連載

学習者コーパス入門

学習者コーパスのあるべき条件とその作成の具体的方法は?

英語学習者の発話や作文などの performance data を
コンピュータ処理し，教育や研究に役立てるために

第5回 実例に見る 学習者コーパス分析 (1)

元東京学芸大学講師／ランカスター大学言語学科博士課程在籍 投野 由紀夫

はじめに

これまで4回にわたってデータ収集の考え方の基礎，および収集したデータの書き起こし，フォーマットの方法，ファイル管理の実際を解説してきた。とくにフォーマット部分では，コーパス・データ管理の最先端の方法を紹介してきたのでご自分のデータをお持ちの先生方は本格的な学習者コーパスの構築が可能なはずである。

さて，最後の2回は今までのデータ構築の基礎的な部分からコーパス分析の具体例に移って解説を行う。ここではすでに筆者が構築中の JEFLL Corpus をもとにして分析の実例を紹介しながら，比較的容易に分析が可能な項目と複雑なデータ再処理・加工が必要な分析と段階を追って紹介したい。

コーパス・データ再処理の複雑度

Transcription が終わった学習者データをいざ分析しようとすると，「はたして何をやればいいのだろう?」と疑問に思ってしまう人もいるだろう。データはあっても調べたい研究テーマが浮かんでこないということはよくある。しかし，これは学習者コーパスの問題ではなくて，調査する研究者の側の問題だ。学習者コーパスは再三述べているとおりただの学習者の作文や会話データなので，そこから何を知りたいのかは研究者ひとりひとりによってさまざまである。学習者の語彙や文法の獲得に興味がある人もいれば，語用論的な観点から会話分析をしようとする人もいるだろう。作文の文体的な面に興味のある人もいれば，もっと practical な面からトピックによる発信語彙のリストを作りたいという人もいる。

研究分野によるトピックの多様性は当然のことであるのでとりあえず置いておくとして，ここでは目的に応じたデータ再処理の複雑度の観点から少しコメントをしておこう。データ再処理の複雑度，というのはたとえばこういうことである。学習者データがすでに一定量蓄積があるとして，そこから「冠詞の用法」について調べたいとしよう。コーパスがあればすぐに a, the といった冠詞の検索は容易にできる。ところが，これらの冠詞のうち「誤用」について検索できるか，というと，コンピュータはそのままでは冠詞の誤用について判断してくれない。「正用」と「誤用」に関する情報をデータの中に埋め込んでやらなければならない。これがエラータグ付与（error tagging）である。つまり冠詞の誤りをコーパス・データで組織的に調べようとすると，書き起こしたデータにエラー分析の判断をタグ付与という形で新たに与えなければならないわけだ。それだけではない。冠詞の誤りにはいわゆる無冠詞の誤用ということもありうるので，冠詞が出現する部分だけでなく，すべての名詞の前で冠詞が正しく使われたかの確認をしなければ正確に冠詞の正用・誤用の割合はつかめない。こう考えるとエラータグ付与の手間は膨大な時間を要する。

一般にこのようなデータ再処理の複雑度は調査しようとする研究対象によってかなり異なってくる。表1に参考までに比較的調査が容易なものと再処理が複雑なものを段階的に示してみた。

ここですべての事例を紹介することはできないが，今回と次回にわたって4つの事例研究を概観しながら，データの加工を複雑に行っていくことでいかに学習者英語の特徴がいろいろな角度から捉えられていくかをおわかりいただければと思う。

表1

コーパス・データ加工の複雑度	抽出可能な頻度情報・分析結果
書き起こしデータがあれば再加工は原則的に不要	1) 単語頻度リスト作成 2) コンコーダンス作成 3) 共起頻度抽出 4) コロケーション分析 5) 特定の語用の頻度抽出 6) 単語を手がかりとしたクラスター分析 7) 単語を手がかりとしたキーワード分析 8) 単語を手がかりとした文法構造分析
自動処理によるタグ付与必要 ・品詞タグ付与 (POS tagging) ・見出し語処理 (lemmatisation) ・N グラム統計 (n-gram statistics)	1) 品詞別単語頻度リスト作成 2) 見出し語化頻度リスト作成 3) 品詞を手がかりとした語彙・文法構造の分析 4) コリゲーション分析 5) 品詞を手がかりとしたクラスター分析 6) 品詞を手がかりとしたキーワード分析 7) 品詞連鎖をもとにした構造分析
部分的手動タグ付与 ・構文解析 (full parsing) ・意味解析 (semantic tagging) 手動によるタグ付与 ・エラータグ付与 (error tagging) ・韻律・音声情報付与 (phonetic transcription/ prosody) ・語用論的情報付与 (pragmatic annotation) ・意味情報付与 (semantic annotation) ・談話情報付与 (discoursal annotation)	1) 音声・韻律タグ付与による音声・韻律分析 2) 意味タグ付与による語彙分析 3) 語用論的分析 4) エラータグ付与によるエラー分析 5) 構文解析情報をもとにした複雑な文法構造分析 6) 談話情報付与による談話分析

