様式 C-19

科学研究費補助金研究成果報告書

平成 21 年 5 月 30 日現在

研究種目：若手研究（B）
研究期間：2006～2008
課題番号：18720151
研究課題名（和文）　パラレル学習者コーパスの構築とその統語的・語彙的発達の分析

研究課題名（英文）　Compiling a Parallel Learner Corpus and Analyzing Japanese EFL Learners' Lexical and Syntactic Development

研究代表者
木村 恵（KIMURA MEGUMI）
獨協大学・外国語学部・専任講師
研究者番号：60409555

研究成果の概要：本研究は，これまでにない新しい特徴を持った学習者コーパスを構築することを主目的としていた。(1) 目標言語（英語）による発話データ，(2) 学習者発話を訂正したデータ，(3) 学習者が意図していたことを母語（日本語）で表したデータという，3 つのバージョンが並列されている「パラレル学習者コーパス」である。本研究においては，最終的に日本人英語学習者 50 名分の発話を用い，意図する学習者コーパスの作成に至った。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2006 年度	1,400,000	0	1,400,000
2007 年度	1,100,000	0	1,100,000
2008 年度	1,000,000	300,000	1,300,000
年度			
年度			
総　計	3,500,000	300,000	3,800,000

研究分野：人文学
科研費の分科・細目：言語学・外国語教育
キーワード：英語教育・学習者コーパス

１．研究開始当初の背景

（1）本研究の究極の目標は，日本人学習者が英語の文法知識・語彙知識を身に付けていく過程(＝発達プロセス)を解明し，そのプロセスに基づいたより効果的な英語教育法を提案することであった。研究開始当初の日本の中学，高校，そして大学の英語教育現場では，より確実な教育成果を生み出すことを目指し，E-learning を用いた学習を推進したり，診断型のテストが開発・活用されたりと，より個々の学習者の能力に合わせた指導への関心やそれを重要視する傾向が高まってきていた。しかしながら，個々人の能力に合わせた適切な指導を行う際の根底となるべき，① 学習者の診断情報を与えてくれる指標や，② 比較対象である日本人の一般的な英語力の(発達の)傾向といったものについての情報が不足していると考えられた。つまり，学習者の持つ英語力の実態を把握する体制が充分にできていないという，大きな問題点があることが感じられた。

（2）そこで本研究は，日本人英語学習者のより詳細な習得過程の実体を記述（＝把握）するためのデータ構築をその主目的とした。いわゆる学習者コーパス（learner corpus）の編纂は，本研究開始当初から世界中で活発に行われていた（例えば，International

Corpus of Learner English : ICLE)。その動向の始まりは 1990 年代半ばと思われ，2005 年時点ではいくつかの学習者コーパスが一般にも使用可能な状態となっていた。その中でも，日本人英語学習者を対象とした学習者コーパスである The NICT JLE Corpus や JEFLL Corpus は，英語習熟度別のサブコーパスを持つという特徴があり，英語習得の過程を記述するには非常に適した，優れたコーパスであった。

第 2 言語習得研究において我々は常に学習者のアウトプットからその目標言語の知識を推測するという作業を行っている。しかし，それまで行われてきた学習者コーパスを用いた研究を顧みると，特により綿密な分析が必要と思われる下位学習者においては，沈黙や統語構造・語彙選択の誤りが非常に多く，発話者が本来何を意図して構築したセンテンス（あるいはフレーズ）なのかが推測不可能なことがしばしばあった。いわゆる「何を言いたかったのかがわからない」という状態である。一般的に学習者コーパスは，一定量を確保するために可能な限り大人数から効率よくデータ収集を行って作られてきたが，コンピュータデータベース化された時点では，既に学習者の発話の意図に立ち戻ることが不可能という状況になっているという限界点があった。

（3）このような背景を踏まえ，本研究は既存の学習者コーパスのいずれもが持っていない，パラレル形式のコーパスを構築しようという発想に至った。パラレル形式とはこの場合，学習者の目標言語である英語でのパフォーマンスデータを中心として，同じ発話を母語である日本語で表したデータ，そして更に同じ発話を英語ネイティブスピーカーが訂正したデータの 3 バージョンを備えていることを指す。学習者データを収集する際に発話者本人との面談を設け，学習者の発話－日本語バージョンを加えようという試みであった。この日本語バージョンの存在により，ネイティブスピーカーによる訂正バージョンもより正確に作ることができると思われた。これは今まで行われてきた学習者コーパス編纂のノウハウを充分に活かしながらも，その弱点を補うデータの構築を行おうという意図を持ち，更に新しいタイプの学習者データとして新たな情報を提供してくれるものと期待された。

