第二言語としての日本語の習得研究
第10号, 2007年12月, pp. -.

〔誌上講座〕

# 英語学習者コーパス研究
## —その概要と第二言語習得研究への示唆—

投野 由紀夫（東京外国語大学）

## 1. はじめに

学習者コーパス（learner corpus）は，第２言語（または外国語）学習者の発話・作文等の産出データ（production data）を機械可読（machine-readable）形式で収集した言語資料体である。言語産出データの観察そのものは母語習得では19世紀後半から，第２言語習得に関しては20世紀半ばから行われており，特段珍しいことではない。学習者コーパスの研究が特徴的なのは「コーパス」という言葉が示すとおり，この20年ほどで急速に発展してきたコーパス言語学（corpus linguistics）の分野の方法論をその中心に据えている点にある。

コーパス言語学と第２言語習得研究の融合とは一体どのような研究上の革新を意味するのだろうか。従来の研究手法と異なる点は何で，それによって，どのような理論上・方法論上の進展が期待できるのだろうか。ここでは，それらのポイントを整理しつつ，英語における学習者コーパス研究の現状を紹介し，他言語（特に日本語）への応用可能性を示唆しておきたい。

## 2. コーパス言語学の貢献

コーパス言語学の発展により，言語研究に開かれた新たな可能性に関して簡潔にまとめておく。コーパスは機械可読形式で大量に収集された言語データであるから，コンピューター処理の特性を活かした以下のような分析の手法や観点が得られる。

### 2.1 頻度と分布

コーパスから得られる最も基本的な情報は，自分の知りたい言語特徴（linguistic features）に関する頻度（frequency）と分布（distribution）である。20世紀後半の言語研究の大勢は，Chomskyの影響で，文法性（grammaticality），容認可能性（acceptability）という「ある構造（表現）が文法的に可能か否か」という点を中心に理論構築が行われてきたが，コーパス言語学の発展により「どのような言語形式がどのくらい多いか

（少ないか）」という頻度の視点が新たに注目されるようになって来た[1]。この頻度の視点は，理論構築上も極めて重要な意味を持つ。認知科学全般では経験が認知的表象（cognitive representations）に多大な影響を与えることは広く受け入れられた事実であり，経験する事象の出現確率（probability）によって，生物は生き延びるための行動原則を選び取っていく，という前提に基づく認知モデルの研究が盛んである。言語習得においても，言語事象の頻度は人間の認知システムの言語構造発現（emergence of linguistic structure）の仕組みと関係しているという主張は近年注目を集めている（cf. Bybee & Hopper, 2001）。また言語処理モデルの構築上も，コーパスから得られる語の頻度やコロケーション頻度のような確率情報が処理プロセスに関与しているとする一連の研究がある（cf. Bod, Hay, & Jannedy, 2002）。また生成文法の枠組みの中でも，頻度情報の影響を考える動きが始まっている（REF）。コーパス言語学の発展により頻度と分布情報をより明確に示せるようになり，上記のような言語習得モデルの構築の基礎情報を与えることに貢献している。

## 2.2　言語注釈付け

コーパス言語学が可能にした新たな可能性のもう１つの側面は，言語データを大量に集めるだけでなく，そのデータに言語注釈付け（linguistic annotation）を施すことで，より有益な言語特徴を抽出できるようにしたことである。言語データに注釈付けをするプロセスをタグ付与（tagging）と呼び，現在は英語に関しては品詞のタグ付与（part-of-speech tagging）はほぼ95～97%以上の精度で自動処理が可能である。日本語の場合は分かち書きと品詞同定などを含めた形態素解析（morphological analysis）という工程を踏み，こちらも同様の精度で自動化できるレベルに達している。

さらに，言語注釈は目的に応じて以下のような種類が用いられる。

① 見出し語化（lemmatization）
② 構文解析（syntactic analysis, parsing）
③ 意味注釈づけ（semantic annotation）
④ 照応関係注釈付け（coreference annotation）
⑤ 語用論的注釈付け（pragmatic annotation）
⑥ 文体注釈付け（stylistic annotation）
⑦ エラータグ付与（error tagging）

（cf. McEnery, Xiao, & Tono 2006）

## 2.3　量的分析手法

コーパス言語学の貢献の３点目は，大量の言語データを処理する分析手法を発展さ

せた点にある。頻度を異なるコーパス・サイズ間で比較する相対頻度（relative frequency）の概念や複数コーパス間での分布を数値化する分布測定（dispersion measure），頻度の差の検定をする手法（Log-likelihood など），単語の共起強度を示す指標（MI-score, t-score, log-log, z-score, Dice coefficientなど），といった言語事象の類似度尺度（similarity measure）や関連性尺度（association measure）といわれるものが開発されてきた[2]。

さらに多変量の言語特徴を解析する手法として，多変量解析（multivariate analysis）の適用が盛んな分野でもある。データの分類・集約を行うクラスター分析，コレスポンデンス分析，変数の合成を行う主成分分析，共通因子の特定を目的とした因子分析，さらに因果関係をモデル化する重回帰分析や対数線形分析，などが大量の言語データに適用された場合，今まで我々が想像もしなかったような新たな言語使用のパターンが観察される。こういった手法の可能性が，コーパス言語学的手法で学習者データを見ることの１つの大きな魅力になっている。

## 3. 学習者コーパスのタイプ

英語学習者コーパスはその研究目的に応じていくつかのタイプに分けられる。以下にその主要なものをあげて解説する。

### 3.1 書き言葉 vs. 話し言葉

コーパス言語学全体の分類にも当てはまるが，収集するデータが書き言葉コーパス（written corpus）か，話し言葉コーパス（spoken corpus）か，という違いは最も大きな区分である。現在，世界中で作られているコーパスは圧倒的に書き言葉が多いが，この４,５年で話し言葉コーパスへの関心は飛躍的に高まっており，コンピューターでの音声処理の技術の向上と相俟って，会話データを収集する研究者は増えている。日本人英語学習者のインタビュー・テストのデータをコーパス化した The NICT-JLE Corpus は世界で最大規模（200 万語）の単一学習者のレベル別話し言葉コーパスである。書き言葉では International Corpus of Learner English（ICLE），Japanese EFL Learner Corpus（JEFLL），HKUST Learner Corpusなどが代表的なもの。日本語学習者コーパスに関しても，書き言葉では「日本語学習者による日本語作文とその母語訳との対訳データベース」[3]，話し言葉ではKY Corpus[4]，日本語会話データベース[5]，CHILDESの日本語版などが利用可能であるが，全体的にまだコーパス規模は小さい。