〈ケーススタディ1〉 英語学習者のコロケーションの誤り

はじめにもっともデータ加工の単純なコロケーションの分析を取り上げよう。コロケーションは現在、語彙習得の分野でも非常に関心の高い研究分野である。とくに大規模コーパスの利用が可能になって、ネイティブ・スピーカーの言語使用の状況を客観的に調査することが可能になってくると、語と語の結び付きに関する情報が単純に「A と B の結び付きが可能か? (possibility の問題)」という質問だけではなく、「A と B の結び付きはどの程度普通か? (probability の問題)」という確率的な情報も提供できるようになってきた。コロケーション情報は日本人英語学習者にとって自然な英文を発信する際に必須であり、最近の英語学習辞典にはコロケーションの分析結果を反映しているものも増えてきている。

学習者コーパスは、ネイティブの用いる自然なコロケーション頻度に対して、学習者英語がどのようなコロケーションを好むか、それがネイティブのパタンとどのように異なるか、といった情報を提供してくれる。とくにネイティブが用いない不自然なコロケーションや誤った語と語との結びつきを身に付けている場合には、その特徴を分析し原因を探ることで、言語獲得のメカニズムや学習上の困難点などに迫る研究が可能になってくる場合もある。データそのもののの加工があまり複雑でない分、コロケーション統計など共起度を表す統計的データの部分に多少の専門的な知識が必要になるだろう。

Classの分類表現

さて、早速実例を見てみよう。私が中心で構築しているJEFLL Corpus の中学1年〜3年の国立附属のデータ約2300名分の英作文 (15万語強) を使ってみる。専門的なコロケーション分析は紙面が限られているのでできないが、身近な例として class という名詞を取り上げて、そのコロケーションを WordSmith で見てみることにしたい。

Class のコンコーダンスラインを取り出した後、その直前の単語でソートした結果を collocates として一覧表にすると図1のようになる。これを見ると上位3つは"our class"、"my class"、"every class"などでとりわけ不自然ではないが、その後に B や A というのが出てくるのでおやっと思う。そこで直接該当する例文に飛んでみると図2のような例文が出てくる。

かなりの生徒が「B組」という意味を"B class"というように表現していることがわかる。このような class の前に数字を持ってくる用法は全813例中、51例 (6%) に上った。our、my、every などの次に高頻度で出てくる用法である。

図1

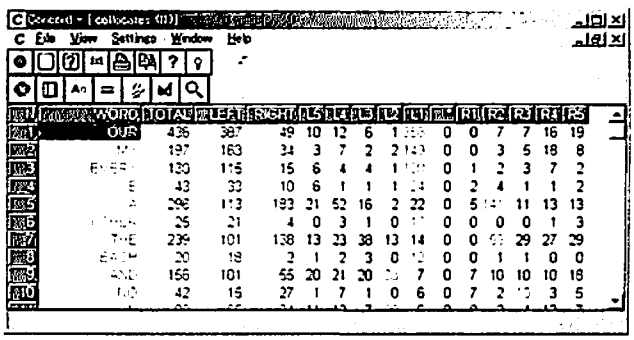
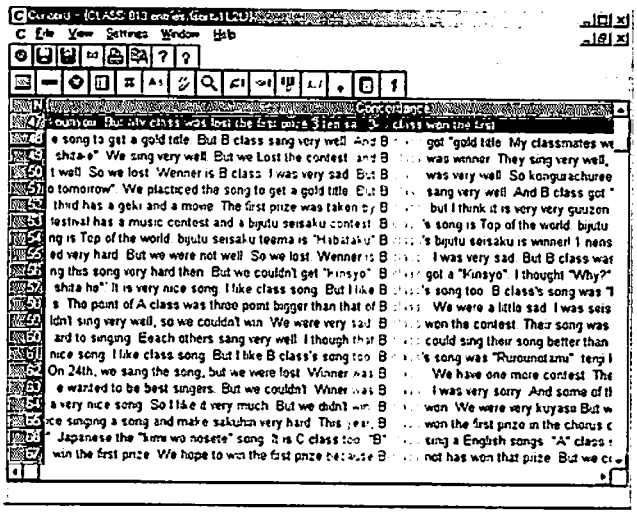


