題となっていた。

そこで，本稿では，新たな作文データをもとに学習者コーパスを２種類構築し，４ヶ月間に渡るライティングの授業を受けた結果，使用した表現にどのような変化が生じたか，また，上級レベルの学習者データとの比較を通し，どのような違いが見られるかを明らかにする。

２．学習者コーパスを用いた学習者の言語使用に関する先行研究

代表的な学習者コーパスとして，世界最大規模の International Corpus of Learner English（ICLE），アジア圏10か国の英語学習者（日本人英語学習者も含む）から収集した International Corpus Network of Asian Learners of English（ICNALE），日本人英語学習者に特化したコーパスには，話し言葉を集めた NICT JLE（Japanese Learner English）Corpus，書き言葉から構築された Nagoya Interlanguage Corpus of English（NICE）や JEFLL（Japanese EFL Learner）Corpus が挙げられ，これらのコーパスを用いることで，学習者による英語の使用実態をより客観的に観察できるようになってきた。

学習者コーパスを用いた研究は1990年代後半から盛んに行われるようになり，De Cock, Granger, Leech, and McEnery（1998）や Aijmer（2002）などの先行研究から明らかなことは，１）母語話者も学習者も高頻度で使用する表現があること，しかし，２）その多用された表現の種類と用法は母語話者と学習者では異なっていることの２点に集約される。このような傾向は，日本人学習者コーパスを分析した研究においても見られることである（阪上・古泉，2008；阪上，2012）。

しかしながら，これまでの研究で，阪上（2013）のように，同じ学習者が書いた複数の作文を比較し，使用した表現にどのような変化が見られるかを観察した事例は多くない。そこで，本研究では，中級レベルの日本人英語学習者が異なる時期に書いた２つの英文エッセイを比較し，どのような変化が見られるか，また，上級レベル学習者との比較も行うことで，どのような共通点・相違点が見られるかを明らかにする。

３．データの収集とコーパス化の手順

3.1 対象者および作文時の条件

本研究では，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89名分のデータを使用する。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520.9点（標準偏差は52.2）であり，最低点は390点，最高点は670点であった（1名は記録が無く，88名分での集計結果)。図１にスコア取得者のヒストグラムを示す。

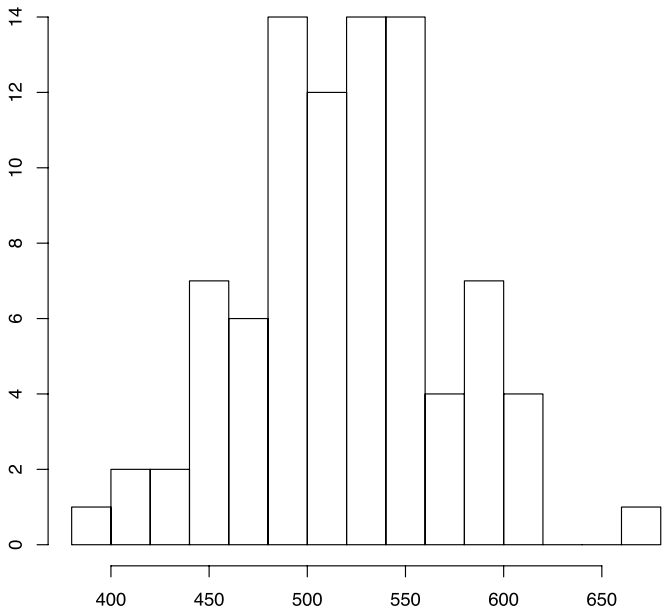

図１　分析対象となった学習者が取得した TOEIC IP テストのスコアのヒストグラム

分析データとして利用したのは，学習者が第２回の授業（Week 2）および第15回の授業（Week 15）で書いた２種類のエッセイである。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。

表１　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 2	１回目の作文収集	50分間	なし	テレビの暴力シーン
Week 3-7	中間エッセイの執筆と提出	なし	あり	携帯電話の利用
Week 8-14	期末エッセイの執筆と提出	なし	あり	早期英語教育の是非
Week 15	２回目の作文収集	50分間	なし	学校教育

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）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をするだけでなく，相互批評の機会を設けて，他の受講生からの内容面・形式面についての助言も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，辞書や機械翻訳を用いた結果と思われる英文も含まれているため，中間エッセイや期末エッセイについては，必ずしも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 Week 2および Week 15の作文データを利用することにした。

3.2 データの処理とコーパス化

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た[1]。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。