### 3.2 母語固定 vs. 母語変動

対象となる英語学習者の母語に関して，１つの母語に固定するか，複数の母語話者

を対象とするか，でデザインが異なる。前者の場合は，単一母語話者の第２言語習得に関心があるのに対し，後者の場合は母語の違いによる言語習得の影響に関心がある。ICLE は後者の代表的なプロジェクトで，現在 20 近くの異なる母語を持つ大学英語専攻３，４年生の英作文データを収集比較することを目的にデータ構築中である。

### 3.3　英語力レベル固定 vs. 変動

異なる国の英語学習者の比較や，母語の影響に注目したい場合，学習者の英語力レベルを統制することがある。ICLE はこの代表的なもので，被験者に大学３，４年生という外的基準を設けてレベル統制をしている。これに対して，NICT-JLE Corpus や JEFLL Corpus などは「レベル差」を比較変数にしているので，インタビュー・テストのレベル判定の結果（NICT-JLE）や中学・高校の学年（JEFLL）といった基準を設けている。

### 3.4　エラータグ付与の有無

学習者データにエラータグが付与されているかどうかも大きなポイントになる。エラータグには包括的タグセット（generic tagset）を用いて，すべてのエラーを統一的に分類しようとするものと，問題別エラータグ付与（problem-oriented error tagging）を用いる場合とがある。ICLE, NICT-JLE などは前者の包括的エラータグセットを考案し，特に後者は全データの約１割にエラータグ付与を施して公開している（和泉他, 2004）。

### 3.5　研究用 vs. 商用

最後に，コーパスが研究用か商用かの違いがある。世界的に大規模コーパスは商用のものが多く，Cambridge Learner Corpus は現在 2000 万語の規模を誇るが，社内の辞書・教材作成用のみに利用されており公開されていない。Longman Learner's Corpus も商用の 1000 万語規模のコーパスであるが，こちらは社内リソースとして以外に，研究用に有償で利用が可能だ。その他，JEFLL, NICT-JLE, ICLE などは研究用でライセンスと実費で利用が可能。

## 4. 学習者コーパスの主な研究領域

学習者コーパスを用いた研究はさまざまな可能性があるが，主要な研究領域としては以下のような分野がある。具体的な研究成果や実例に関しては，本研究会で講演を行った際に詳しく紹介した。論文集としてはGranger（1998），Granger et al.（2002）にICLE関連を中心とした研究成果がまとめられている。また包括的な文献リストがICLEのwebサイトに掲載されている[6]。

### 4.1 NS vs. NNS の言語使用比較

英語学習者のさまざまな言語使用を母語話者と対比して研究する。単に誤り分析だけではなく，過剰使用（overuse），過少使用（underuse）も重視される。扱う言語構造は多岐にわたるが，特に談話標識，コロケーション，動詞関連（時制・相，態，動詞型）などが中心。

### 4.2 母語の相違による第2言語習得・使用への影響

英語学習者の母語の違いによる習得への影響を探る。これは特に ICLE が目的として強調している研究手法で，Contrastive Interlanguage Analysis（CIA）と呼んでいる。研究手法として興味深いが，本格的な研究は今後を待たねばならない。

### 4.3 第2言語習得研究仮説の大規模データによる検証・追試

学習者コーパスを用いて，既に提唱されている第2言語習得の仮説や知見を検証する試みがある。文法形態素の習得順序（Tono, 2000a），非対格動詞の習得（Oshita, 2000）などがある。

### 4.4 大規模データによる習得プロセスの記述研究

コーパス言語学の多変量解析手法を用いて，異なる学習段階の学習者コーパスからの言語統計を比較分析する。品詞の n-gram（n 個の連鎖）統計の比較（Aarts & Granger, 1998; Tono, 2000b），エラー分析比較（Abe, 2003），動詞の補部構造の習得（Tono, 2004）などがある。

## 5. 学習者コーパスを用いた研究の具体例

ここでは英語学習者コーパスにおける具体的な研究方法を筆者の研究グループの事例を中心に紹介する。

### 5.1 コンコーダンスによる使用例観察

まずもっともオーソドックスな手法として，コーパスから自分が興味を持っている語彙や文法に関する用例を抽出し，それを観察するという方法がある。その際に，コーパスから目的の構造を正確に検索できるように，正規表現などの基礎知識が必要になる。またコーパスそのものがどういう言語注釈（品詞や見出し語情報など）が施されているか，といった情報も検索の際に知っておく必要がある。