図2



「2組」とか「B組」というときには、class の前と後ろのどちらに数字を置くのが普通なのだろうか？ 中学以上になると、ほとんどの欧米の学校が科目によって教室を移動するようになるので自分のクラスという概念が多少日本と異なるのだが、日本ではたとえば「2年4組」のことを英語で言う必要が出てくるだろう。案外、こういう身近な表現が英和辞典の class の項目には解説されていない。

そこでネイティブ・スピーカーのコーパスを参照してみることにした。British National Corpus という90年代イギリス英語の書き言葉、話し言葉を組織的に収集した1億語のコーパスで class を検索し、名詞の直前にくる数字の表現を抽出したところ表2のようになった。表の total は全コーパス中での単語個別の出現頻度、n col. が class の直前に出現する頻度、n txt. が出現したテキストの数である。左側上位3位の one, two, three はクラスの分類の表現ではなく、「1つのクラスでは」と単純に数を表していて該当するコロケーションではない。これを見るかぎりでは、class をカテゴリー化して「2年A組」のように表現する際には圧倒的に class 1 / 2 / 3 とか class 2-A といった言い方が普通で、class の前に数字を置くのは稀であることがわかる。この一見、単純なコロケーションの誤りのように見える現象は、日英の名詞句構造の相違とそれが第2言語習得に及ぼす影響といったより一般的な問題へと発展していくのだが、ここでは詳細には触れない。

表2:BNCのclassの前後に出てくる数詞のtop10の頻度

Rank	Left				Right			
	word	total	n col.	n txt.	word	total	n col.	n txt.
1	one	190499	145	111	ii	8629	165	26
2	two	156111	10	9	1	38384	102	36
3	three	79759	7	6	2	34394	65	27
4	2251	5	3	2	i	8509	61	17
5	fifty	8579	2	1	4	20017	50	26
6	i	8509	2	2	5	17557	30	15
7	1	38384	2	2	3	25040	25	12
8	2-6-0	12	2	1	two	156111	20	15
9	500	2352	2	1	iii	4853	19	12
10	250	1160	2	2	three	79759	19	19
Total			177				556	

✱ 動詞comeの用法

2つめのコロケーションの実例は come を見てみることにしよう。これは正確に言えば collocation ではなく、動詞の持つ文法的な特性なので、イギリスのファース学派風に言えば colligation、アメリカの言語学の潮流で言えば argument structure とか subcategorization という問題になるだろうか。しかし、コーパスの処理上は項構造（argument structure）を機械的に特定する処理は行わずに単純に表面上の単語連鎖をもとに分析するので、このセクションでは「文法的なコロケーション」としておきたい（項構造のより専門的な分析の例は最終回で解説する予定）。

同じ中学1〜3年の英作文データで come の次にくる要素を調べてみた。Come の全用例は389例で WordSmith のコロケーションのアウトプットは表3のようになる。これを見るかぎりでは、come to, come back, come home, come here などが高頻度の結び付きであることがわかる。10位〜20位を見ると、come の直後に所有代名詞 our, my などがくる例がある。これらは *come my house 型の誤りで、全体の割合は5%ほどだが come の項構造の獲得を考える上では非常に重要な誤りのパタンである。

表3:
comeの右1番目のコロケーション

N	WORD	FI
1	TO	108
2	BACK	71
3	HOME	31
4	HERE	17
5	AND	12
6	THERE	12
7	IN	10
8	INTO	9
9	FROM	8
10	COME	7
11	*OUR	7
12	TRUE	6
13	*MY	4
14	AT	3
15	*HIS	3
16	UP	3
17	WITH	3
18	AFTER	2
19	*HIM	2
20	ON	2

実はこの現象はもう少し根が深い。一見正しい形に見える come to のコロケーションをより詳細に調べてみると *come my house 型の誤りとは逆のケース、つまり *come to home / here のように come to+名詞の用法を副詞にまで過剰般化（overgeneralize）してしまっている誤りがやはり6〜7%見受けられる（図3）。

図3:come toの後に副詞が来る誤りの例

1	at festival.It practice is very hard. I **come to** *home 6:45 in the evening.
2	mata, There is this dream, too. I **came to** *home slowly the other day.
3	get up my family, but I didn't. thief **came to** *here. I fought him in order to sa
4	o watarouto shita. suruto, Car is **come to** *over there. This car is aka shingo
5	ime-sama said, "Please don't open." He **came to** *home, and opened of tamatebako.
6	e got to Ryugujo he said to monban, "I **come to** *here to see Otohime-sama. An
7	le around me. I was lonley. Nobody **come to** *here. It was a very bad dream.