２．研究の目的

本研究の目的は，日本人英語学習者の実態をより正確に記述するための学習者のパフォーマンスデータを収集し，学習者コーパスを構築することである。そのコーパスは以下の 3 部構成をとる。

（1）学習者の目標言語である英語によるスピーチ（LC-OR : learner corpus, original version）

（2）学習者の目標言語である英語によるスピーチを，英語母語話者が訂正したもの（LC-CR : learner corpus, corrected version）

（3）学習者の目標言語である英語によるスピーチを，日本語に変換したもの（LC-JP : learner corpus, Japanese version）

加えて，得られた学習者データを語彙的・統語的に分析することも，目的の一つである。

３．研究の方法

（1）データ提供者

本研究のデータ提供者は，総計大学生 50 名となった。そのうち，英語を専攻している学生が 31 名，その他の専攻が 19 名である。彼らの英語力の幅は大きく，データ収集時点で最新の TOEIC のスコアが 300 点台から 800 点台までの学生がいた。学生に大学での授業等を通じて研究への協力を呼びかけたり，既にデータ収集に参加した学生に友人の紹介を依頼したりすることで，一定数の協力者を集めるに至った。呼びかけの際には，謝金が支払われることも伝えた。

また，研究に参加しデータ提供をしてくれた学生たちには，彼らの発話を研究目的のために使用することを承諾する「同意書」への理解，署名・捺印を依頼した。参加した全ての学生が，同意書に応じてくれた。

（2）データ収集のための道具

発話データは，英語による 30 分間のインタビューにより収集した。インタビューの構成，使用した道具（絵，ロールプレイカード）は，（株）アルクが米国機関 ACTFL と共同開発した The Standard Speaking Test（SST）で用いられているものを使用した。本研究の実施者は，(株)アルクが提供する SST 面接官，評価者の資格を得るためのセミナーを受講し，本研究におけるインタビューを行った。

（3）データ収集の手順

データ収集は，一人につきおよそ 2 時間をかけた。本研究実施者とデータ提供者のみが個室にいる状態で行った。まずは，データ提供者の緊張を解し，またデータ提供への理解を得てもらうため，挨拶，数分の雑談，研究の主旨説明を行った。その後，30 分間の英語によるインタビューを行った（注：通常の SST はおよそ 15 分間だが，本研究ではより充実した発話収集のため，Stage 2 と Stage 4 のタスクを一つずつ増やし 30 分間をかけた）。英語によるインタビューはその後「沈黙」や

「言いよどみ」,「理解不能な発話」の理由や意図を学習者自らに質問し，その際に「考えていたこと」や「本当は言いたかったこと」を答えてもらうため，ビデオカメラに録画をした。このことはインタビューを始める前にデータ提供者に説明し，了承を得ている。
30 分の英語によるインタビュー後, 録画したビデオをモニターに映しながら日本語による聞き取りを行った。この際は音声のみを録音した。データ提供者には，ビデオを見ながら沈黙時に考えていたこと，英語で上手く言えなかったこと，本当は言いたかったが回避したことを自発的に言ってくれるように依頼した。ただし，データ提供者がビデオに見入ってしまい，なかなか発話がない時は研究実施者の方から質問をした。ビデオの再生は適宜止めながら，およそ 1 時間 15 分ほどの振り返りを行った。
最後に，① データ提供者の英語学習についての情報と，② インタビューを受けての感想を口頭アンケート形式で収集した。これによりおよそ 2 時間のデータ収集を終えた。

（4）データの書き起し（LC-OR の作成）
学習者の英語でのオリジナル発話コーパス（LC-OR）を作成するため，インタビューを録画したビデオを見ながら書き起しを行った。書き起しは発話者のフィラー (filler, 例 : uh，uhmm，errr）やくり返し，言い直し，インタビュー者との発話の重なりなども忠実に再現するよう留意した。沈黙は短い沈黙（2，3 秒程度）と長い沈黙（それ以上）に分けて，該当箇所を特定できるようタグ付与を行った。書き起しやそれに伴う発話タグの付与は，SST を基にした学習者コーパスである，The NICT JLE Corpus の書き起しガイドライン（和泉他，2004）に従った。

