学習者音声コーパスから見えてくるもの

朝尾幸次郎

ABSTRACT

This article discusses the interlanguage of Japanese learners of English as exemplified in the spoken learner corpus that was experimentally developed. The findings reported here are serendipitous, being a by-product of an attempt to create an audio corpus that contains both speech sound and its transcription. During the process of transcribing learners' speech it was noted that the corpus data contained frequent errors that rarely appeared in the traditional written learner corpora. Learner corpora that have been used for second/foreign language acquisition research so far are essentially a collection of written discourse. Which should we look at in order to investigate learners' interlanguage, written or spoken discourse? The findings reported here indicate that a spoken corpus directly reflects interlanguage whereas traditional written corpora represent linguistic performance as a result of monitoring.

Keywords：学習者コーパス，中間言語，話しことば，書きことば，エラー

はじめに

近年，外国語習得研究に学習者コーパスは必須のものとなっている。形態素の習得順序，母語話者と比べた外国語学習者の言語の特徴などが実証的に明らかされたのには学習者コーパスの貢献が大きい。これを支えてきたのが，個人的に構築されてきた数多くの学習者コーパス群であり，現在，完成に向かって進行しつつある国際学習者コーパス（ICLE）である。

ここに報告する内容は音声コーパスを構築する過程で思いがけなく発見されたものである。この研究は当初，音声コーパスの作成を目的としたものであった。英語による学習者の自由な発話を音声として記録し，それを文字に書き起こした後，文字列から音声を検索するしくみを構築することがその構想であった。ところが，発話を文字に書き起こしていくと，そこにはこれまでの学習者コーパスにはまず現れることのないエラーが頻出することに気がついた。中間言語の特徴はエラーに典型的に現れることを考えると，従来の書きことばによるコーパスが中間言語を正しく反映しているものか，疑問が生じた。中間言語を見るには書きことばコーパス，話しことばコーパスのどちらを見ればいいのだろうか。このような観点から，話しことばコーパスと書きことばコーパスを比較した。その結果，中間言語を直接反映しているのは話しことばであり，書きことばは相当な量のモニタリングを経た結果であるとの結論に至った。

1. 音声コーパスの作成

音声コーパスの作成には書きことばコーパスとは違う制約が多い。書きことばコーパスの場合，極端に言えば，題材を指定して学習者に作文をさせたデータを収集するだけで構築は可能である。しかし，音声による発話の収集は，発話してもらうことそのものにまず困難がある。トピックを制限せず自由に決めさせても，データとして十分な一定の時間，学習者に英語を話してもらうことはむずかしい。調査者と会話すれば，ある程度ことばのやりとりを続けることはできる。しかし，学習者の応答は調査者の英語に大きく影響を受けるであろうから，それをコーパスデータとするのはためらわれる

このため，学習者に一定の時間，自由に英語で発話してもらえるよう，次の手順により音声を収集した。

1．次のように指示を与える。「これからできるだけ長い時間，英語で話をしてもらいます。時間は3分間が目標です。しかし，3分を超えても自由に話してかまいません。話すトピックは自由です。もし，適当な話題がみつからない場合は，次のふたつのうちからひとつ選んでください。」
   （1）これまで見た映画のなかでおもしろかったもののストーリーを話す。
   （2）これまででかけた旅行のうち，印象深かった経験を話す。
2．話してもらう前にウォームアップとして調査者と英語でその話題について英語で会話をする。”Was this your first experience going abroad?” “What happened after that?” のような問いかけにより，学習者の記憶を呼び起こし，話す内容を豊富にさせる。
3．録音を始めた後は，学習者に自由に話させる。調査者は相手が話しやすくなるよう，しかも相手の発話の表現，内容に影響を与えないよう，”Is that so?” “That’s good.” のような簡単なあいづちを返すにとどめる。

被験者は英語専攻の大学3年生と4年生7名である。うち6人が海外旅行の経験を，残りひとりが映画のストーリーを話した。このようにして収集した音声を文字に書き起こした。

