明海大学大学院応用言語学研究科紀要
応用言語学研究 No.6 抜刷
(2004年3月)

第6回明海大学大学院応用言語学研究科セミナー　講演

コーパス言語学の手法を応用した第二言語習得研究

投野　由紀夫

Graduate School of Applied Linguistics
Meikai University

コーパス言語学の手法を応用した第二言語習得研究

投野　由紀夫

1. はじめに

最近言語学の研究方法の一分野としてコーパス言語学が注目を集めてきている。本論はコーパス言語学の手法を第二言語習得研究に応用する可能性に関して論じ、その具体的な分野の１つとして英語学習者コーパス（EFL/ESL learner corpus）の研究動向を概観する。まず学習者コーパスの定義を行いその特徴をまとめ、次に方法論的な視点から第二言語習得研究における学習者コーパスの使用事例を概観する。最後にさらに具体的な研究例として、筆者の動詞下位範疇化情報の獲得に関する研究を紹介する。

2. 学習者コーパスとは

2. 1. コーパス全体の中での位置づけ

学習者コーパスは「外国語学習者によって産出された言語のコーパス（コンピューター・テキスト・データベース）」（Leech（1998））である。通例、現代のコーパス言語学では「コーパス」はある特定の研究目的の下に、コーパス・サイズ（corpus size）、テキスト標本抽出（sampling）、代表性（representativeness）といった一定のコーパス設計計画（design scheme）に従って収集されたテキストの集合体を言う。

学習者コーパスはコーパスのタイプ的には特殊コーパス（specialized corpora）の一種である。図１は主要な英語コーパスの種類を概観したものである。図の右左に「一般（general）コーパス」対「特殊（specialized）コーパス」の区分を配した。また上下には「書き言葉（written）」と「話し言葉（spoken）」の軸を設けた。一般コーパスとしては1980年頃までに盛んに開発された100万語規模の英語変種コーパス（例えば Brown, LOB, Frown, FLOB[1]）が有名で、その後いわゆる１億語規模の mega-corpora として Bank of English[2] や British National Corpus[3]（以下 BNC）が作成された。また話し

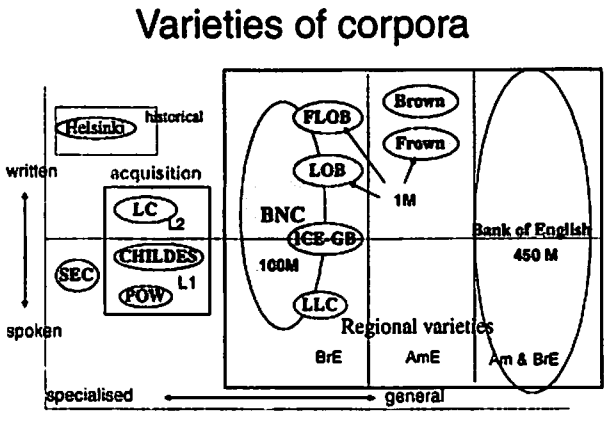

図１　主要英語コーパスの種類

[1] Brown、LOB、Frown、FLOB に関しては ICAME（http://helmer.hit.uib.no/icame/cd/)参照。
[2] Bank of English に関しては COBUILD 公式ページ（http://www.cobuild.collins.co.uk/）参照。
[3] British National Corpus 公式ページ（http://www.natcorp.ox.ac.uk/index.html）参照。

言葉コーパスとしてはICE-GB（半分が話し言葉）およびLondon Lund Corpus（LLC）がSurvey of English Usageの一部として有名である。

特殊コーパス分野では、図1に挙げたように歴史的な英語の変遷を知るための通時コーパス（Helsinki corpusなど）、ESPなどの特定の分野コーパス（PERC, MICASE, BASEなど）、そして言語習得データをコーパス化したものが主流である。特に言語習得データでは、母語の幼児データをコーパス化したもの（代表的なものとしてCHILDES）、そして本論で取り上げている第2言語学習者コーパス（L2 learner corpus）がある。

2. 2. 研究データの性質的に見た学習者コーパスの特徴

学習者コーパスは言語習得研究におけるデータの性質的に見るとどのような特徴があるのだろうか？ Ellis（1994: 670）によれば第二言語習得研究における実証データは次の3種類に分類される：

(1) (a) 言語使用（language use）
(b) メタ言語的判断（metalingual judgment）
(c) 自己報告（self-report）

(1a) の言語使用データはさらに理解（comprehension）と産出（production）に分かれ、またそれらの言語使用の環境設定として、誘引された言語使用（elicited language use）と自然な言語使用（natural language use）とに分けられる。学習者コーパスのデータはこの分類で行けば、自然な言語使用あるいは誘引された言語使用のカテゴリーに入る。言語使用データの長所・短所を他のデータタイプと比較しながら完結にまとめてみると表1のようになる。

	長所	短所
自然な言語使用	他のデータの基盤となりうる	収集が困難；標本が不統一
	自然な習得状況が観察可能	観察数が確保しにくい
	言語使用の諸要因を考慮	コストがかかる
		回避現象を防げない
誘引された言語使用	変数を制御して確実に見る	不自然な使用になりがち
	観察数を十分に確保できる	習得状態の定義が不明確
	コストがかからない	
メタ言語判断	文法知識の確認に適している	判断の信頼性に問題あり
	非文の判断は他ではできない	判断からの類推が困難
自己報告	学習者の内面に迫れる	恣意的である
		どこまで顕在化可能か疑問
		信頼性が低い

表1：言語習得研究におけるデータの種類とその長所・短所

自然な言語使用は可能な限り人工的な設定を廃した自然な言語使用場面を対象としているので、習得の最も自然な、ありのままの状況を観察することが出来る。よって自信のないあるいは不必要な語彙・表現は使わないという過少使用（underuse）や回避（avoidance）の現象が見られるし、正しい用法でも本国人の使用に対して著しく使用頻度が高くなってしまう過剰使用（overuse）のような現象も観察できる。このような利点がある反面、自然な言語使用のデータを採取するには、出来るだけ英語による自然な会話の機会を捉えてデータ収集を行わねばならず、日本の英語教育の現場において生徒同士の自然な会話を採取しようとすると、環境的・物理的・経済的にかなりの時間と労力をかけないと集まらないという現実がある。学習者コーパスはこのような制限の中でも学習者の言語使用のありのままの姿を反映するように学習者データを集めようとする試みであり、さらにその規模的にも70年、80年代のデータ収集よりも組織的かつ大規模に行われる。そういった意味で貴重な言語資料であるといえよう。

Rod Ellisは2001年10月に昭和女子大学で行われた学習者コーパスのワークショップにおいて" Real Data and Real Pedagogy" と題する講演を行い、その中で第二言語習得研究用のデータとしての望ましいコーパス・データとして以下の3種類を挙げている：

(2) (a) Corpora of 'authentic' native speaker language use
(b) Corpora of learner language use
(c) Corpora of native speaker language use with learners (i.e. samples of foreigner talk).