（5）データの訂正（LC-CR の作成）
学習者のオリジナルの発話は多くの英語的誤りや不自然さを含んでいる。LC-CR は，そのような「学習者の英語」と「規範的な英語」を比較分析するための参照コーパスの役割を担うものである。
英語の訂正は，まず本研究実施者が行い，その後英語母語話者が確認するという手順で行った。訂正する英語は，語彙的・統語的に規範的な英語と異なるもののみを対象とし，場面において不適切さ（inappropriateness）は対象としていない。英語母語話者に確認を依頼する際，その点に留意してもらった。
また，LC-CR は，LC-OR とは別個のテキストファイルを用い，ファイル名とテキスト内のヘッダー情報により対応関係を把握できるようにした。

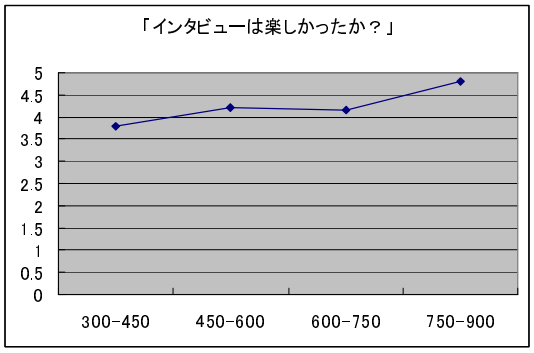
（6）データの翻訳（LC-JP の作成）
学習者のオリジナル発話は多くの沈黙や理解不能な英語を含んでいる。LC-JP は，そのような実際の英語による発話と「学習者の意図」を比較分析するための参照コーパスの役割を担うものである。
日本語への翻訳は，本研究実施者が行った。コーパスファイルごと（データ提供者ごと）の翻訳のズレを極力避けるため，同じ SST タスクで使用される単語は，全て同じ日本語に翻訳するなどの配慮を行った（例 : man/guy/boy → 男性，music player/audio player/compo/audio set → オーディオ）。また，発話の語尾は全て「～です」「～ます」に統一した。その他，助詞・助動詞を省略しない，発話者が男性であっても I は全て「私」と翻訳するなど，一定のガイドラインに従った。ガイドラインは翻訳作業を行いつつテキスト化し，既に作業を終えたものに対しても不統一が生じないよう，くり返し見直し・訂正を行った。
また，沈黙時に意図していたことや，「英語ではこう言ったが，本当はこういうことを言いたかった。だが，英語での言い方がわからなかったのでこのような英語を言ってしまった」という「回避」については，コーパスファイル内に適宜記述をし，該当箇所には新規の<int>...</int>というタグを付与した。

（7）付加情報（言語タグ）の付与
英語コーパスである LC-OR，LC-CR には品詞タグを付与した。英国ランカスター大学のコーパス研究機関（UCREL）によって開発された自動品詞タグ付与プログラム CLAWS を用い，詳細な品詞情報を与えてくれる C7 版のタグを付与した。

４．研究成果

（1）提供者によるデータ収集の感想
本研究においては，学習者が英語発話時に「意図していること」を引き出すことが重要な作業となる。しかしながら一般的に，学習者は自らの発話の内容やその時点で意識していたことを内省することを得意としていない。これは英語力が低い学習者ほど顕著な自己モニターに対する困難だといえる。そこで本研究では，データ収集時に① 研究主旨を説明し,「英語が苦手でも全く構わない」ということを伝えリラックスできる状態を作り，さらに② 英語でのインタビュー直後に録画したビデオを見ることで，その時点で考えていたことを思い出し易くするという，データ収集の工夫を行った。これら方法が効果的に機能したかどうかを，研究初年度（平成 18 年度）のデータ提供者へのアンケート調査によって知ることができる。
全てのデータ収集終了後に，学習者の英語学