図２　英文収集用プログラムの画面

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，阪上（2013）の手順に倣い，英文に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（*N*-gram 抽出）を行って，頻度データを集計した。

４．結果と考察

4.1 語彙頻度一覧および使用表現の経時変化

まず，Week 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，その結果を表２に示す。なお，語数の集計には，UNIX コマンドである wc（word count）コマンドを使用した。

表２　各エッセイの総語数・最大語数・最小語数・平均語数（*N* = 89）

	Week 2	Week 15
総語数	13,141	17,333
最大語数	243	322
最小語数	63	68
平均語数	147.7（*SD* = 38.0）	194.8（*SD* = 50.9）

次に，各学習者の Week 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。補助線よりも左側にあるプロットは，Week 15の総語数が Week 2の総語数を上回っていたことを意味する。この結果から，学習者が教室内での指導および教室外での学習の結果として，同一時間内により多くの語を産出することができるようになったことが分かる。

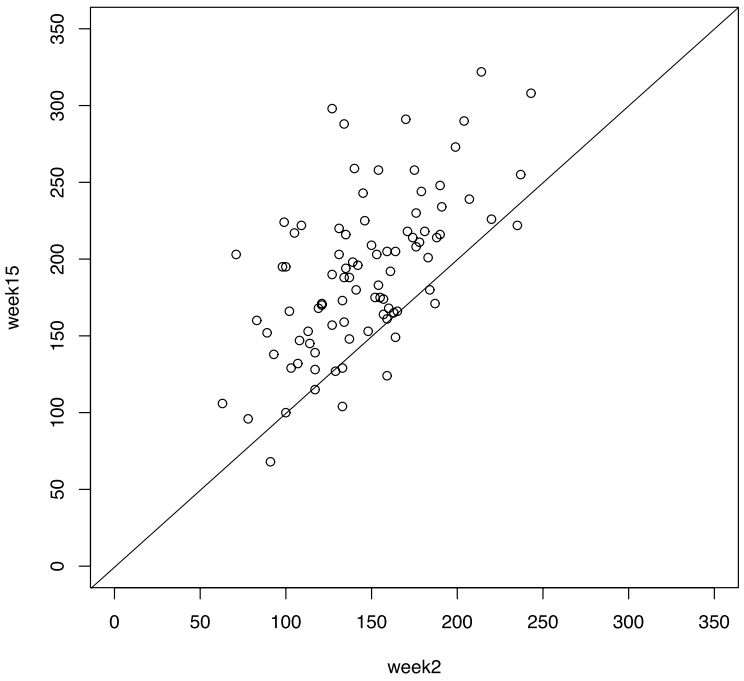

図３　Week 2および Week 15で書かれた作文の総語数の散布図

中間エッセイを集めたコーパスと，期末エッセイを集めたコーパスから得られた語彙頻度一覧から上位30語を表３に示す。なお，データ処理の結果として，すべての英文字が小文字に変換されているため，*I* や文頭で使われていた単語もすべて小文字で表記している。

表３　各エッセイのコーパスから得られた語彙頻度一覧の比較（上位30語）

	Week 2		Week 15	
順位	頻度	表　現	頻度	表　現
1	421	tv	664	the
2	398	the	663	to
3	372	is	443	of
4	336	violence	443	in
5	323	to	423	is
6	306	that	413	school
7	300	i	335	and
8	294	children	311	that
9	220	think	294	students
10	218	of	275	a
11	205	and	244	not
12	198	it	231	i
13	194	not	230	are
14	190	they	222	they
15	185	in	207	children
16	183	are	205	we
17	183	a	186	it
18	168	bad	181	study
19	165	people	169	for
20	163	for	168	education
21	152	on	160	have
22	134	we	157	their
23	129	watch	152	do
24	118	scene	150	can
25	117	program	137	should
26	115	but	114	high
27	111	so	113	there
28	100	have	110	teachers
29	100	be	109	think
30	97	violent	107	people

表３から，Week 2のコーパスでは，*TV* や *violence* といった作文のトピック直接関係する高頻度で使用されていることが分かる。一方で，Week 15のコーパスでは，Week 2の場合とは異なり，定冠詞 *the* や *in* や *to* といった前置詞が上位を占めている。さらに，Week 2の第9位にある *think* については，*I think that* のような表現を多用したことにより，高頻度語として現れたものであるが，Week 15のデータでは，第29位に現れており，時期を経て使用頻度が下がっていることが見て取れる。