現在、(2a) に関してはさまざまな大規模標準コーパスの構築が行われており、(2b) の分野として学習者コーパスの一連の動きが活発になってきている。(2c) はまだなかなか本格的なものがない[4]。今後、このようなタイプのデータが相互に比較できるようになることで第二言語習得の観察データとしての活用が期待される。

2. 3. 主要な学習者コーパス

学習者コーパスの先駆的な試みは筆者の調査した限り、コペンハーゲン大学で1970年代に行われたPIF（Project In Foreign Language Pedagogy）Corpus（Faerch, Haastrup & Phillipson 1984）が最初である。その後、70年代後半から80年代になってドイツの移民の第二言語習得を調査したZISA Project（Clahsen 1980; Meisel, Clahsen, and Pienemann 1981）がPienemannのMultidimensional Modelなどで話題をさらったが、PIF、ZISA両者ともコーパスとして資料は公開されなかった。

現在のようにデータ共有の発想から公開に踏み切ったものとしては、欧州科学財団（ESF）がサポートしたPerdueらのESF Database（欧州5カ国を対象にその国で移民として学ぶ成人学習者の習得データを採取したもの、詳細はPerdue（1993）参照）が初期のもので、現在世界的に著名になった幼児の母語習得のデータベースCHILDES（MacWhinney 1995）は比較的早く（80年代後半）からデータを自由に交換できる方式を作ったが、第二言語習得

[4] 例外としては少し古いがSinclair & Coulthard（1975）がこの種のデータの最も優れたかつ古典的なもの。

データは最近になって一部含まれるにとどまっている。

以上のデータは主として母語・第二言語習得研究の分野での観察データをコンピューターに入れて再利用するという心理言語学分野の研究者らの試みであったが、反対にコーパス言語学の発展と共に学習者データを対象としたコーパス構築の提案が行われたのもほぼ80年代終わりくらいである。現在、学習者コーパス研究をリードするInternational Corpus of Learner English（ICLE）がスタートしたのもこの頃であった。ICLEはSydney Greenbaumらがロンドン大学のSurvey of English Usage（SEU）という大規模な語法研究の延長で世界の変種英語を収集するというICE（International Corpus of English）プロジェクトの傘下で「学習者英語（learner English）を変種の一種として集める試みが始まりで、その後Louvain大学のGrangerを中心に独立して国際的なプロジェクトとして発展したものである(Greenbaum 1992; Granger 1998)。ICLEは現在15カ国以上の異なる母語を持つ英語学習者（主として大学3・4年生）を対象に1件500語以上の論説文（argumentative essay）を収集している。このように異なる母語を持つ学習者データを集めるのは、特に商用の学習者コーパスで盛んに行われており、例えば、Longman Learner's CorpusやCambridge Learner's Corpusではそれぞれ28から75といった多様な国籍の英語学習者データを1000万語超の規模で集めている（表2参照)。

	公開状況	サイズ (M=100万語)	被験者の英語レベル	L1の数	written/spoken
ICLE	R	2.5 M	advanced	Multi (11)	written
Longman LC	C	10M	all	Multi (28)	written
Cambridge LC	int	15M	all	Multi (75)	written
HKUST	int	10M	advanced	mono	written
JEFLL	R	1M	all	mono	written
SST	R	1M	all	mono	spoken

表2：主要な英語学習者コーパス

（注：R＝研究用に公開、C＝商用、int＝内部利用のみ）

これら異なる母語を持つ学習者データを集める背景には、これらの学習者データを比較分析することで、どの学習者にも共通に起こっているエラー（普遍的エラー：universal error）とある特定の母語を持つ話者にのみ特徴的なエラー（個別的エラー：local error）とを峻別し、その結果から第二言語習得における普遍的要因と個別言語に関する周辺的要因を見分けたいという関心があるからである。ICLEの最初のデータは2002年秋にすでにCDとして公開されており、11カ国の学習者計200万語が利用可能である。

この10年間ほどで整備が着々と進んできているその他の学習者コーパス・プロジェクトとしてHKUST、JEFLL、そしてSSTを紹介しておく。HKUSTは香港科学技術大学のJohn Miltonが収集している中国人大学生の英作文コーパスで1990年代には同一学習者グループのコーパスとしては世界最大規模であった。John Miltonはこれらの作文データの誤り分析を行い、それに基づく文法・作文のweb教材などを開発している（Milton 2001)。

JEFLL、SSTは投野が中心で日本の英語学習者データをコーパス化したもので、JEFLL

は中学1年～高校3年までの英作文データ（辞書を用いず20分間で書く；テーマは共通で6種類の説明文および物語文；文中で英語に表現できない部分は日本語で書いてもよい）を約100万語規模で収集し2005年春までに公開予定、またSSTは(株)アルクが開発した15分間のインタビューによるスピーキング・テストStandard Speaking Test（SST）の音声データ約1200人分を通信放送機構（TAO）の援助のもとに書き起こしたもので、これは2004年夏ごろを目処に公開の予定。この2つのプロジェクトの特徴はICLE、HKUSTなどと異なり、単一の学習者グループを英語力のレベル別にデータ収集を行っている点で、これによりレベルの異なる学習者間の英語使用の違いを科学的に捉え、語彙・構文の発達やエラーの経緯を調べ、中間言語のプロセスの解明を試みようというものである。