習状況と本研究におけるインタビューに対する感想を尋ねたところ，「今回のインタビューは楽しかったですか？5 段階評価で表わしてください」という質問に対し，平均して



4.24 の高い（楽しかった）という回答を得た。これを提供者の英語力（直近の TOEIC のスコア）別に見てみたところ，以下図 1 のようになった。

図 1　TOEIC スコア別インタビューの楽しさへの評価（max. 5）

英語力が低い学習者の数値は若干低いものの，彼らの多くは「楽しかったけど，上手く話せなくて大変だった」というものであった。一般的に学習者の発話を収集することは，提供者に心的負担をかけると懸念されがちだが，ここで得られた結果を見る限りは，30 分間の英語でのインタビューとそれを録画したビデオを鑑賞することは，十分に実行可能だということがわかった。

（2）各コーパスの概要

以下表 1 は，3 つのサブコーパスの総語数（LC-JP の場合は総字数）を示したものである。英語の訂正を加えた LC-CR は，LC-OR よりも総語数が増えているが，これは LC-OR に見られる不完全なセンテンスに対して，単語等を補うことによってそれらを修正したためだと思われる。また，LC-JP の文字数には，学習者が沈黙時や別の発話中に意図していた発話である<int>…</int>箇所は含まれていない。LC-OR と LC-CR に対応する箇所のみの結果となっている。

	LC-OR	LC-CR	LC-JP
総語数	130,671語	138,294語	182,939字

表 1　各サブコーパスの語数情報

この総語数から換算すると，データ提供者一人あたりの平均総語数は 2,613.42 語となり，1 分間あたりは 87.114 語話していることになる。これは，例えば The NICT JLE の Level 4（中級レベル）の学習者の平均発話が 1,542.1 語であるのに対し，1.5 倍以上の語数となっている。これは，インタビューにおけるタスクの数を 5 つから 8 つに増やし，時間も 15 分から 30 分に延長したための当然の結果であるが，一人の学習者からより充実した発話を得る結果となった。

（3）今後行う分析の予定

本研究の開始当初の目的は，大別すると 2 つあった。一つはパラレルコーパスの作成で，もう一つは語彙的・統語的分析であった。前者はデータ数こそ多くはないが，作成に際しての困難点・留意点の特定や，ガイドラインの作成も含め，一定の成果を挙げることができた。しかしながら後者については，英語版の 2 つのサブコーパスに対する品詞タグの付与にとどまり，十分な分析には至らなかった（ただし，語彙使用状況については，英単語の習得困難度特定の研究に応用済み：2009 年 6 月論文刊行予定）。この点を踏まえ，今後は作成したコーパスの特徴を生かし，以下の研究を進めて行きたいと考えている。

① 語彙分析：LC-CR において高頻度かつ LC-OR において低頻度の単語は何か＝学習者が使用できない単語の特定

② 語彙分析：LC-JP 内の<int>…</int>内に見られる高頻度単語の特定＝学習者が使用できない単語の特定

③ 統語分析：品詞連鎖について，LC-CR において高頻度かつ LC-OR において低頻度の連鎖は何か＝学習者が使用できない統語構造の特定

④ 統語分析：LC-JP 内の<int>…</int>内に見られる文法事項の特定＝学習者が使用できない統語構造の特定

これら以外にも，学習者コーパスデータの存在により，さまざまな観点からの分析およびそれによる日本人学習者の英語習得における困難点の特徴が探求できることと思う。今後も引き続き，コーパスデータの整備・増強，およびデータの特性を生かした分析を続けていく予定である。

５．主な発表論文等

〔雑誌論文〕（計 1 件）

① 木村恵，「学習者と教材間の距離：学習者コーパスを基礎データとした教材作成の意義」，外国語教育研究，10 号，90-96 頁，2007 年，査読有．

〔学会発表〕（計 1 件）

① 木村恵，「パラレル学習者コーパス構築の試み」，関東甲信越英語教育学会，2007 年 8 月 19 日，千葉商科大学．

〔図書〕（計 1 件）

① 投野由紀夫編著，『日本人中高生一万人の英語コーパス』，38-44 頁，67-87 頁，118-123

頁，2007 年，小学館.

６．研究組織

(1)研究代表者

木村 恵（KIMURA MEGUMI）
獨協大学・外国語学部・専任講師
研究者番号：60409555

(2)研究分担者

なし

(3)連携研究者

なし