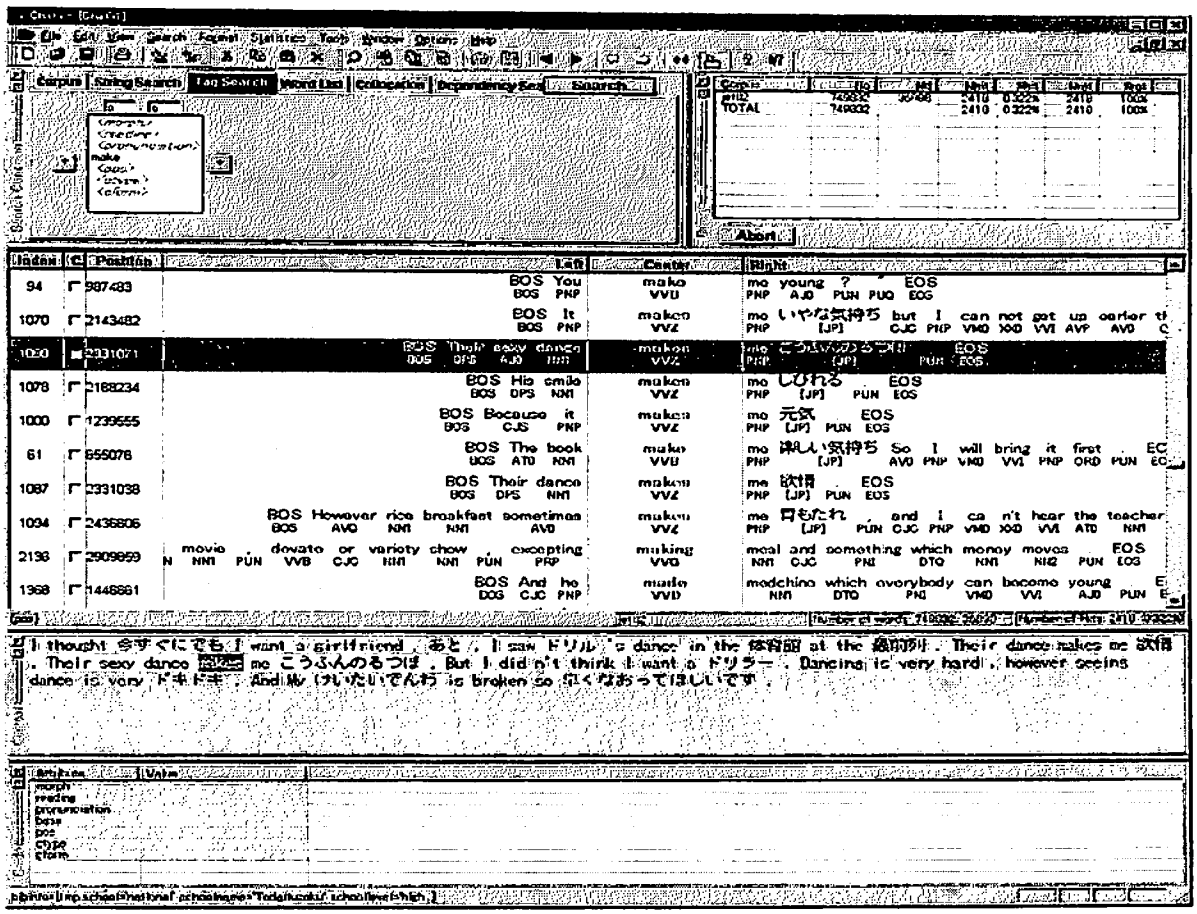


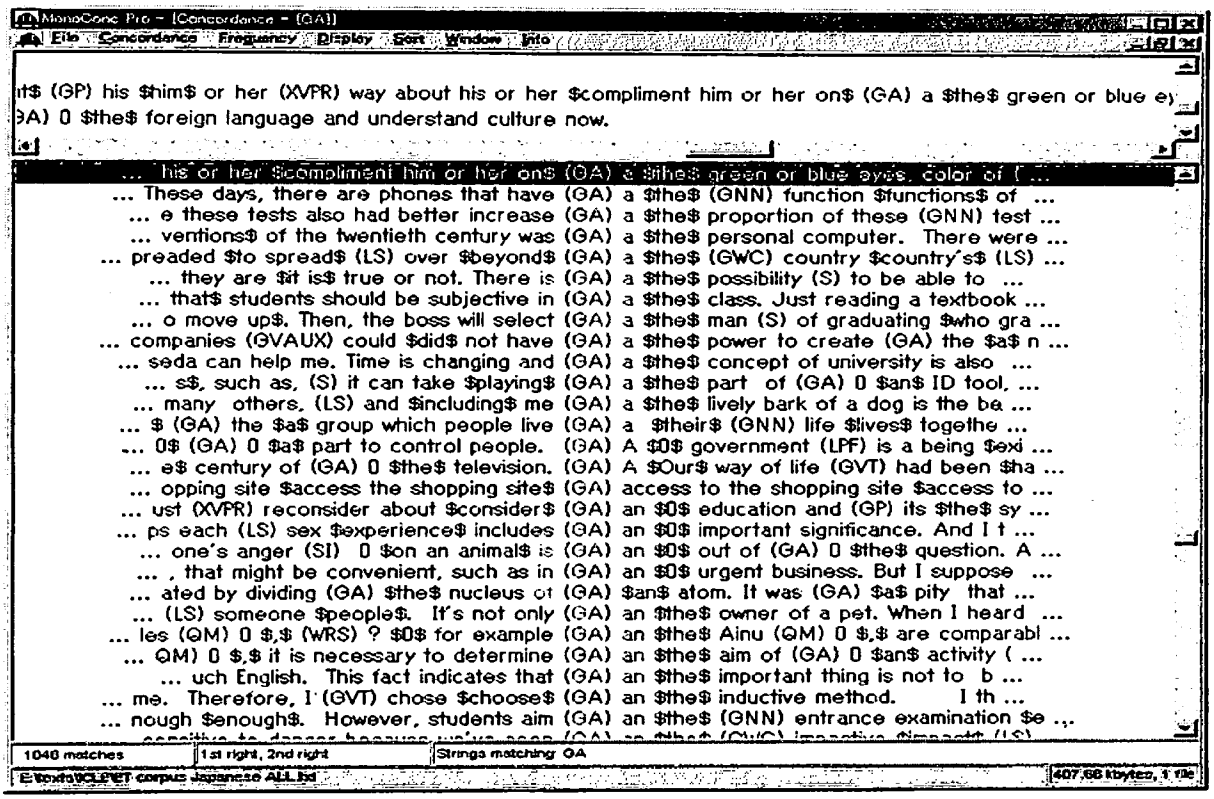
図 1　JEFLL Corpus で make の後の日本語使用の状況を調べたもの（ツール：茶器）

図 1 は，投野が中心となって構築している中高生 1 万人以上の英作文コーパス JEFLL Corpus を ChaKi（奈良先端科学技術大学院大学松本研究室を中心に開発されている言語注釈付きコーパス検索ツール）で検索したものである。

JEFLL は英作文中にどうしてもわからない表現があれば，日本語の使用を許している。動詞 make の後に続く日本語のパターンを見ることで，make の項構造の習得でどのような困難点があるのか，といった点がわかってくる。

次ページの図 2 は ICLE という国際的な学習者コーパス・プロジェクトの日本人データのエラー検索画面である。(GA) は 文法エラー（Grammar）-冠詞（Article）という意味で，学習者データ中の冠詞エラーを一貫して取り出せるようになっている。このようなエラー情報は人手によってタグ付与を行ったうえで検索することになるが，学習者コーパスに特有の貴重な分析方法の 1 つである。コンコーダンスで文字列としてエラータグを抽出することによって，その文法エラーの生起している頻度や環境を調査することができる。

図2　ICLE の冠詞エラーの検索例（ツール：MonoConc Pro）




## 5.2　コロケーション分析

語と語の共起関係の分析も重要な研究分野である。一般的なコーパス分析ツールはたいてい検索語を対象とした共起関係を抽出する手法を実装している。表 1 は JEFLL Corpus（日本人中高生）と ICLE 全体（世界 10 カ国余の大学生英語学習者）の，動詞 make の後の名詞の共起語の上位 5 位までの比較である。

表 1　make + 名詞のコロケーション上位 5 語の比較

| | JEFLL | ICLE-all |
|---|---|---|
| 1 | money | use |
| 2 | foods | money |
| 3 | breakfast | a difference |
| 4 | friends | decisions |
| 5 | movies | an effort |