文字に起こす作業でむずかしかったのはセンテンスの切り分けである。So, and this was my first trip, so I was little nervous.のような発話では2番目に現れるsoの前でセンテンスとして終わっているのかどうか，判定がむずかしい。これはandで発話がつながっている場合も同様である。このような場合，センテンスとしての発話が終了しているかどうかの判定は音調を基準とした。

このようにセンテンスを単位として発話を切り分け，音声ファイルと対応させたものをPerlで記述したCGIにより文字列から検索し，音声を聞くことができるように設定した。この実験コーパスは次に公開している。

http://www.eng.ritsumei.ac.jp/asao/corpus/ej.html

2. 話しことばと書きことば

話しことばコーパスに現れた発話の特徴を見るため，書きことば学習者コーパスと比較した。比較に用いた書きことばコーパスはTravelingというトピックで英語専攻の大学生が12分間で作

文したものである。題材が旅行なので，今回作成した話しことばコーパスと比較するのに好都合である。話しことばコーパスの総語数は4,285語で，これと同等になるよう，書きことばコーパスからは最初の4,186語を比較対象として抜き出した。

表1はトークンとタイプの数，タイプ・トークン比である。タイプ・トークン比の数値は語数により影響を受ける。ここではコーパスの総語数はほぼ同じにそろえたので，そのまま比較してさしつかえないだろう。念のため，1,000語あたりに換算したタイプ・トークン比も示した。標準タイプ・トークン比として示したのがそれである。

	話しことば	書きことば
トークン	4,282	4,186
タイプ	758	804
タイプ・トークン比	18	19
標準タイプ・トークン比	29	33

表1　トークン，タイプ，タイプ・トークン比

話しことばは書きことばと比べてタイプ・トークン比がやや小さく，話しことばの方が冗長で，繰り返しが多い傾向にあることを示している。

表2は両コーパスにおける単語頻度の上位25項目の比較である。話しことばで最も頻度の高い語はandで，出現回数は215である。書きことばでもandの頻度は高く，出現頻度は114で第3位である。しかし，話しことばの方が書きことばよりも頻度は2倍である。このうち文頭にAndが現れるものは話しことばコーパスで50.2% (108回)，書きことばコーパスで17.5% (20回)である。

これを英語母語話者のコーパスのデータと比べてみた。英語母語話者の話しことばコーパスとして使ったのは100万語からなるCorpus of Spoken Professional American English (CSPAE)で，これはアメリカの大学における会議のやりとりを記録したものである。CSPAEではandの出現数は50,639，そのうち文の冒頭に現れる例は13,450で，and全体の26.6%である。英語母語話者の書きことばコーパスとして用いたのはBrown Corpusで，これも総語数は100万語である。Brown Corpusでandは28,708回現れ，そのうち文の冒頭に現れる例は851で，andの出現数のうち3.0%にすぎない。Andで発話を始めるのは話しことば的特徴である。日本人英語学習者の場合，発話の冒頭にAndを使う例は英語母語話者と比べてきわめて多い。英語母語話者と比べた場合，書きことばにその差が極端に表れている。発話の冒頭における，このAndについて言えば，学習者の英語は話しことばと書きことばの区別が希薄である。

発話の冒頭，あるいは途中に接続語を使って発話をつなげるのにはandだけではなく，soも高い頻度で使われている。表2でsoの頻度は話しことばでは第6位で126，書きことばでは第13位で52である。このうち，話しことばでsoが接続語以外の用法で使われているのは10回のみで，実に残り92%が接続語としての用法である。

順位	話しことば		書きことば	
1	*and*	215	*I*	294
2	*the*	206	to	178
3	to	195	*and*	114
4	*I*	160	traveling	112
5	we	151	is	85
6	*so*	126	in	77
7	she	82	*the*	74
8	a	79	travel	65
9	went	64	a	63
10	was	62	we	59
11	*but*	57	go	57
12	very	55	for	53
13	is	50	*so*	52
14	of	49	very	52
15	*ah*	43	my	49
16	in	42	it	48
17	he	35	of	47
18	were	34	many	46
19	you	32	people	39
20	my	29	there	38
21	want	28	*but*	37
22	after	27	that	37
23	that	27	can	35
24	it	25	have	35
25	at	24	think	34