図4:*become toの誤用例

1	eienni. And turu to kame **became** *to medetai mono
2	But became his around. He didn't **become** *to young man. He knew th
3	the morning. Because if I don't, I'll **become** *to be hungry. But, becaus
4	the butcher became like that, Wilbur **became** *to 10 tears old and was in Ryug
5	as very happy. Because he want to **become** *to be torendiina shibui rojin
6	He died, but he **became** *to a turu. Now in Jap
7	ox? It is very cheap now." He **became** *to sell the box. The man s
8	I like it better. Sometime have **became** *to the karee in the next mo
9	If there are a lot of money, I only **become** *to be able to buy a lot of thing.
10	ney. Nobody know the young man **became** *to. Then, you know what

これらの誤用例を見てみると「～に来る」という意味を“come to ～”という表現に1対1で置き換えて規則化した場合に *come to home 的な誤りが生じ、逆に come home をデフォルトとして規則化すると *come my houseのような誤りが出てきてしまうと予想される。しかし、現象としては確かにそうだが、そのように学習者が誤った規則を立ててしまう背後には、はたしてどのようなメカニズムが存在するのであろうか?

さらにこの中間言語の過程を興味深いものとするいくつかの事実がある。come to には「～するようになる」という意味で to 不定詞をとる用法があり、中学3年生くらいになると部分的にこのパタンを使えている学習者もいる。実はこの come to do の表現は圧倒的に *become to do という誤用で用いられるケースが多い（図4）。

Become は下位範疇として to 不定詞はとらないのだが、学習者が come to do に見られるようなある種の動詞の項構造をインプットから観察し、かつ日本語での「～になる」という意味と become がほぼ同一視されることを考えると、この誤用にはかなり複雑な日英の動詞の意味特性に関する相違が第2言語としての英語の習得に影響を及ぼしているということが考えられる。コーパスそのものは決してこの背景にある理論に対して直接的な答えを与えてはくれないが、中間言語の示す学習者の持つ言語知識の証拠は多くの問題点を指摘してくれるわけである。もちろん、この事実はそういった言語習得理論との関連で研究される以外にも、学習上の困難点として文法・語彙指導、参考書や辞書への記述など、いろいろな方面に示唆を与えることができる。

〈ケーススタディ2〉自動品詞タグ処理を施した学習者データを分析する

続いて、もう少し書き起こしデータの中身に手を加えた例をお見せしよう。コーパスのテキスト中への情報付与（annotation）にはいろいろなレベルが考えられる（表4）。この中で品詞タグ付与に関しては、かなり精度が高くなってきているので、学習者英語に自動品詞タグ付与を試みて、そのデータをもとに分析をする例を紹介したい。

表4:コーパスへの情報付与

レベル	自動化	レベル	自動化
orthographic	M	syntactic parsing	A / M
phonetic / phonemic	M	semantic	A / M
prosodic	M	discoursal	A / M
part of speech	A	pragmatic / stylistic	M

A:ほぼ自動化されている　M:手作業が普通
A / M:自動処理一部可だが修正必要

自動品詞タグ付与（Automatic POS tagging）

さて、ある程度まとまった量のテキストを自動で品詞タグ付与をするには、それなりのプログラムを使わなければならない。

日本ではまだ品詞タグ付与に関して本格的に解説した言語学の本はないが，興味がある方はGarside et al (1997)，または国内では自然言語処理の分野で長尾(1996)，北(1999)の一部の記述が参考になる。

一般の中高の先生方が自分のデータに品詞タグを付けたいとすると，大別して2通りの方法が考えられる。1つは一定料金で自動タグ付与サービスを供給している大学などの研究機関に依頼する方法がある(表5参照)。もう1つは自分で利用可能なタグ付与プログラムをインストールして使うという方法があるが，これには Brill's Tagger を除いてほとんど UNIX の知識が必要なので，ここでは扱わないでおく。