日本人中高生のコロケーションにはほとんどが具体的な物事を表す名詞が来ているのに対して，世界の大学生レベルの上級英語学習者は make use of, make a difference, make decisions, make an effort などの make の軽動詞（light verb）の用法を多用している。

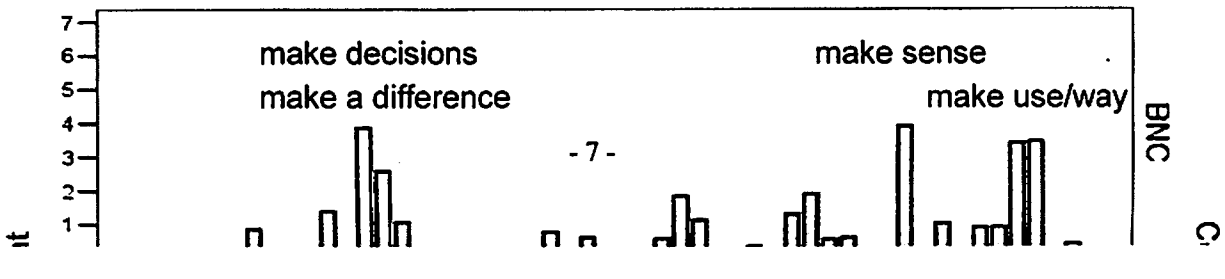


**図3　JEFLL と BNC　の make のコロケーションの出現状況の比較**

図3は同様に JEFLL を母語話者のコーパス（British National Corpus）と比較して，コロケーションの出現度合いを視覚化したものである。図3でも BNC の頻度の高い部分は JEFLL ではコンスタントに頻度が低い。日本人英語学習者がいかに目標言語の自然なコロケーションの獲得が出来ていないか，ということを示している。

さらにコロケーションの強度を示す統計にはいろいろな種類がある。単純頻度，相対頻度，T-score, MI-score, Z-score, Log-log score, Dice coefficient など。内山他（2004）の研究により，これらのコロケーション統計には統計量間に特性があり，たとえば比較的コーパス中での生起頻度が高いコロケーションを取ってくるためには T-score などがよいが，結びつきの強度を重視して低頻度のものでも特に結び付きの強さをもとに抽出したい場合には MI-score の方が適している。さらに T-score と MI-score のそれぞれの欠点を補正して中間的な特徴共起語を抽出しようという場合には log-log score などが好まれる。

### 5.3　相対頻度比較

コーパス全体から語彙リストを作成し，そのリストを相互に比較することもよく行われる分析手法の1つである。例えば，学習者データをレベル別や習得段階別で比較したり，学習者データと母語話者データを比較したり，といったケースである。

この際に相対頻度の比較を行うために対数尤度比（Log-likelihood coefficient）がよく利用される。かつては頻度の有意差検定には z 検定やカイ2乗検定がよく用いられた

が，Dunning（1993）で指摘されたように，コーパス・サイズが小さく低頻度語の検定を行う場合にはｚ検定やカイ２乗検定の条件を満たさない場合が多い。その際に，対数尤度比を利用すれば，それらの条件の制約を受けずに頻度検定が可能になる。表 2 に対数尤度比（通称 Dunning G2）の式を示す。

**表２　対数尤度比 G2 の算出方法**

| | コーパス 1 | コーパス 2 | 合計 |
|---|---|---|---|
| 調査対象の単語の頻度 | a | b | a+b |
| それ以外の語の総頻度 | c-a | d-b | c+d-a-b |
| コーパス全体の頻度 | c | d | c+d |

期待度数の計算：E1 = c*(a+b) / (c+d) ；E2 = d*(a+b) / (c+d)

G2 = 2*((a*LN (a/E1)) + (b*LN (b/E2)))　　※ LN(x) =x の自然対数

これにより例えばレベルの異なる学習者グループを比較して，特徴語の抽出を行うことが出来る。表３は会話コーパスNICT JLE Corpusのレベル１，５，９の語彙頻度リストをコーパス全体と比較して対数尤度比を産出したものである。レベル１ではほとんどの単語は日本語のフィラーが中心だが，レベル５になるとsoなどが表現れ，レベル９ではem, urm, wellなど英語らしいフィラーになる。レベル５では冠詞のtheが現れているが，冠詞の出現は中位から上位にかけてコンスタントに特徴語として現れる。さらにof, toといった前置詞のパターンがレベル５には特徴的に出てきているが，特にto不定詞の使用などが発話中に有意に多くなってくる。レベル９ではI think that...

**表３　NICT JLE Corpus のレベル別特徴語（G2= 対数尤度比）**

| レベル１ | G2 | レベル５ | G2 | レベル９ | G2 |
|---|---|---|---|---|---|
| MMMM | 106.53 | AHH | 261.78 | EM | 397.87 |
| UMM | 62.58 | SO | 200.33 | WAS | 157.07 |
| WAKANNAI (J) | 54.26 | UHH | 102.58 | URM | 152.27 |
| EETO (J) | 53.28 | AH | 91.38 | REALLY | 93.21 |
| EETTO (J) | 47.84 | UH | 70.91 | THE | 92.73 |
| EH (J) | 44.34 | OF | 59.28 | THAT | 92.27 |
| HAI (J) | 41.73 | FOR | 50.42 | WELL | 77.03 |
| MMM | 27.17 | TO | 48.77 | YOU | 75.12 |
| SUIMASEN (J) | 26.25 | THE | 41.01 | POLICE | 69.96 |
| | | SOME | 39.32 | JUST | 61.77 |
| | | PARTY | 36.8 | WOULD | 49.42 |
| | | ALSO | 32.76 | GET | 49.15 |

のような従属節の that，法助動詞の would などが現れる。このように，単純な頻度表

を見ただけでは発見できないような，サブコーパスの持つ語彙的特徴を捉えるような方法として対数尤度比は有効である。

### 5.4　多変量解析を用いたコーパス分析

コーパスからの頻度をいろいろな言語特徴に関して集計した場合，それらを何らかの形で整理したり分類したりしたい，というニーズがある。多変量解析はこのようなデータの整理統合といった用途に適した手法が用意されている。その代表的なものは主成分分析，因子分析，コレスポンデンス分析などである。