表2　単語頻度

文と文をつなぐはたらきをする語をさらに見てみると，butの頻度は話しことばでは第11位で57回，書きことばでは第21位で37回である。これらand, so, butの頻度をすべて合計すると，話しことばコーパスでは398，実に総語数の9.3%となる。書きことばでも頻度は203，その割合は4.8%にのぼる。英語母語話者の書きことばコーパスであるBrown Corpusではこれら3語の全体に占める割合は3.5%である。これらすべてが接続語の用法であるわけではない。しかし，接続語and, so, butの多用は話しことばに典型的にみられる特徴であることを考えると，学習者の英語は母語話者の英語と比べ，きわめて話しことば的と言える。

これは表3，表4からもあきらかである。表3，表4は学習者と英語母語話者についてso, and, but, because, if, whenの頻度を比較したものである。英語母語話者の話しことばコーパスとして選んだのは上で触れたCSPAEである。英語母語話者の書きことばコーパスとして用いたの

はAnn Landersの身の上相談の文章である。この身の上相談の文章は日常的な話題を扱っており，書きことばながら比較的口語的な文章であることから，学習者の英語に近い性質をもつと考えられる。それぞれ学習者コーパスの語数にみあった分量を取り出して比較に用いた。

	and	so	but	because	if	when
話しことば	215	126	57	17	6	14
書きことば	114	52	37	17	15	27

表 3　学習者の英語

	and	so	but	because	if	when
話しことば	150	38	28	15	21	4
書きことば	116	16	20	9	25	13

表 4　母語話者の英語

母語話者ではandとsoは書きことばよりも話しことばで多用されており，これが話しことば的特徴となっている。学習者コーパスは話しことばにおいても書きことばにおいても母語話者コーパスに比べand, soに関し極端な過剰使用がみられる。とりわけ話しことばでそれが顕著である。

学習者の書きことばが話しことば的であるという特徴は1人称代名詞Iにもみられる。一般に英語母語話者の話しことばで頻度が最も高い語は1人称代名詞Iである。ここで分析している学習者による話しことばコーパスではIの頻度は第4位で160，書きことばコーパスでの頻度は294で，第1位である。母語話者コーパスと異なり，書きことばでの頻度の方が高い。これは作文のトピックによる影響が考えられるものの，学習者のことばは話しことばも書きことばも基本的に差が見られないのである。英語学習者は一見，話すように書いているように見える。しかし，これは学習者のことばは話しことば，書きことばが未分化の状態であるととらえるのが正確であろう。

発話の成熟度を測る指標としてセンテンスあたりの語数，またT-unitあたりの平均語数を見たのが表5である。発話の成熟度を測るにはタイプ・トークン比の他，センテンスの長さがめやすとなる。長いテキストや発話であっても，ひとつひとつの文が短ければ言語運用の成熟度は低い。ただし，andやbutを連ねれば，実質的には短い文であってもみかけ上長くできる。これを補正するために提案されたものがHunt (1965) によるT-unitである。T-unitとはHuntの定義によればminimal terminable unitということである。たとえば，So, and this was my first trip, so I was little nervous.という発話はSo, and this was my first tripとso I was little nervousに分けることができ，それぞれが独立した文として機能しうる。つまり，この文はふたつのT-unitから成ると考えられる。これに対し，Last year, I went to New York with my friend.という発話はセンテンスであると同時にT-unitでもある。これ以上，切り分けると単独で文として成立しない。

	話しことば	書きことば
文	10.11 (SD 5.74)	9.42 (SD 4.35)
T-unit	8.39 (SD 5.02)	9.05 (SD 4.56)

表5　文とT-unitの平均語数

表5からわかるとおり，文レベルで見れば，話しことばの方が書きことばよりも平均語数が多い。これは上で見たとおり，発話がand, but, soでつながれ，みかけ上，文が長くなっているためである。ところがT-unitで見ると，逆に話しことばの平均語数が少なくなる。ただし，全体を通してみれば，その差は少ない。結局，この指標からも学習者の発話は話しことば，書きことばは均質で，いまだ未分化の状態にあることがうかがわれる。