3. 学習者コーパスの特色

学習者コーパスは現代のコーパス言語学の成果を学習者データに応用しているが、具体的にどのような点が通常の学習者データと異なるのか、簡単にまとめてみることにしよう。

3. 1. 文書ヘッダ（Document header）

コーパスは一般の利用者が個々の目的に応じて検索したいデータを選択するのが普通である。そのため通例、コーパスの元になっているテキストの書誌情報などをヘッダ部分に詳細に格納する。学習者コーパスは通例このようなコーパス整備の方法論を応用して、データの採取状況や被験者情報に関して精密なヘッダが付与されている（図2参照）。

このヘッダ情報を基に特定のファイルのみを検索対象にして分析を行うことや、ある一定の条件で絞込みや統合が可能になり、コーパスをデータベースのように使うことが出来るようになる。

```
<header>
<textnum>0057</textnum>
<filedesc>
<name>Hanako Yamada</name>
<grade>10</grade>
<proflevel>step pre2</proflevel>
<date>1999-07-10</date>
</filedesc>
<textdesc>
<medium>essay</medium>
<domain>informative</imaginative>
<genre>student writing</genre>
<region>Japanese EFL</region>
<title>my family</title>
</textdesc>
.....
</header>
```

図2：学習者コーパスの文書ヘッダの例

3. 2. テキスト内構造

3. 2. 1. ブロックレベル要素とインライン要素

コーパスはヘッダ部分のようなテキスト外情報（extra-textual information）が充実している

だけでなく、テキスト内情報に構造化したタグを有する。大きくは話者情報や段落・文情報のようなブロックレベル要素（block-level elements）と文の内部に付く品詞タグやエラータグのようなインライン要素（inline elements）とに分けられる。ブロックレベル要素は書き言葉であれば見出し（<h>）、段落（<p>）といった大きなテキストの構造の区切りを指す。図3はSST Corpusのテキスト内構造の実例である。

```
<text>
<body version="2.1.1">
<stage1>
<A><F>Um</F> O K. Thank you for coming.</A>
<B><OL>You're welcome</OL>.</B>
<A><OL>My name is</OL> Hanako Yamada.
      Can I have your name, please?</A>
<B><F>Ur</F> My name is Ichiro Tanaka.</B>
<A><F>Uhu</F> How are you today?</A>
<B>O K. <OL><nvs>laughter</nvs></OL></B>
<A><OL><nvs>laughter</nvs></OL></A>
<B>A little nervous, I mean.</B>
<A><F>Oh</F></A>
</text>
```

図3：テキスト内構造（SSTデータの例）

話者タグ（<A>、<B>）といった単位はブロックレベル要素であるが、それ以外の発話中の重なり箇所（<OL>）、つなぎ語（<F>）、非言語情報（<nvs>）といった基礎情報はインライン要素と見なせる。これらが明示的に付与されていることで、会話の流れをテキスト構造的に再現することが可能になる。またそれだけでなく、このテキスト内の単語1つ1つに関してさらに詳しい言語情報を付与することが出来る。それが以下に紹介する品詞情報、意味情報、および構文解析情報である。

3. 2. 2. 品詞/意味/構文解析情報

テキスト内情報として話者・段落・文などの大きな単位のまとまりを示す以外に、より言語的な注釈（annotation）をインライン要素として付与することが可能だ。このような技術は工学系の自然言語処理の分野で高い関心を持たれているが、もともとはBrown Corpusに自動品詞タグ付与を施した研究がきっかけであり、欧州のコーパス言語学の分野では品詞タグ付与に関して過去30年間さまざまな試みがなされてきた（概要はGarside et al. 1997参照）。

個々の単語に品詞情報を付与することは現在の自然言語処理の技術で95－97％の精度で可能である。これにより同綴異品詞語などを見分けることが可能になるし、見出し語とその活用形のグループ化処理などの基礎データを作ることも可能になる。学習者データの場合は母語話者の英文に比して誤りなどが多く含まれるため品詞自動タグ付与の精度が落ちるという報告がある（van Rooy and Schäfer 2002）。しかし筆者の経験からは修正が困難なほど精度が低いということはなく、むしろ自動でタグ付与を行ってから誤認識パタンを抽出して自動修正するような補正プログラムを書く方がずっと効率的に品詞情報を付加でき

る。図4は学習者データにCLAWSというタグ付与プログラムで品詞を付与したものである。

```
0000020 042 ------------------------------------
0000020 050 I                    93 [PPIS1/100] MC1@/0 ZZ1@/0
0000020 051 'm                   03 VBM
0000020 060 from                 93 II
0000020 070 Hawaii               93 NP1
0000020 071 .                    03 .
0000020 072 ------------------------------------
0000020 080 I                    93 [PPIS1/100] MC1@/0 ZZ1@/0
0000020 081 'm                   03 VBM
0000020 090 junior               93 [JJ/100] NN1@/0
0000020 100 high                 93 [JJ/99] NN1%/1 RR@/0
0000020 110 school               93 [NN1/100] VV0%/0
0000020 120 student              03 NN1
0000020 121 ,                    03 ,
0000021 010 too                  93 [RR@/100] RG/0
0000021 011 .                    03 .
0000021 012 ------------------------------------
```

図4：品詞タグ付与された学習者データ

詳細は省くが、この縦型フォーマットでは第3カラム目の単語列（I, 'm, from, Hawaii,...）に対して、推定される品詞候補がその確率情報つきで第5カラムに列挙される。たとえば最初のIの場合は“[PPIS1/100] MC1@/0 ZZ1@/0”となっているがこれは「人称代名詞（PP）1人称（1）、主格（S）のI（末尾の1がIで2がweを指す）である可能性が100％、その他Iが数詞の1（MC = singular cardinal number）である可能性やZZ1（singular letter of the alphabet）である可能性はゼロである、という意味である。また次の行の'mの場合は、一意にbe動詞のam（VBM）であると決まるので候補は示されない。このような品詞情報が各単語に付与されていれば、後に述べるような品詞連鎖を抽出して構文発達を調べたり、単語＋品詞の文法コロケーション（grammatical collocation）の獲得を見るのに役立つ。また学習者集団の全体像を単に語彙リストで知るだけでなく、使用語彙の品詞リストを見ることで発見できることもあろう。