次ページの図4はNICT JLE Corpusの9段階のレベル別の語彙頻度表から上位100語を抽出して，それらの頻度変化をもとにコレスポンデンス分析を行ったものである。行プロットと列プロットを重ね合わせてみると，レベル別にどのような単語が特徴的に出現しているかを視覚的に把握することが出来る。またレベルもDimension 1 に沿って，右から左へレベル別に配置されていることから，~~あるいは~~100 語程度の高頻度語の使いこなしで，レベルの差が現れることを示している。

コレスポンデンス分析は 1990 年代~~台~~からコーパス言語学の論文でも使われ始め，学習者コーパスに最初に適用したのは，Tono（1999）であるが，その後同様の手法を用いて投野の研究グループがさまざまな記述的研究をしている。例えばKaneko (2006) では，名詞句の構造に関してNICT JLE Corpusのレベル別特徴をコレスポンデンス分析により分析している。◉ページ図5にそのプロット図を示す。名詞句のタイプとしては単純な名詞のみ（N），冠詞＋名詞（art/det+N），数詞/所有格＋名詞（num/pos+N），（副詞）＋形容詞＋名詞（adj+N），名詞＋前置詞句（N+PP），名詞＋従属節（N+clause）に分れる。Dimension 1 に沿って，IL → IM → IH → A → NS とレベル別になっており，またそれに対応してどのような構造が来ているかを視覚~~資格~~化してみることができ，興味深い。特にIL（中の下）レベルでは名詞単独の用法（Nのみ）が，上級レベル（A）では N+PPやN+clause のような複雑な構造が近い関係にあるのが見て取れる。

Abe and Tono（2005）ではJEFLL，NICT JLEという2つの学習者コーパスのデータにエラータグを付与して，そのエラータグと2つのコーパスの分布関係をコレスポンデンス分析で見ている。◉ページ図6にその概略をプロット図で示す。1軸に沿って書き言葉のJEFLL（J1～S3までのレベル）が左側に，話し言葉のNICT JLE（SST2/3～SST8/9）が右側に配置されている。さらに2軸に沿って，上側に初級者グループが下側に上級者グループが配置されている。言語特徴を見ると，上側が動詞関連のエラーが多く，下側が名詞関連のエラーが多い。これによって，大まかなエラー傾向としては会話でも作文でも初級～中級にかけて動詞関係のエラー（時制や数の一致などの形態的エラーが中心）が頻発し，それが上級になるにつれて，より名詞関係のエラーの割合が顕著になる，ということがわかる。

このようにさまざまな言語特徴やエラー傾向と学習者レベルの関係を見ていくことで，学習レベル別の文法や語彙の使用状況の分析が可能になる。これらを蓄積していくことで，ある程度の学習プロセスの客観的な記述が可能になることが期待される。