3.　定型表現の多用

学習者の英語にみられるもうひとつ顕著な特徴は定型表現の多用である。これはトライグラム，すなわち3語連結の連鎖の出力結果から知ることができる。表6，表7は学習者コーパス，英語母語話者コーパスにおけるトライグラムの上位20項目である。

表6に示した学習者のことばの特徴は，どちらも上位7位，8位までに現れるトライグラムの頻度が高いという点である。これは表7の母語話者のデータと比べるとよくわかる。書きことばAnn Landersの文章でトライグラムの頻度第1位はdear Ann Landersの16である。しかし，これはAnn Landersの身の上相談コラムに毎回，定型的に現れる表現なので例外である。

話しことばコーパス，書きことばコーパスに現れるトライグラム上位10の頻度の合計を見てみよう。Ann Landersの文章は例外である第1位をはずし，上位2位から11位までを対象とする。学習者の話しことばコーパス，書きことばコーパスに現れるトライグラム上位10の頻度の合計はそれぞれ113，129である。これらはそれぞれ3語連結の表現なので，語数で数えた合計はそれぞれ3倍となり，339，387語である。これらに重複がないと仮定すれば，これらの表現だけで総語数に占める割合はそれぞれ7.9%，9.2%である。実に全体の10分の1近くが定型表現で占められていることになる。これに対し，母語話者コーパスでは話しことばコーパス，書きことばコーパスでその割合は150語，78語である。話しことばコーパスの方が書きことばコーパスの数値の2倍あるのは，話しことばの方が繰り返しが多く，冗長であるためであろう。学習者のことばは母語話者のことばに比べ冗長さが2倍から5倍ある。また，学習者コーパスで話しことばと書きことばの上位トライグラムの頻度に差はほとんどない。これもまた，学習者の英語は基本的に話しことば的で，話しことばと書きことばが未分化であることを示す証左である。

定型表現の多用は初級者になるほど顕著である。表8は高校3年生が「国際人」というトピックで自由英作文した6,012語のコーパスからトライグラムを出力し，その上位20項目を示したものである。トライグラム上位10項目の合計は229で，語数に直せば687語，これが全体に占める割合は11.4%である。上位20項目の合計は339で語数に換算すれば1,017語，この全体に占める割合は16.9%である。大学生の作文では書きことばコーパスに現れるトライグラム上位10項目がコーパス全体に占める割合は9.2%である。高校生の作文での割合11.4%はこれを上回る。

順位	話しことば		書きことば	
1	we went to	24	I want to	23
2	went to the	20	I went to	16
3	I went to	18	and so on	12
4	after that we	10	a lot of	12
5	and after that	9	want to go	12
6	we were very	7	I think that	11
7	said to me	7	I do n't	10
8	n't want to	6	to go to	10
9	that we went	6	when I was	9
10	do n't want	6	I like traveling	9
11	was very surprised	6	want to travel	8
12	so we were	6	traveling very much	8
13	I was very	6	traveling is very	8
14	many kind of	5	this summer vacation	7
15	Ariel 's father	5	like traveling very	6
16	but it 's	5	over the world	6
17	went back to	5	to go abroad	6
18	and we went	5	think that traveling	6
19	to the station	4	it is very	6
20	I really like	4	with my friends	6

表 6　学習者コーパスに現れるトライグラム

順位	話しことば: CSPA		書きことば: Ann Landers	
1	I do n't	6	dear Ann Landers	16
2	the national test	6	do n't have	4
3	one of the	6	I 'd like	3
4	back to the	6	I do n't	3
5	and I think	5	thank you for	3
6	in terms of	5	you know what	3
7	do n't know	4	trying to sell	3
8	it 's a	4	'd like to	3
9	be able to	4	you do n't	3
10	I think we	4	my ability to	2
11	I have a	3	one of his	2
12	in that case	3	confidence in my	2
13	goes back to	3	our Christmas gift	2
14	and there is	3	dan and I	2
15	would like to	3	people with epilepsy	2
16	of the examination	3	did n't know	2
17	I think it	3	realized that I	2
18	of the test	3	do n't want	2
19	if it 's	3	rice at weddings	2
20	on the NAEP	3	for a thank	2