4.2 中級レベル学習者と上級レベル学習者の共起表現の比較

表現レベルでの違いを観察するために，*N*-gram の生成を行った。なお，比較のために，上級レベルの学習者データとして，NICE の学習者データ（NICE-NNS）を使用する。それぞれの作文データの収集条件を表４に示す。

表４　各コーパスのデータ収集条件

条件等	Week 2/15	NICE-NNS
データ数[2]	89	96
制限時間	50分間	60分間
参照物	なし	なし
TOEIC 平均スコア	520.8（*SD*＝52.2）	849.8（*SD*＝70.8）

NICE-NNS の収集条件について，Week 2および Week 15の条件と異なるのは，制限時間と作文のトピックである。制限時間は, NICE-NNS の方が10分間長くなっている。また, 作文のトピックとして，NICE-NNS の中には，スポーツ，学校教育，死刑問題の是非などが含まれており，Week 2や Week 15のトピックと類似するものもあれば，そうでないものもある。トピックに応じて使用される単語に違いが出ることが予想され，特に名詞については，使用頻度に偏りもしくは差が見られると予想される。

続いて，各コーパスの概要を表５に示す。

表５　各コーパスの概要

	Week 2	Week 15	NICE-NNS
総語数	13,141	17,333	41,923
平均語数	147.7 （*SD*＝38.0）	194.8 （*SD*＝50.9）	436.7 （*SD*＝118.0）

表５を見ると，まず，Week 2と Week 15の比較から，すべての値が上昇していることが分かる。さらに，これらの値と NICE-NNS の値を比較すると，制限時間が10分長いという違いはあるものの，平均語数は中級レベルの学習者のデータを上回っており，２倍以上の差がある。

次節以降で，各エッセイのコーパスから得られた 2-gram，3-gram の結果を報告するとともに，その傾向を記述する。

4.2.1 共起表現の比較：2-gram 表現の比較

各エッセイのコーパスから得られた2-gram の表現を表６に示す。

表6　各エッセイのコーパスから得られた2-gram の比較（上位30表現）

順位	頻度	表現（Week 2）	頻度	表現（Week 15）	頻度	表現（NICE-NNS）
1	128	i think	108	it is	197	it is
2	103	think that	95	do not	187	of the
3	89	it is	79	in the	132	in the
4	59	tv program	73	there are	121	i think
5	51	there are	71	go to	96	have to
6	50	is not	65	lot of	91	there are
7	47	on tv	65	a lot	87	to be
8	45	violence is	60	high school	80	high school
9	40	the violence	51	want to	72	a lot
10	40	that violence	48	to study	71	to the
11	39	on tv.	48	i think	65	lot of
12	37	want to	46	of the	63	to do
13	35	of violence	42	think that	63	people who
14	33	tv programs	39	have to	61	is the
15	33	the tv	36	that the	61	and the
16	33	for example,	36	is not	60	for the
17	33	do not	36	for example,	58	is not
18	32	to watch	35	to the	57	one of
19	31	in tv	35	school education	56	they are
20	30	that the	34	school and	55	on the
21	29	watch the	32	is that	55	do not
22	29	violence scenes	30	we can	50	that they
23	29	of tv	30	to school	49	that the
24	29	lot of	29	to be	48	think that
25	29	a lot	29	junior high	47	to make
26	28	tv programs.	28	we should	47	to learn
27	28	they are	28	the school	47	like to
28	27	of the	28	school is	46	is a
29	26	violence scene	27	number of	45	want to
30	25	in the	26	to learn	45	i have

表6から，Week 2の2-gram には，暴力シーンというトピックに直接関わる *violence* を含んだ表現が上位にあることが分かる。一方で，Week 15や NICE-NNS の2-gram には，*in the* や *of the* といった機能語のみで構成される2-gram が上位に現れているとともに，*of* を中心とした前置詞を含んだ2-gram の種類が増えていることが分かる。また，すべてのコーパスに共通しているのは，*I* を主語として文を産出すること，また，*it is* や *there is/are* を使った構文を使用していることである。こうした傾向は，参照物があり，時間制限のないデータを分析した阪上（2013）でも確認されている。

なお，Week 2のデータでは *I think* を使用した表現が上位に現れているが，Week 15では使用頻度が下がっている。しかし，上級レベルの学習者のデータを見ると，上位に現れており，上級者でも比較的使用頻度の高い表現であることが分かる。Week 15における *I think* の使用頻度が他のデータに比べて高くなっていない要因として考えられるのは，指導の影響である。学習者が