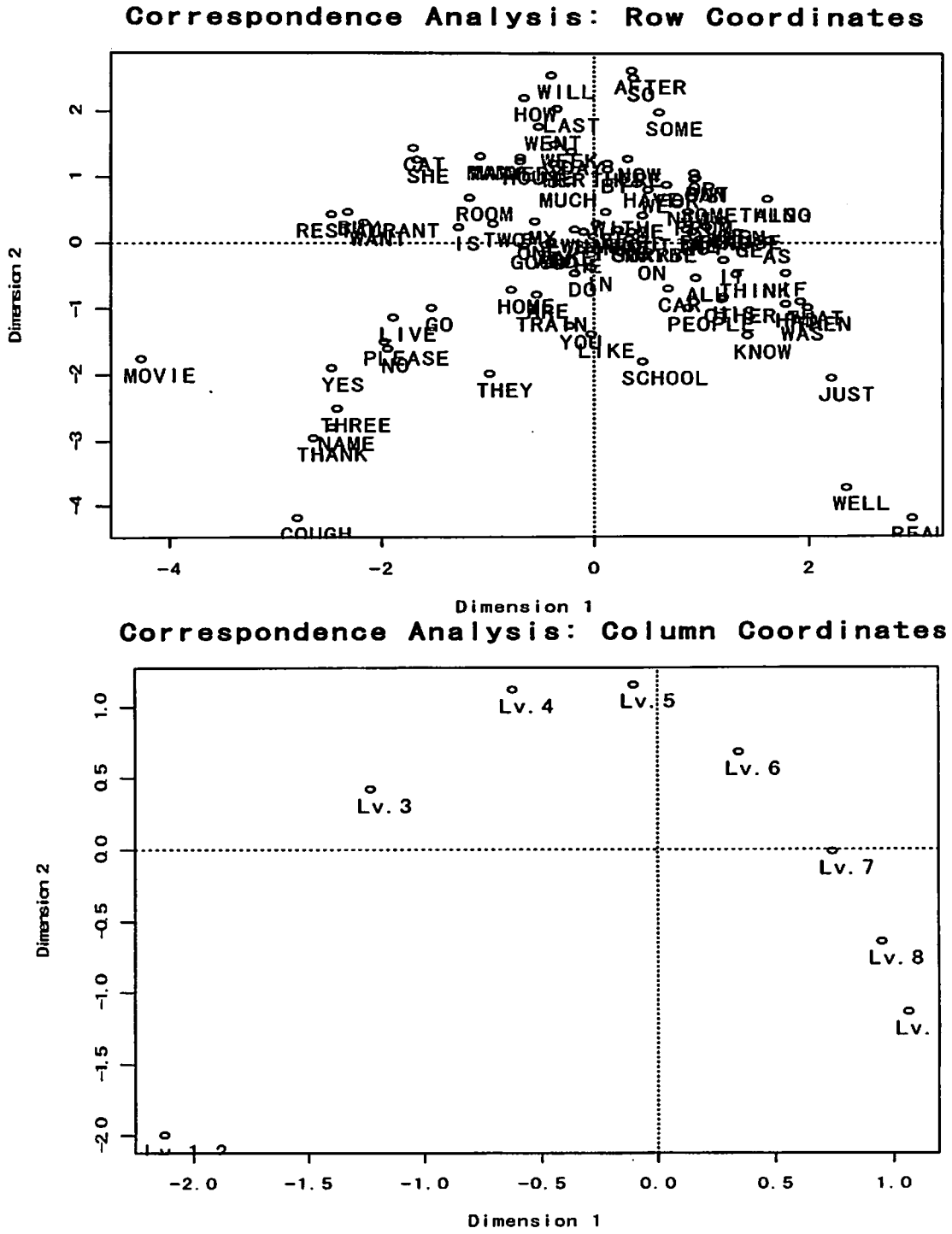



図4　NICT JLE Corpus の各９レベルの上位 100 語のコレスポンデンス分析

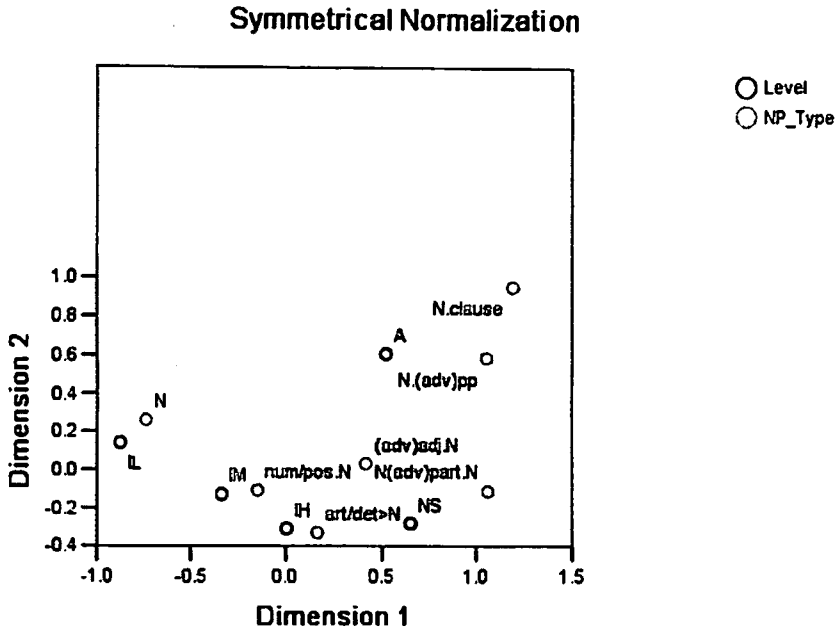



図 5　NICT JLE Corpus のレベルと名詞句構造の関係（Kaneko 2006）

注：IL=Intermediate Low; IM=Intermediate Middle; IH=intermediate High; A=Advanced; NS=Native Speaker

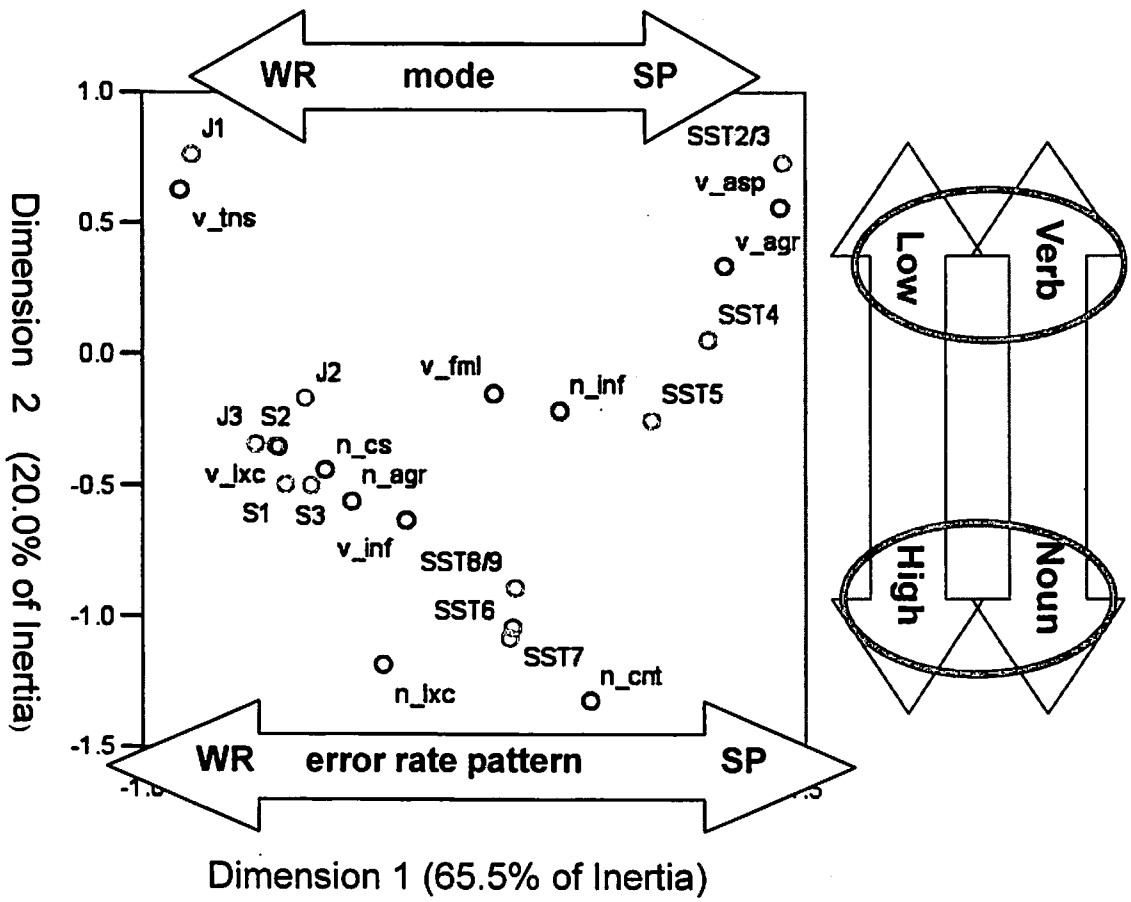



図 6　NICT JLE および JEFLL におけるエラー分布の比較（Abe & Tono 2005）

## 5.5 エラータグを用いたエラー傾向の総合的分析

他のコーパスと比べたときの学習者コーパスの最大の特徴は，コーパス・データにパフォーマンス上のエラーが混入しているということである。それはいわゆる母語話者でもやる言い誤り（mistake）ではなく，中間段階の習得途上の文法システムの発現としての組織的な誤り（error）である。これらエラーの特徴と頻度を正用法と共に組織的に記述していくことで，中間言語の文法体系の推移が記述できる可能性がある。

表4はCambridge Learner Corpusにおける，日本人英語学習者コーパス40万語にみられるエラー頻度と英語学習者全般（600万語）に観察されたエラー頻度を順位で比較したものである。比較してまず明らかなことは，日本人英語学習者の最も苦手な項目は「冠詞」であり，この脱落エラーに関しては一般の英語学習者全般に比しても極めてエラー頻度が高い，ということである。さらに第9位の「名詞の単複」の間違いも英語学習者全般（第15位）に比して順位が高い。これも名詞の複数形などの形が日本語にはないことと関係がありそうである。その他の傾向は比較的，両者共に似通っていることを思うと，動詞がらみのエラーは比較的世界的な傾向として認められ，名詞関連のエラーに関しては母語からの干渉が特に顕著である，と言えそうである。