表 7　母語話者コーパスに現れるトライグラム

1	I think that	34	11	with foreign people	13
2	I want to	33	12	to communicate with	13
3	I think it	30	13	I think international	12
4	a lot of	26	14	it is very	11
5	think it is	23	15	international person is	11
6	all over the	20	16	can communicate with	10
7	over the world	19	17	I ca n't	10
8	it is because	16	18	to be a	10
9	want to be	14	19	ca n't speak	10
10	an international person	14	20	international people are	10

表8　高校生の英語に現れるトライグラム

定型表現の多用は学習者の英語が「プレハブ英語」であることを示している。家を建築するには材木を寸法通りに切り取り，壁や床や屋根を作っていく。これに対し，プレハブ工法ではあらかじめ工場で用意された，できあいの壁や床を組み合わせて家を建てる。学習者の英語にも同じプロセスが見られる。あらかじめ用意された定型表現を繰り返し使うことで発話を構成している。このため，発話の一部を見れば文法的で適格な表現が使われているけれども，全体を通してみると自然さに欠けた表現になる。

この定型表現の多用は日本人英語学習者に限らず，各国の英語学習者に広くみられる現象のようだ。Weinert (1995) はformulaic languageということばで，第2言語としての英語学習者にみられるこの特徴を指摘している。このプレハブ的発話が学習者に共通するものであり，言語習得の普遍的なプロセスであるなら，学習者が言語的インプットによって文法体系を作り上げていくという，現在，主流となっている認知的言語学習観の見直しを迫るものになろう。

4. フライングとエラー

書きことばコーパスが中間言語を反映したものと言えるか，その妥当性に疑問を投げかけるのが話しことばコーパスに見られるフライング（false start）とエラーである。どちらも話しことばコーパスには頻出するのに，書きことばコーパスにはそのような例は多くはみられない。

次は話しことばコーパスに現れるフライングの例である。最初の例はvery enjoyと言ったところで，enjoyが動詞であるためveryと共起しないことに気づき，veryを言わずにwe enjoyedと言い直したものである。最後の例は受動態を意識して Continental Airline *was putted* と put に語尾-edをつけて発話したものの，その誤りに気づき言い直したものである。ここで学習者が教わったことのないputtedという表現が使われていることに注目したい。幼児の第一言語としての英語習得においては，動詞の活用の習得について3つの段階があることが知られている。まず不規則動詞が習得される。次に不規則動詞の過去形語尾-edが習得される。しかし，この動詞語尾は規則動詞だけにつくのではなくgoなどの不規則動詞に対しても使われ，goedのような形で現れる。規則動詞と不規則動詞が区別されて現れるのはこの時期を経てからのことである。上

のContinental Airline *was putted*の例は外国語習得の過程にも同じ現象が起こることを示すもので，興味深い。

フライング

… so *we were very enjoy*, we were very ... ah ...*enjoy* ... we *enjoyed* this place.
And Tiffany shop ... Tiffany shop *was* ... *has* four floor ... *had* four floor.
So we *impressed*, we *were impressed*.
But, but, but at last we *choice, choose chose* the design.
… and the shop clerk *come, came* here washing my hair.
And we *are,* we *were* in the same class.
And we together ... ah ... *go* to, *went* to the airport.
Next day she, she *is*, she *was* still bad in condition.
After that we ... *went* ... water, *move, moved*.
And one day she *give* me a, *gave* me a invited card.
Continental Airline was *putted* ... *put* on TV all of seats, so we were surprised.