表5:品詞タグ付与サービスをしてくれる代表的なサイト

UCREL tagging service (CLAWS) http://www.comp.lancs.ac.uk/computing/research/ucrel/claws/tagservice.html
AMALGAM email tagging service (Brill's Tagger) http://www.scs.leeds.ac.uk/amalgam/amalgam/amalgtag3.html
Corpus Reserch Group at Birmingham University http://clg1.bham.ac.uk/tagger.html
MBT Tagger Page: http://ilk.kub.nl/~zavrel/tagtest.html

タグ連鎖の抽出

学習者コーパスに自動品詞タグを付与すれば，品詞情報を基にしたいろいろな分析を行うことができる。私がランカスターで行った分析の1つを紹介しよう(詳細は Tono 2000 参照)。

まず学年別に採取した英作文データを CLAWS というランカスター大学の開発した品詞タグ付与プログラムによって品詞タグ付けを行い，そこから品詞タグの連鎖を抽出して，その連鎖の頻度が学年を追うごとにどのように変化するかを見てみた。

専門的には CLAWS の vertical output を利用して，そこから perl などでスクリプトを書いて品詞連鎖の頻度集計をするのだが，そのような技術的なことを知らなくても WordSmith を用いることで比較的同じような分析をすることができる。その方法をここでは紹介したい。

品詞タグ付けをしたデータは図5のような出力になっている。これを「秀丸」などのテキストエディタに読み込んでから，「置換」処理をして単語部分を削除してしまい，品詞タグの

図5: CLAWSで品詞タグを付けた学習者データ

```
^ When_CS I_PPIS1 am_VBM so_RG busy_JJ and_CC
sick_JJ ,_, I_PPIS1 always_RR see_VV0 that_DD1
dream_NN1 ._.
^ It_PPH1 is_VBZ terrible_JJ for_IF me_PPIO1 ._.
^ Once_CS@ I_PPIS1 lived_VVD by_II the_AT river_NN1
._.
^ The_AT river_NN1 (_( )_) under_II the_AT road_NN1
._.
^ The_AT dream_NN1 which_DDQ I_PPIS1 always_RR
see_VV0 is_VBZ terrible_JJ because_CS I_PPIS1 fail_VV0
the_AT river_NN1 ._.
```

図6

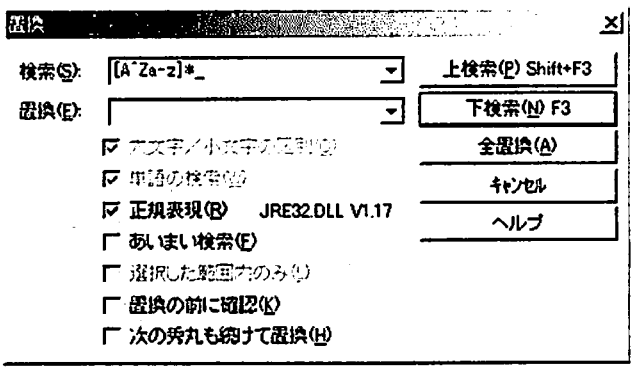
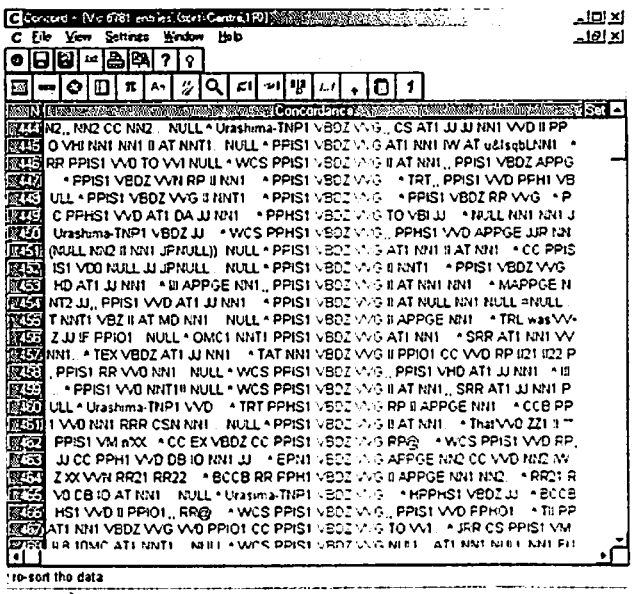