品詞タグ付与ほど進んではいないが、個々の単語または語句に付与できる情報としては意味情報と構文解析情報がある。意味情報は各単語の意味領域を文脈から推定して自動付与する研究が行われており（たとえばランカスター大学のSEMTAG）、これを応用すると、bankでも「川の土手」の意味と「銀行」の意味を分けて例文検索が出来たりする（図5参照）。また構文解析データの場合は文の木構造のより高次の構成素（例えば名詞句、動詞句、前置詞句など）を特定してそれらの境界をテキスト内に示して検索に利用する（図6参照）。これらの情報があれば、学習者の使用語彙の意味領域に関する分析や構文発達の研究にも利用できる。

```
0000009 082  ----   ----
0000009 090  PPIS1  I          Z8mf
0000009 091  VBM    'm         Z5 A3+
0000009 100  VVGK   going      T1.1.3[i4.2.1
0000009 110  TO     to         T1.1.3[i4.2.2 Z5
0000009 120  VVI    tell       Q2.1 Q2.2 X3 A10+
0000009 130  PPY    you        Z8mf
0000009 140  II     about      Z5
0000009 150  APPGE  my         Z8
0000010 010  NN1    host       S2mf B1@ Y2@ S9%
0000010 020  NN1    family     S4c A4.1c
0000010 021  ,      ,
0000010 030  AT     the        Z5
0000010 040  NN2    Saitos     Z99
0000010 041  .      .
```

図５：意味タグ付与された学習者データ

```
A01:0250j   PPH1   It          it       [S[Ni:s.Ni:s]
A01:0250k   VVDt   urged       urge     [Vd.Vd]
A01:0250m   CST    that        that     [Fn%:o.
A01:0250n   AT     the         the      [Ns:s.
A01:0250p   NNL1c  city        city     .Ns:s]
A01:0250q   YIL    <ldquo>     -        .
A01:0250r   VV0v   +take       take     [V.V]
A01:0260a   NNL2   steps       step     [Np:o.Np:o]
A01:0260b   TO     to          to       [Ti:c[Vi.
A01:0260c   VV0t   remedy      remedy   .Vi]
A01:0260d   YIR    +<rdquo>    -        .
A01:0260e   DD1i   this        this     [Ns:o.
A01:0260f   NN1c   problem     problem  .Ns:o]Ti:c]Fn%:o]S]
A01:0260g   YF     +.          -        .O]
```

図６：構文解析情報が付与されたデータ（Suzanne Corpus の例）

3. 2. 3. エラー情報

学習者コーパスに特有のインライン情報としては学習者の犯す英語の誤りに関する情報が挙げられよう。エラー情報を組織的に付与する試みは HKUST、ICLE、Cambridge LC、SST、JEFLL などで行われてきているが、まだ国際的に統一基準となるエラータグセット (error tagset) は開発されていない。中間言語のエラーは非常に複雑な現象であるので、タグセットの開発そのものの信頼性や妥当性といった問題は避けて通れない。しかし、現在は基本線として以下のような点が研究者相互で合意されつつある：

(a)　形態・統語を中心とする表面上のエラー（surface error）に特化してタグセット

を作るべきで、エラーの原因に関してはデータ採取の状況などが具体的に分かっているものなどに限定するべきである。

(b) 誤り文があった場合、可能な限り局所的エラー（local error）の組み合わせで誤り文を説明できるように原因を切り分けてからタグを付与すべきである。さもないと、全体的なエラー（global error）として誤文訂正をしただけでは具体的にどのような文法・語彙知識が欠如していたのかがコーパスから抽出不可能になってしまう。

(c) エラータグ付与は目視で行うために、統一的に全部のエラーをコーパス全体に付与することは現実的でない。むしろ研究目的に応じて部分的に付与していき、徐々にエラータグ付データを充実させるべきである。

4. 学習者コーパスの第二言語習得研究への貢献

4. 1. 言語使用実態の観察：品詞連鎖の推移

実際に学習者コーパスを用いることでどのような知見が第二言語習得に関して得られるのであろうか？おそらく最も基礎的な学習者コーパスの利用方法は、研究者が関心を持つ言語の諸側面に関して英語学習者の使用実態を観察したい場合に、当該の英語表現や語彙・構文を検索し用例を見る、ということであろう。仮説検証の前にまず言語事実の把握ということを考えれば、大量の学習者データはそれだけで貴重なリソースとなろう。

このような「言語使用実態の把握」という視点で考えると、学習者コーパスからどのような情報が抽出可能かをまとめたのが表3である。

注釈なしコーパスの場合	注釈付コーパスの場合
頻度表（異なり語）	頻度表（見出し語・品詞・構文）
語・句検索（コンコーダンス） 単純な文字列検索＋正規表現[5]	語・句検索（コンコーダンス） 単語＋タグ（品詞）＋正規表現
N-gram 分析 単語＋単語	N-gram 分析 単語＋単語／品詞＋品詞／単語＋品詞
キーワード分析（2つの語彙リスト比較） 特徴語；使用語彙の推移；文体比較	キーワード分析 品詞レベルで複数コーパス比較が可能
構文分析 精密には難しく正規表現検索のみ	構文分析 構文タグがあれば精密な検索集計が可能

表3：学習者コーパスから抽出できる情報

表3を見るとわかるように、言語学的注釈付け（linguistic annotation）がタグの形で付与されているコーパスの場合は、一般の単語情報のみならず、品詞統計、見出し語レベルでの集約などが出来る。また単純な語・句検索だけでなく、n-gram という文字列の連鎖

[5] 正規表現は文書処理、フォーマット変換、検索処理のための技術の名称。通常、コーパス検索ツールには文字列を柔軟に扱えるように正規表現検索が実装されていることが多い。

(cluster) を集計する機能を用いれば、単語の共起パタン（= collocation)、単語と品詞との共起パタン（= colligation）を調べたりすることも可能だ。