書いた作文については，教員からの助言や指導が加わるため，その内容に応じて表現を変えた結果だと言える。つまり，Week 2以降の作文に，*I think* が多用されており，その過剰使用を避けるよう指導を受けた結果として，*I think* の使用頻度が下がり，他の表現への置き換えが行われたものと考えられる。

4.2.2 共起表現の比較：3-gram の表現の比較

各エッセイのコーパスから得られた3-gram の表現を表７に示す。

表７　各エッセイのコーパスから得られた3-gram の比較（上位30表現）

順位	頻度	表現（Week 2）	頻度	表現（Week 15）	頻度	表現（NICE-NNS）
1	72	i think that	65	a lot of	65	a lot of
2	29	a lot of	31	i think that	42	one of the
3	20	think that violence	23	go to school	33	would like to
4	15	think that the	22	the number of	33	i think that
5	15	that violence is	21	to go to	31	i would like
6	13	there are many	20	high school and	23	the other hand,
7	13	i don't think	19	students do not	23	on the other
8	12	scene on tv	19	and so on.	21	there are many
9	12	is not good	18	they do not	21	is one of
10	11	tv program is	17	there are many	20	think it is
11	11	to watch the	17	it seems that	20	junior high school
12	11	reason is that	17	it is important	20	i think it
13	11	on the tv	15	when i was	19	it is very
14	11	is not good.	15	in high school	19	it is not
15	10	we have to	15	education of school	19	and so on.
16	10	violence on tv	15	do not have	17	we have to
17	10	the violence seen	14	there are two	16	when i was
18	10	that the violence	14	reason is that	16	they have to
19	10	it is not	14	junior high school	16	in elementary school
20	10	i do not	14	it is not	15	that it is
21	9	violence scenes in	13	japanese young people	15	in order to
22	9	violence is bad.	12	school education is	15	english education in
23	9	they do not	12	high school students	15	be able to
24	9	the scene of	12	first of all,	15	a part time
25	9	on tv is	11	the school education	14	we need to
26	9	of violence on	11	in the future.	14	to learn english
27	9	not good for	11	in order to	14	they want to
28	9	and so on.	11	i want to	14	the people who
29	8	violence scene on	11	go to the	14	reason why i
30	8	violence is not	11	go to school.	14	is very important

表７から，2-gram の一覧と同様に，暴力シーンというトピックに関わる *violence* を含んだ3-gram 表現や，学校教育に関わる *school* や *students* などを含んだ3-gram が上位を占めていることが分かる。また，2-gram 表現でも *think* を含んだ表現が上位に現れていたが，3-gram 表現についても，各コーパスに共通して使用されていた。ちなみに，Week 15のデータでは，類似表現として *it seems that* が見られるが，これは学習者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* の多用を避け，学習者が新たに学んだ代用表現として産出したものと思われる。

しかしながら，NICE-NNS では，*I think that* の使用頻度が上位にあり，中級レベル学習者の初期の傾向と類似している。その要因として，データ収集時の環境が影響している可能性がある。中級レベル学習者のデータは，英語で文章を書くための継続的な指導を受けた後で Week 15に作文を行うため，授業で学んだ *it seems that* という表現が記憶として長く保持され，なおかつ，*I think that* を多用しないよう指導された結果として *I think that* の使用頻度が低くなるという一時的な現象に至ったのではないかと考えられる。

一方で，NICE のデータは，継続的な指導がなく，その場ですぐさま書くことを要求されているため，作文のための準備がほとんどできず，より即時的な処理が必要となり，接触頻度の高さなどが影響して *I think that* を多用することになったものと考えられる。上級レベル学習者であれば，*it seems that* という表現をまったく知らない可能性は低いと考えられるが，継続的な指導がなくすぐさま書かなければならない環境では，*I think that* が *it seems that* などの別の表現よりも優先的に処理された可能性がある。したがって，一定時間を経て，なおかつ，継続的な指導が途切れた状態で，再度作文をすることを要求した場合には，習熟度に関係なく *I think that* が多用されるだろうと予測されるが，この点については，数年単位での長期的な調査が必要となるだろう。

阪上（2013）では，時間を経ることで，前置詞の使用傾向に変化が見られ，*in*，*to*，*of*，*at* といった多様な前置詞を含む表現が観察されたと報告している。しかし，時間制限があり，参照物がない状態では，顕著な違いが見られず，各コーパスに共通して *of* や *on* を中心とした3-gram 表現が観察されている。時間などが限られた状況では，使用する前置詞に偏りが出る可能性があると考えられるが，どのような要因が働くことで，使用する語に影響が現れるのかについては，さらなる調査が必要である。