さらに詳しく冠詞エラーについてみると，冠詞の脱落エラーは学習者全般が40,644回（1,000万語で正規化）エラーがあるのに対して，日本人英語学習者は85,621回。冠詞の選択エラーに関しては，一般が9740回に対して日本人は19,684回，と極めて

**表4　日本人英語学習者特有の誤り傾向について**

| 日本人（CLC　40万語） | 英語学習者全般（CLC 600万語） |
|---|---|
| 1. **冠詞の脱落** | 1. スペリング・ミス |
| 2. スペリング・ミス | 2. 句読点の脱落 |
| 3. 句読点の脱落 | 3. 前置詞の選択 |
| 4. 前置詞の選択 | 4. 動詞の選択 |
| 5. 動詞の選択 | 5. 句読点の選択 |
| 6. 動詞の時制 | 6. 動詞の時制 |
| 7. 句読点の選択 | 7. **冠詞の脱落** |
| 8. その他の語彙の選択 | 8. 名詞の選択 |
| 9. **名詞の単複** | 9. その他の語彙の選択 |
| 10. 前置詞の脱落 | 10. 語順 |
| 11. 名詞の選択 | 11. 動詞の相 |
| 12. 冠詞の余剰 | 12. 冠詞の余剰 |
| 13. 語順 | 13. 前置詞の脱落 |
| 14. 前置詞の余剰 | 14. 前置詞の余剰 |
| 15. 動詞の相 | 15. **名詞の単複** |

多いことがわかる。こういった観察が可能なのも，コーパス・データに冠詞エラーのタグが組織的に付与されていることによる。

Tono (2000a) は，このようなエラータグを用いて第２言語習得研究で有名な文法形態素の習得順序研究を学習者コーパスによって追試を行った。Burt, Dulay, and Krashen (1982) の一連の研究を最終的にまとめた普遍的習得順序に対して，JEFLL Corpus にエラータグ付与を行い，それらの集計を Bilingual Syntax Measure の計算方法と同様に行った結果，JEFLL Corpus では冠詞の the/a が最も遅く，代わりに Burt, Dulay, and Krashen (1982) で最も遅く習得される項目の１つである所有格に関してはかなり早く獲得されることが確認された。これは，一連の日本人英語学習者を用いた他の研究（Hakuta，白畑，冨田など）と同様の研究結果であり，普遍的習得順序という考え方に対して部分的に疑義を呈するものであるといえよう。

### 5.6　「英語にしにくい表現」の分析

特殊なアノテーションを用いた分析例として，JEFLL Corpus にある英作文中に使った日本語の分析を紹介する。JEFLL Corpus は 20 分間辞書なしの自由英作文課題であるが，中学１年生から高校３年生まで書かせるために，どうしても英語が出てこない場合の日本語使用を認めている。その日本語部分に関して集中的に分析を加えることで，「英語にしにくい表現」の分析が可能になる。次ページの図 7 は JEFLL の日本語部分を抽出した例である。

清水 (2007) は日本語部分の語彙分析を行い，5.3.で紹介した対数尤度比をもとに表 5 のような各レベルの日本語の特徴語抽出を行った。各学年を比較すると，高３レベルではほとんど名詞の変換のみが問題なのに対して，中学１年では「だから，や」といった接続詞系，「ない，も」(副詞系)，「で」(前置詞系) といった機能語群が大量に現れている。中１レベルでは文法の骨格がまだ出来ていないためにこのような文の骨組みを構成する概念を英語でどう表現するかが十分にわかっていないことが観察される。

このような分析を深めていくことで，初期の中間言語の文法体系がどのような特徴を有しており，それにどういった指導を行えば習得を助けることが出来るのか，といった興味深い仮説へと導いてくれる可能性がある。

## 6.　まとめ

本稿では現在注目を集めている学習者コーパス研究について，英語の分野における利用できる学習者コーパスの種類，構築方法，具体的な研究方法や事例に関して，筆者の研究グループの活動を中心として最新動向の解説を行った。