このような言語使用実態の観察の一例として、投野が行った英語の幼児の言語習得データと日本の英語学習者（中学1～3年）のデータを比較したものを紹介しよう。英語データはCHILDESの中から選んだBelfast Corpusといわれる2歳から4歳半までの幼児数名の母親との対話を採取した縦断的データ、また日本人は某国立大学附属中学の中学1年から3年生までの各40サンプルに4コマ漫画を見て描写するタスクを課したものである。これらの発話に品詞タグを付与し、それぞれの品詞連鎖を1個の組（モノグラム monogram)、2個組（バイグラム bigram)、3個組（トライグラム trigram）と頻度集計した[6]。

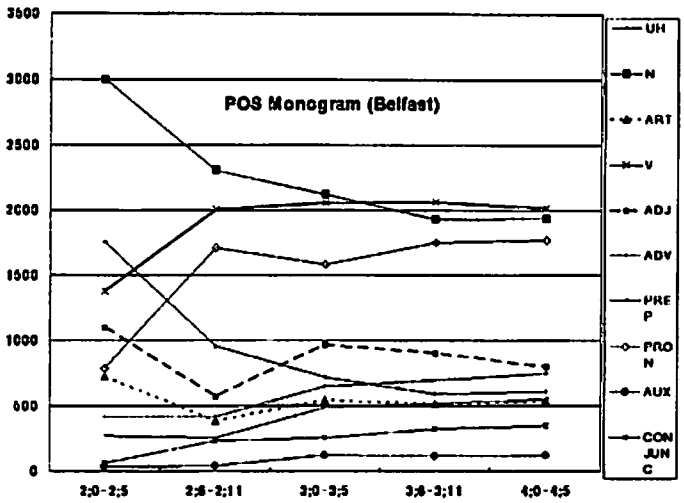

図7: 幼児の英語習得データに見る品詞頻度の推移（CHILDES Belfast Corpus）

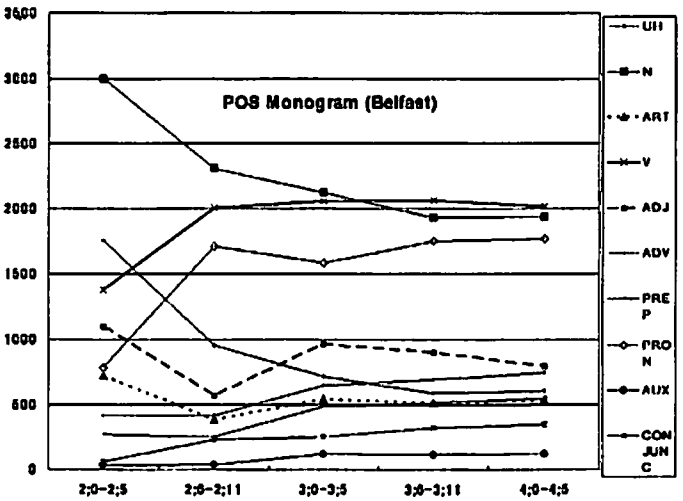

図8: 日本人中学生の英語発話データに見る品詞頻度の推移

[6] ここでは紙面の都合上、トライグラムは省略する。

品詞単体の頻度を示したのが図7および図8である。これを見ると、英語を母語とする幼児の発話の初期（2歳の約1年間）には品詞頻度に大きな揺れが見られ、それが徐々に一定のパタンに落ち着くことがわかる。初期の発話は1語文、2語文に見られるように名詞が中心であり、3歳前後までに動詞および代名詞の頻度が高くなって、文の形成の兆候が見られる。一方で、日本人英語学習者の発話を同様の手法で見てみると、品詞頻度は学習初期からほとんど変化がない。母語の習得と違って教室環境の外国語学習では最初から文の要素や概念が教えられるので、比較的初期から整った文単位の発話を行っていると推測できる。またこれは現実の教室内での会話練習場面で片言の英語で（例えば名詞だけを使って）話す、というような状況がないことを示唆している。すなわち、教室環境での習得は比較的最初から文単位の発話を要求され、そういった指導が発話の特徴にも現れている。

同様の傾向は品詞のバイグラムでも見られる。図9はBelfast Corpusのバイグラムの推移を上位10組に関してみたものである。興味深いことに、品詞2個組の連鎖を見ても2歳から3歳半くらいまでは品詞の組み合わせに一定のパタンがなく、この時期に脳の中で成人の文法に近い知識を身につけていく過渡的な状態が観察される。また最終的に安定した品詞連鎖の頻度を示すようになる3歳半ごろのデータを見ると、最も高頻度で見られたのは「代名詞＋動詞」（PRON+V）の連鎖、その次が「動詞＋代名詞」（V+PRON）で、代名詞使用が本国人の会話データにおける重要な特徴であることが分かる。

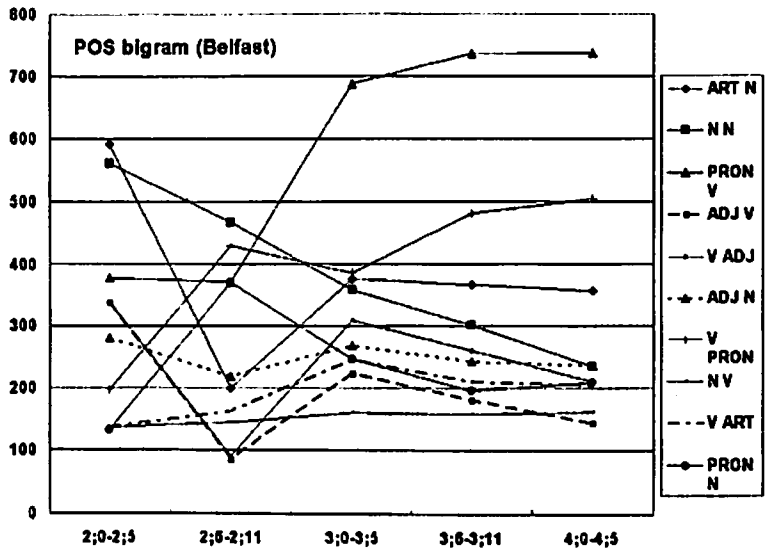

図9：品詞頻度の推移：品詞バイグラム（CHILDES Belfast Corpus）

これに対して、図10は同様のバイグラム・データ上位10組を日本人英語学習者（中学1～3年）に対して取ったものである。モノグラムの場合と同様、日本人データの場合には品詞連鎖の若干の揺れがあるものの母語話者のデータに比べると使用頻度がはるかに安定している。これは前述のような文レベルでの発話指導があるからであろう。最も頻度の高い連鎖は「名詞＋動詞」、次は「代名詞＋名詞（my bookなど）」であった。母語の習得と比べて著しく異なる点は、代名詞の使用頻度が相対的に低いことと、冠詞がまったく現れ