このようなフライングは書きことば学習者コーパスには現れない。学習者は英語を書くとき，上と同じような心的プロセスをたどっている。しかし，作文として最終的に現れるのはKrashen (1982) の言うモニタリングを経た結果であるため，そのプロセスがコーパスに反映しないのである。

フライング以外に，書きことばコーパスにはまず現れないけれども，話しことばコーパスに頻出するエラーには次のようなものがある。

動詞の人称の一致に関する誤り

So Tako *change* the Ariel to human.
But she tells her that she *have* only three days.
And this three days, she *have* to kiss the human
Otherwise she *become* a very small ‘ikimono’.
But then, and she *lose* her voice and she cannot talk and act only she *have* to kiss to the man.
But he, he *speak* too long, so we, we were very, very tired.
How about ... how about ... she *suggest* to us what do you think of going out by motorcycle.
Because the drama *come* out on TV.
Same as Japanese machine, so machine speak Japanese, …
And this summer he *want* to come here, but it’s too expensive.
And my father *treat* her legs.
Smithsonian Museum *have* a lot of many pictures.

–*ing*直前のbe動詞の脱落

And he don't want to 'seifuku,' the tako wants 'seifuku', so *he going* to fight the tako and he won.

And *we just looking* many kind of thing, sweets or ice cream or book, magazine.

不規則変化動詞の誤った活用

We *putted* on put on black hose and put set hair and pass, put on pumps and like Audrey Hepburn, …

And I said, "Thank you so much." And she *runned*.

不定詞の誤り

… this February I, I decided *to went* to Taiwan with my Japanese friend.

単数形と複数形の誤り

Then, there I, I made, I made *a friends* with Chinese friends.

So she, they asked very *many question*.

And looking around *many shop*, …

まとめ

話しことばコーパスと書きことばコーパスの比較から導かれる知見は3つある。ひとつは学習者の英語は話しことばと書きことばが未分化の状態にあることである。話しことばコーパスと書きことばコーパスを比べると，言語使用の実態がよく似ている。このため，一見，学習者は話すように書き，書くように話すように見える。

もうひとつの知見は，学習者の英語はいわば「プレハブ表現」ともいうべき定型表現で構成されるところが多いという事実である。英語学習者が用いる定型表現に関する先行研究では，学習者が母語話者よりも定型表現を多用するという明確な結論は得られてはいない。De Cock et al. (1998) はautomated phrasiconということばで，英語学習者と母語話者の定型表現の使用を比較している。そこでは定型表現の使い方と種類について差はみられるものの，学習者の方が母語話者よりも定型表現を多用すると述べてはいない。ただし，この調査の被験者はフランス語を母語とする学生であり，日本人英語学習者とは事情が違うことを考慮に入れなければなるまい。プレハブ表現の多用が日本人英語学習者に特有なものか，母語にかかわらず普遍的な事実であるのか，さらに研究が必要であろう。

3つめは話しことばと書きことばのどちらが中間言語の反映であるかという疑問である。これまで学習者コーパスは学習者の作文を収集することにより構築され，書きことばであることを暗黙の了解としていた。また，中間言語ということばで語られる実体は目に見えないものであるものの，それがどのような形で言語運用に反映されて現れるものかについては議論はされてこなかった。話しことばコーパスは書きことばコーパスに現れない情報を数多くもっている

ことを考えると，中間言語は話しことば，書きことば，どちらに典型的に現れるの，あらためて議論が必要であろう。

参照書目

De Cock, S., Granger, S., Leech, G., & McEnery T. (1998). An automated approach to the phrasicon of EFL learners. In S. Granger (Ed.), *Learner English on computer* (pp. 67-79). Harlow, Essex: Addison Wesley Longman.

Hunt, K. W. (1965). *Grammatical structures written at three grade levels*. NCTE Research Report No. 3. Champaign, IL: NCTE.

Krashen, D. S. (1982). *Principles and practice in second language acquisition*. Oxford: Pergamon Press.

Weinert, R. (1995). The role of formulaic language in second language acquisition: A review. *Applied Linguistics, 16* (2) : 180-205.

※本研究は立命館大学国際言語文化研究所／言語教育情報研究科主催シンポジウム「言語理論と英語教育，そしてコーパスの融合を目指して」（2006年12月3日）での発表をもとにしたものである。