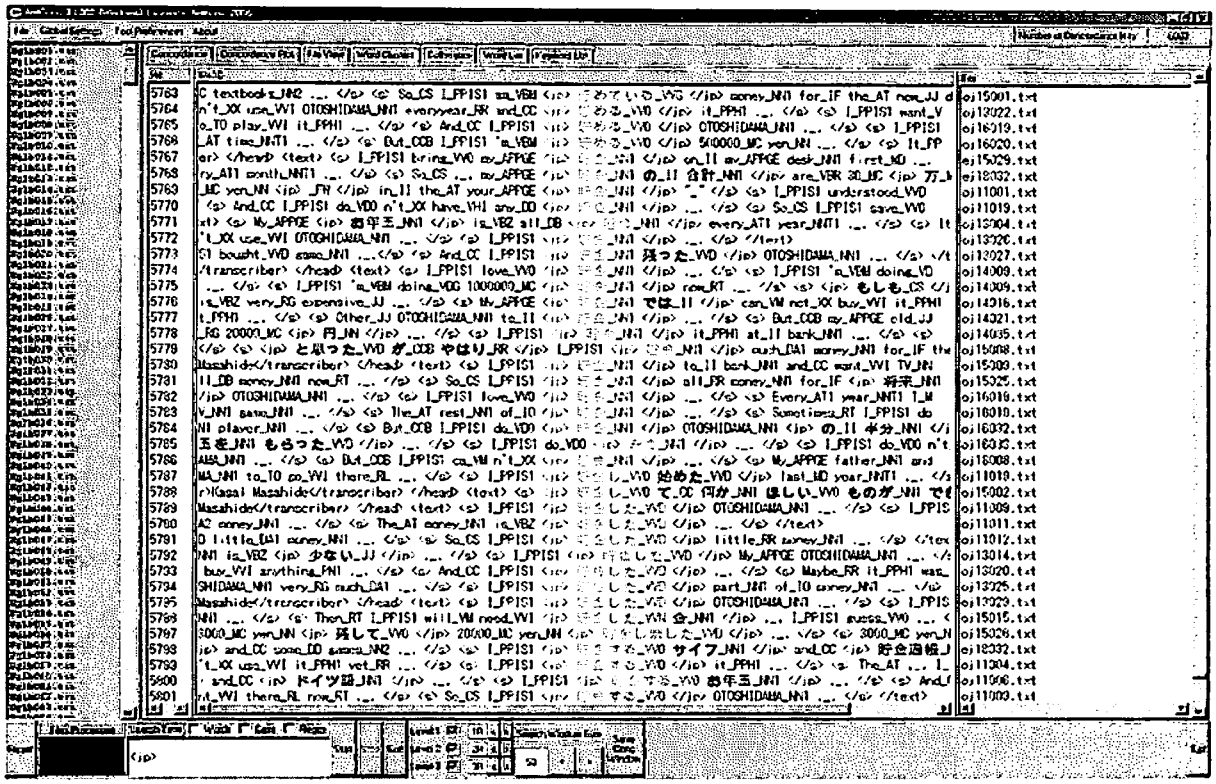


図7　JEFLL Corpus 中の日本語部分の抽出（ツール：AntConc）

表5　対数尤度比を用いた JEFLL Corpus ３学年の特徴的な日本語

| | 中学１年 | 中学３年 | 高校３年 |
|---|---|---|---|
| 1 | だから（接続詞） | 亀（名詞） | お年玉（名詞） |
| 2 | 高い（形容詞） | の（前置詞） | 焼きそば（名詞） |
| 3 | 貯金（名詞） | 劇（名詞） | もち（名詞） |
| 4 | だった（動詞） | 怖い（形容詞） | 出し物（名詞） |
| 5 | ない（副詞） | する（動詞） | お化け屋敷（名詞） |
| 6 | も（副詞） | 手帳（名詞） | ダンボール（名詞） |
| 7 | テーマ（名詞） | 言った（動詞） | おにぎり（名詞） |
| 8 | おいしい（形容詞） | その（決定詞） | 体育館（名詞） |
| 9 | で（前置詞） | 昔（名詞） | シリアル（名詞） |
| 10 | や（接続詞） | 仙人（名詞） | うどん（名詞） |

（注：カッコ内の品詞はそれを英語に変換した場合の該当する品詞を充てている）

学習者コーパスは基本的には習得データであるので，言語習得の知見や仮説をもとに分析の観点を決めていくことになる。よって言語習得理論の十分な理解が必要になる。同時に，コーパス・データとして処理する際の方法論や統計処理などの技術的なことを熟知することで，より分析の切り口も鋭くなる。自然言語処理の技術は日進月歩で日本の研究水準も極めて高い。そういった工学系の分野との連携も徐々に始まっ

ている。学習者コーパスの構築が進み，日本語学習者の習得の実態に関しても，画期的な研究の前進が見られるよう期待しつつ，今後も英語の分野との研究交流を密にしていっていただきたいと願うものである。

**注**

1. Chomsky 以前の構造言語学では，言語データの精密な観察と記述の洗練された手法が培われてきたが，大量の言語データから頻度を算出する観点が弱かった。また生成文法の枠組みでも有標性（markedness）の観点は頻度と関係している。
2. これらの多くは，自然言語処理の分野において考案されたものが多く，コーパス言語学は言語処理上の方法論は自然言語処理から恩恵を得ている。自然言語処理の詳細は長尾（1999）参照。
3. 国立国語研究所日本語教育基盤情報センター評価基準グループの web ページで詳細を公開（http://www2.kokken.go.jp/~smudr/public/sakubun/）
4. 詳細は http://opi.jp/shiryo/ky_corp.html 参照。
5. 北九州市立大学上村研究室（http://www.env.kitakyu-u.ac.jp/corpus/）
6. http://cecl.fltr.ucl.ac.be/learner%20corpus%20bibliography.html

**参考文献**

和泉絵美・内元清貴・井佐原均編著. (2004).『日本人 1200 人の英語スピーキングコーパス』 東京: アルク.

内山将夫・中條清美・山本英子・井佐原均 (2004).「英語教育のための分野特徴単語の選定尺度の比較」『言語処理』484: 165-197.

清水伸一 (2007).「英語になりにくい日本語（語彙レベル）」投野由紀夫編著『日本人中高生 1 万人の英語コーパス—中高生が書く英文の実態とその分析—』 東京: 小学館, pp. 33- 36.

長尾真編 (1996).『岩波講座ソフトウェア科学 自然言語処理』 東京: 岩波書店.

Aarts, J., & Granger, S. (1998). Tag sequences in learner corpora: A key to interlanguage grammar and discourse. In S. Granger (Ed.), *Learner English on computer* (pp. 132-141). London/New York: Addison-Wesley Longman.

Abe M. (2003). A corpus-based contrastive analysis of spoken and written learner corpora: The case of Japanese-speaking learners of English. In D. Archer, P. Rayson, A. Wilson, & T. McEnery (Eds.), *Proceedings of the Corpus Linguistics 2003 Conference (CL 2003), Technical papers, 16,* 1-9. Lancaster University: University Centre for Computer Corpus Research on Language.

Burt, M, Dulay, H., & Krashen, S. (1982). *Language two.* New York. Oxford University Press.

Bybee, J., & Hopper, P. (Eds.) (2001). *Frequency and the emergence of linguistic structure.* Amsterdam: John Benjamins.

Granger, S., Hung, J., & Petch-Tyson, S. (Eds.) (2002). *Computer learner corpora, second language acquisition and foreign language teaching.* Amsterdam: John Benjamins.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book.* London: Routledge.

Oshita, Y. (2000). What is happened may not be what appears to be happening: A corpus study of 'passive' unaccusatives in L2 English. *Second Language Research, 16 (4)*, 293-324.

Tono, Y. (2000a). A computer learner corpus-based analysis of the acquisition order of English grammatical morphemes. In L. Burnard, & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective: Papers from the third international conference on teaching and language corpora* (pp. 123-132). Frankfurt am Main: Peter Lang.

Tono Y. (2000b). A corpus-based analysis of interlanguage development: Analysing part-of-speech sequences of EFL learner corpora. In B. Lewandowska-Tomaszczyk, & P. J. Melia (Eds.), *PALC'99: Practical applications in language corpora* (pp. 323-340). Frankfurt am Main: Peter Lang.

Tono Y. (2004). Multiple comparisons of IL, L1 and TL corpora: The case of L2 acquisition of verb subcategorization patterns by Japanese learners of English. In G. Aston, S. Bernardini, & D. Stewart (Eds.), *Corpora and language learners* (pp. 45-66). Amsterdam/ Philadelphia: Benjamins.

原稿受付: 2007.10. 8
掲載決定: 2007.10.15