ていないことである。前者は英語学習者データ全般に特徴的な事実で、初級・中級の学習者は代名詞の使用がうまく出来ないため、往々にして代名詞で言えるような場面でも普通名詞で言ってしまう。代名詞の適切な場面での多用は1つの発達指標の重要な要素といえよう。また冠詞に関しても極端な過少使用（underuse）の傾向が見受けられる。母語の習得データでは3歳くらいですでに相当数の冠詞が表れていることを考えると、日本人の英語冠詞習得はやはり相当に難しいということが観察できる。

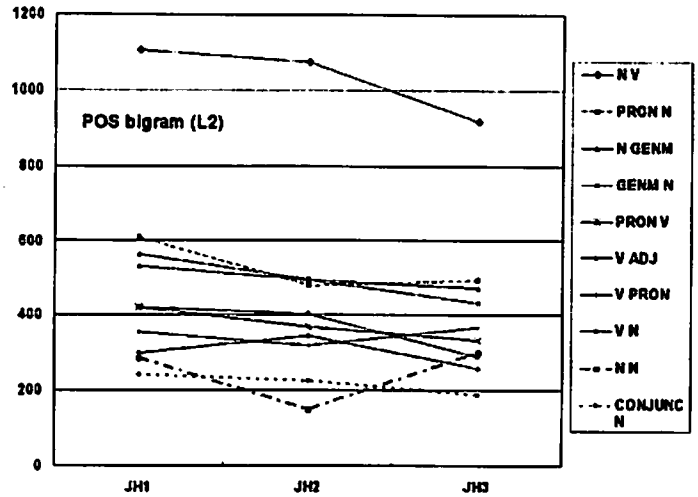

図 10 ：品詞頻度の推移：品詞バイグラム（日本人中学生）

4. 2. 習得順序仮説の検証

学習者コーパスは、習得段階別のデータをコーパス化することにより、習得段階別の記述的な研究のみならず、過去に議論された第二言語習得のさまざまな仮説の再検証を行うことが出来る。例えば、Tono（2000a）は文法形態素の習得順序の再検証を JEFLL Corpus を用いて行った。JEFLL は主として中学1年から高校3年生までの6年間のデータを横断的に集めたもので、現在日本で最も大規模な英作文コーパスである。詳細は Tono（2000a）に譲るとして、結果を図 11 に示した。Dulay、Burt そして Krashen らがまとめた形態素習得順序が上半分に、JEFLL の結果がその下に示されている。これを見ると、第一に目を引くのは日本人学習者の場合冠詞の習得が一番遅い、という事実である。前節の図9の習得段階の記述的なデータでも冠詞の過少使用が指摘されていたが、ここでも冠詞の誤用の多い日本人の習得の特徴が浮かび上がった。面白いことに、PELCRA などのポーランド人英語学習者のデータを見ても、冠詞の習得は遅いことが報告されている（Mason & Uzar 2000）。ポーランド語も日本語も冠詞という品詞の概念がないことを考えると、やはりこれは母語の言語体系に関連した問題ではないかと思われる。

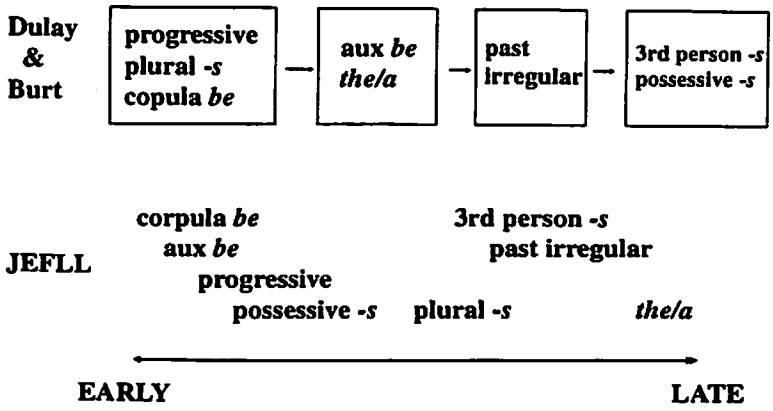

図 11 ：文法形態素の習得順序の比較（Tono 2000a に基づく）

4. 3. 言語習得段階を特徴付ける発達指標の特定

学習者コーパス研究の貢献できる第 3 の領域は正用法・誤用法を習得段階と関連付けて調査し、どのような語彙・文法・形態のエラーが習得段階を判定するのに貢献するかを特定する「発達指標（developmental criteria)」の研究である。この分野も現在目覚しい勢いで発展しているが主要な研究として Tono（2000b)、Supnithi et al.（2003)、 Abe（2003）などが挙げられる。図 12 は SST Corpus からのレベル別エラー発生状況を図示したものである。

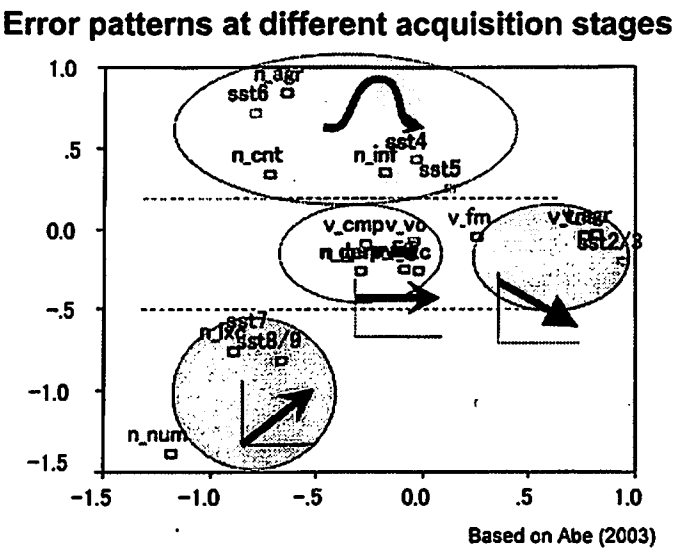

図 12 ：話し言葉コーパスの習得段階別エラーの発生状況 （Abe 2003 に基づく）

これは SST という会話インタビュー・テストの結果 SST Level 2 から 9 までに判定されたグループ（図の sst2 - sst9）の各発話コーパスにエラータグを付与し、そのエラータグの発生頻度とレベルの関係をコレスポンデンス分析で調査したものである。図の右側の方がレベルの下位のものに特徴的なエラー、左に行くにつれて上級レベルに顕著なエラーが要約さ