５．おわりに

本稿では，中級レベルの日本人英語学習者によって書かれた英作文データをもとにコーパス化した上で，４ヶ月間に渡るライティングの授業を受けた結果，どのような変化が生じたか，また，上級レベルの学習者データとの比較を通し，どのような違いが見られるかを調査した。その結果として，中級レベルの学習者が初期に書く作文では *I think* が多用されるものの，時間を経て，代用表現が使えるようになっていること，そして，上級レベルの学習者データとの比較から，２語・３語単位であれば，使用する表現に共通点が多く見られることがわかった。

最後に，今後の課題を２つ挙げる。１つ目は，個人ごとの表現の変化の分析である。縦断的なデータとはいえ，今回はすべてをまとめた状態で分析しているため，使用される品詞の種類や頻度，文の構造や文章の展開方法などがどのように変化するのかについて，個別に分析することも必要となる。２つ目に，インプットの影響による作文の変化の分析が挙げられる。学習者は，授業の中で受ける教科書の例文，教員からの助言，Web 上にある情報の検索結果といったインプットを受けながら作文を書くわけであるが，そのインプットの影響により，作文にどのような変化が生じているのかを観察する必要もあるだろう。これらは今後の課題としたい。

付記　本稿は，科学研究費補助金 若手研究（B）24720259による研究成果の一部である。

注

1）http://sugiura-ken.org/wiki/wiki.cgi/exp?page=KLogS を参照されたい。

2）Week 2と Week 15は異なる学習者89名がそれぞれ作文をしているが，NICE-NNS に含まれる96個のデータは，一人で複数回作文したデータが含まれているため，延べ人数を示す。

参考文献

Aijmer, K. (2002). Modality in advanced Swedish learners' written interlanguage. In Granger, S., Hung J. & Petch-Tyson S. (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 55–76). John Benjamins.

De Cock S., Granger S., Leech G. and McEnery T. (1998). An automated approach to the phrasicon of EFL learners. In Granger, S. (Ed.), *Learner English on Computer* (pp. 67–79). London: Addison Wesley Longman.

阪上辰也・古泉隆（2008）.「学習者コーパス「NICE」と ANC および BNC おける N-gram 表現の比較」杉浦正利（代表）（2008）『英語学習者のコロケーション知識に関する基礎的研究』（pp. 15–52）平成17～19年度 科学研究費補助金（基盤研究（B））研究成果報告書.

阪上辰也（2012）.「日本人英語学習者による接続詞の使用における母語の影響－英語学習者コーパスと日本語コーパスの比較から」『日本語とX語の対照２－外国語の眼鏡をとおして見る日本語－対照言語学若手の会シンポジウム 2011 発表論文集』13–22.

阪上辰也（2013）.「日本人英語学習者のエッセイに見られる共起表現の分析」『広島外国語教育研究』16, 159–169.

ABSTRACT

An Analysis of Chronological Change of Fixed Expressions in Intermediate-level EFL Learners' Writings with a Longitudinal Learner Corpus

Tatsuya SAKAUE
Institute for Foreign Language Research and Education
Hiroshima University

The purpose of this study is to examine the use of words and fixed expressions by intermediate-level EFL learners through the analysis of learner corpora. Previous studies have revealed that second language learners of English (NNS) used a wide variety of fixed expressions in the same manner as native speakers (NS). However, their types were different from NS, and some phrases were either overused or underused. There are many cross-sectional learner corpus research studies, but few studies on longitudinal design have so far been conducted. In the present study, I would like to analyze a new longitudinal corpus, which consists of writing data compiled by 89 intermediate-level EFL learners.

To conduct the present study, three corpora were used: one is the corpus of writing in which the topic was "violence scenes on TV programs" by intermediate learners (Week 2 Corpus); the second consisted of writing in which the topic was "school education" by the same learners (Week 15 Corpus); and the third was NICE-NNS, which is part of the Nagoya Interlanguage Corpus of English and consists of the writing data of advanced learners. From each corpus, the *N*-gram expressions were extracted, and I analyzed the similarities and differences between them.

The results showed that some fixed expressions were changed throughout the four months of classroom instructions. In Week 2 Corpus, the phrase "I think" was frequently used throughout the instruction; the frequency of "I think that" was reduced, and instead, "It seems that" was used. In comparison to NICE-NNS, the types of expressions used were different, and advanced learners tended to use a greater variety of fixed expressions and prepositions than intermediate learners.