れている。これを見ると興味深いのは初級のレベルと深い関係があるのは動詞の数・人称などとの一致（v_agr）および動詞の時制（v_tns）に関するエラーが頻出する点である。これらはしかし図の矢印が示すように、レベルが上がるにつれて減少してしまう。一方、中級レベルと関連の深いエラーは動詞ではなく名詞の呼応（n_agr)、加算不加算（n_cnt)、そして屈折（n_inf）などで、これらは中級レベルで多く出現するが上位レベルになるとやはり減退する。そして上級レベルに関連が深いのが名詞の数（n_num）および語彙的エラー（n_lxc）である。これらの関係から、大まかに言えることは、習得の初期は文の組み立ての中心となる動詞を多用することでそれに関連するエラーが増えるが、そこがある程度落ち着くと今度は動詞の周囲の名詞を使いこなそうとしてエラーが発生してきて、これらの屈折や呼応・加算などの形態的な要素を獲得すると、最後には語彙選択のエラーなど上級レベルではより高度な使いこなしを要求される部分でエラーが発生してくる。

Abe（2003）の調査で非常に興味深いのは、習得段階とそれを特徴付けるエラー出現率に一定の傾向があるという点で、このような分析を大量のデータでより精密に行っていくことによって、習得段階を特定する発達指標の同定が可能になってくる。またエラーのみならず、一部の文法事項（例えば動詞の補部（v_comp)、態（v_vo）など）はエラー傾向を見てもレベルとの相関は強くなく、むしろこれらの項目は正用法の使用頻度を見ることでレベル特定に役立つことがAbe（2003）の結果から読み取れる。このような文法項目の習得段階との関係のメリハリがわかることで、今後のSLAの記述研究は新しい領域に踏み込める、という期待感を抱かせる。

4. 4. 多重比較アプローチ（Multiple Comparison Approach）

最後に学習者コーパスを用いた多重比較アプローチを見よう。Tono（2002）は日本人英語学習者の習得研究をコーパス言語学の手法を活かして行う1つの試みとして、英語の項構造（argument structure）の習得を調べた。英語学習者コーパス、母語としての日本語コーパス、教室内での主要なインプット・ソースとしての教科書コーパスの3種類のコーパス・データを相互に比較しながら10の基本動詞に関して、その動詞の下位範疇化情報の獲得が（a）母語の要因、(b）発達的要因、(c）教室内でのインプットの要因、(d）動詞の意味特性、のうちのどれに最も影響を強く受けるか、を見るために、これらの因子の組み合わせのうち最も経済的なモデルを提案する対数線形分析を用いた。

詳細はTono（2002）に譲るが、日本語と英語の動詞の下位範疇化の分類はそれぞれIPAL、COMLEX Lexiconという機械用辞書を参照し、下位範疇化情報の一致度はそれぞれの項目の有無、および相対頻度を元に数式化した。また英語学習者コーパス（JEFLLの一部；約30万語)、教科書データ（150万語)、および2000万語規模の毎日新聞記事データはそれぞれApple Pie Parserによる構文解析、茶筅による形態素解析を行い、正規表現を用いて動詞型頻度を抽出した。これらを学習者の動詞使用の正用・誤用の情報に個々の動詞の基礎情報を付け加えて、10の動詞1つ1つに関してデータベース化し（図13参照)、動詞の正用・誤用にどのような影響があるのか、その関係をモデル化した。

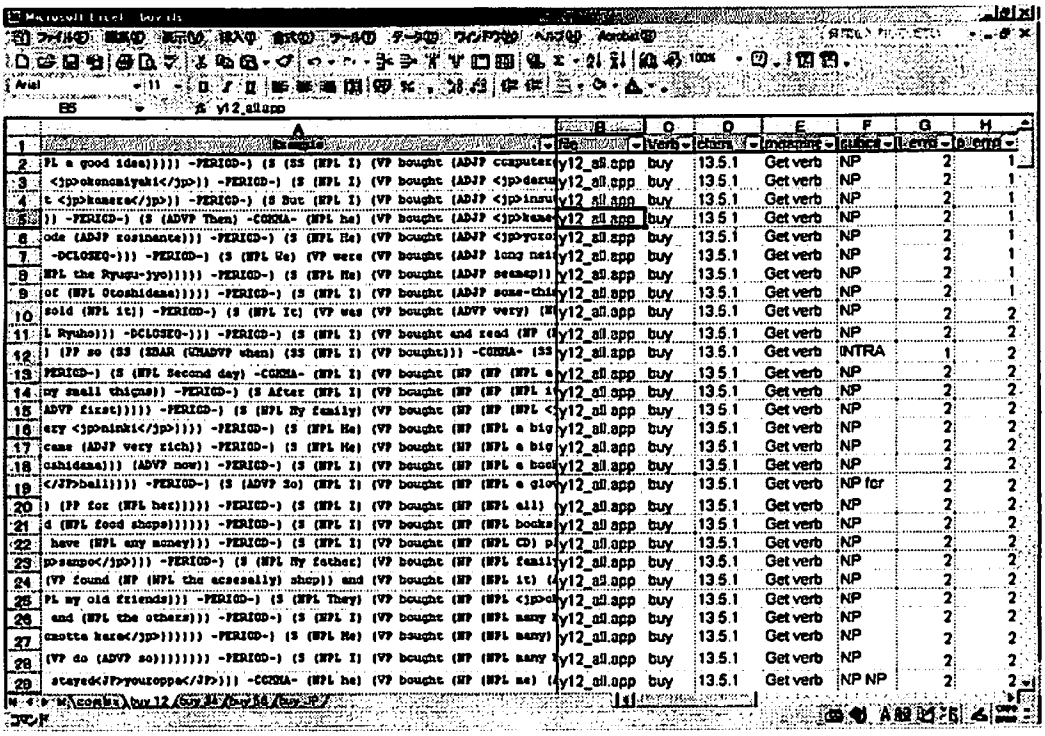

図 13 ：動詞下位範疇化データベース

対数線形分析の結果、動詞の下位範疇化情報の習得に関していくつかの興味深い知見が得られた。第一に、教科書中での使用頻度は動詞型の正誤にはあまり影響を与えず、むしろ過剰使用に強く影響していた。つまり教科書の中に当該の動詞型が多く出てくると、学習者はその型をたくさん使う傾向があるが、その型を正しく身につけるかどうかはまた別問題ということである。第二に、動詞型の誤用に強く関与する要因としては動詞型の日英のパタンのずれ、頻度のずれといった日英対照の要因、および個々の動詞の意味特性といった目標言語に内在する（L2-internal）要因が強いことがわかった。特に一点目の、教科書のインプットと習得の関係が（たとえ項構造の習得という一側面であるとしても）コーパスを用いることでこのように明らかになるのは注目に値する。その他、興味深い結果が動詞ごとにいろいろあるが、ここでは詳細を述べる紙数がないので省略する。

この多重比較アプローチは複数テキストを比べてその差違から統計的に一般的な解釈を導き出そうとするコーパス言語学的な手法に非常に適している。比較する複数コーパスの設計や構築をうまく行えば、研究デザイン的にも多様な研究テーマに利用が可能になろう。そういった意味で、研究用の質の高いコーパスを数多く用意することが急務である。

5. まとめ

本論では、最近進展の著しい学習者コーパス研究を概観し、特に第二言語習得研究への貢献という視点からその可能性をまとめてみた。コーパス言語学が言語習得に貢献するのは、ちょうど脳科学の分野で fMRI などの磁気による脳内観察が言語中枢を生物学・解剖学的

に解明していくのに似ている。理論言語学者のような１つ１つ規則を理詰めで理論化・検証していくプロセスとは異なるが、まったく別の視点・方法から脳内の言語能力の存在に迫ろうとする。コーパス言語学も従来とは比較にならない規模の言語使用データを大量に蓄積し、その言語使用の様相をコンピューターを駆使して時系列的・多次元的に見たときに、理論言語学の視点とはまた異なる新たな事実や発見が必ずある。

まだ若い学問であるこの領域は、発展していくためにコーパス言語学の将来の課題とまったく同じ課題を有している。それは良質なコーパス・データを数多く作ること、そしてコーパスというテキストの集合体から如何にして有意味な情報を抽出するかというテキスト処理と統計をより発展的に融合させること、さらには第二言語習得を含めた関連諸分野の応用例を蓄積すること、である。個人的には良質な学習者コーパスの分析結果によって、近い将来、日本人英語学習者の英語習得過程が克明に記述され、日本のようなEFL環境での英語習得プロセスが飛躍的に解明されることを期待している。

参考文献

Abe, M. (2003). Variability in interlanguage: a corpus-based approach. Paper presented at the 22nd Japan Association of English Corpus Studies Conference.

Ellis, R. (1994). *The Study of Second Language Acquisition*. Oxford University Press.

Færch, C., K. Haastrup and R. Phillipson (1984). *Learner Language and Language Learning*. (Multilingual Matters 14) Clevedon: Multilingual Matters.

Garside, R. Leech, G. and McEnery, T. (1997) *Corpus Annotation*. London: Addison Wesley Longman.

Granger, S. (ed.) (1998). *Learner English on Computer*. London: Addison Wesley Longman.

Greenbaum, S. (1992). A new corpus of English: ICE. In J. Svartvik (ed.), *Directions in Corpus Linguistics*, pp. 171−179. New York: Mouton de Gruyter.

Leech, G. (1998). Introduction to S. Granger (ed.) *Learner English on Computer*. London: Addison Wesley Longman.

MacWhinney, B. (1995). *The CHILDES Project: Tools for Analyzing Talk* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Mason, O. and R. Uzar (2000). NLP meets TEFL: Tracing the zero article. In Lewandowska-Tomaszczyk, B. and J.P. Melia (eds.) (2000), *PALC '99: Practical Applications in Language Corpora*. Frankfurt: Peter Lang, pp. 105−116.

Milton, J. (2001). *Describing and overcoming environmental limitations on the interlanguage of Hong Kong Chinese learners of English: a computational and corpus-based methodology*. Unpublished PhD thesis. Lancaster University.

Perdue, C. (ed.) (1993). *Adult language acquisition: cross-linguistic perspectives*. 2 vols. Cambridge: Cambridge University Press.

Sinclair, J. and Coulthard, M. (1975). *Towards an Analysis of Discourse*. Oxford: Oxford University Press.

Supnithi, T., Uchimoto, K., Saiga, T., Izumi, E., Sornlertlamvanich, V. & Isahara, H. (2003). Identifying the acquisition stage of L2 learners: a learner corpus-based approach. Paper

presented at NLP2003, March 2003 at Yokohama National University.

Tono, Y. (2000a) A computer learner corpus-based analysis of the acquisition order of English grammatical morphemes. In Burnard, L. and T. McEnery (eds.) *Rethinking Language Pedagogy from a Corpus Perspective*, pp. 123−132. Frankfurt: Peter Lang.

Tono, Y. (2000b). A corpus-based analysis of interlanguage development: Analysing POS tag sequences of EFL learner corpora. In Lewandowska-Tomaszczyk, B. and J.P. Melia (eds.) (2000), *PALC '99: Practical Applications in Language Corpora*. Frankfurt: Peter Lang, pp. 323−343.

Tono, Y. (2002). *The Role of Learner Corpora in SLA Research and Foreign Language Teaching: The Multiple Comparison Approach*. Unpublished Ph.D. thesis. Lancaster University.

van Rooy, B. and Schäfer, L. (2002). The effect of learner errors on POS tag errors during automatic POS tagging. *Southern African Linguistics and Applied Language Studies*, 20: 325−335.