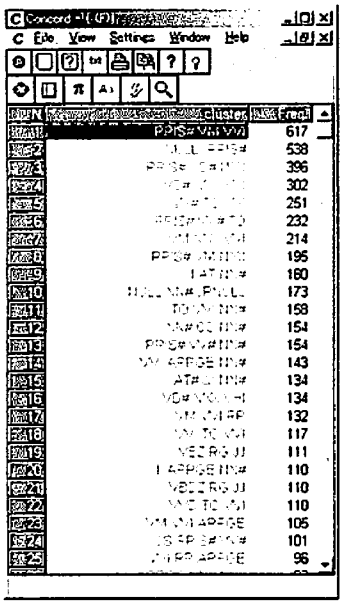
図7



連鎖のみのテキストを1セット作成する(図6の秀丸における正規表現を用いた置換例を参照)。こうするとこの品詞連鎖をあたかも単語のように見なしてWordSmith で検索することができる。

このデータを WordSmith で読み込んで，各文に1度は出

図8



(注) この画面では品詞タグ内の数字が # になって，正しく認識されていない。クラスターは英文字のみを処理するので，品詞タグを事前に数字の含まれない簡易タグに変換してからクラスターをする必要がある。

てくるはずの動詞 (V*) で検索してやると，データのほぼすべてのラインが検索結果として表示される (図7参照)。これをもとにしてクラスタリング (共起頻度をもとにした単語や品詞連鎖の集合を抽出する手法) を行えばよい。クラスターの数を3にすれば 品詞タグの3個組 (trigram) のデータが，2にすれば2個組 (bigram) の統計がとれるというわけだ。クラスターの集計結果のようすを図8に示しておく。

このようにして処理したデータを学年ごとに並べて品詞並びのデータを分析していくとおもしろいことがわかる。それは低学年の英作文ほど"verby" (動詞中心の連鎖) であるということである。第1言語の習得データの場合には，初期の発話の単語連鎖は名詞中心でそれから動詞中心に移行し，さらにまた名詞中心に戻るのだが (cf. Tono 1999)，学習者の英語は少なくとも英作文データに関するかぎり，最初から動詞がきちんと決まって，それを中心に短い名詞句と前置詞句を項構造においた文が成立する。

その後発達段階を経るにつれて，構造は名詞を中心とした連鎖の頻度が増す。それは名詞句が修飾語などを伴って徐々に長くなるためである。大学生レベルになると，徐々に名詞句が複雑になって前置詞の品詞連鎖に占める割合が大きくなるはずなのだが，日本人英語学習者の場合には他国の英語学習者に比べると名詞句の内部構造が十分発達しないままで大学生レベルを終わってしまう。ほんとうに academic writing などをきちんと教えられていないので，複雑な名詞構文などがほとんど使えない状態でボツボツした文を書くわけだ。

このような作文の特徴分析は中間言語の一側面しか示していないかもしれない。しかし，このように中間言語のデータを電子化して文法情報を加えることによって新たな可能性が開けてくる。たとえば，中間言語データをもとにした機械による言語モデルの構築もそのひとつだ。コンピュータに中間言語の文法を確率的に獲得させるようなモデルを評価しながら，言語習得理論と組み合わせて研究していくことによって，新しい第2言語習得のモデル化ができる可能性がある。

＊　　＊　　＊

さて，次回はいよいよ最終回なので，さらにデータ加工を進めて，エラータグ，構文解析を利用した研究事例をご参考までに紹介したい。このような研究の可能性が，ますます中学・高校の先生方の「学習者データ」に対する価値の再認識につながることを期待して，今回はここまで。

参考文献

Garside, R., G. Leech and A. McEnery. (eds.). (1997). *Corpus Annotation*. Longman.

北研二. (1999). 『確率的言語モデル』. 東京大学出版会.

松本裕治 他. (1997). 『単語と辞書』(岩波講座：言語の科学3). 岩波書店.

長尾真 (編). (1996). 『自然言語処理』(岩波講座：ソフトウェア科学15). 岩波書店.

Tono, Y. (1999). Developmental patterns of Japanese EFL learners' POS tag trigrams. *Japanese Society for Language Sciences: First Conference Handbook*, pp. 13-16.

Tono, Y. (2000). "A corpus-based analysis of interlanguage development: Analysing part-of-speech tag sequences of EFL learner corpora." In Lewandowska-Tomaszczyk, B. and P. J. Melia. (eds.). *PALC' 99: Practical Applications in Language Corpora*. Frankfurt: Peter Lang, pp. 323-340.

問い合わせ：y.tono@lancaster.ac.uk

● お詫びと訂正のお願い

「STEP 英語情報」(2000 11・12) p.57 図6で，

記号「w k 9 e t 001.xml」

が，印刷ミスのため脱落してしまい，読者ならびにご執筆の投野由紀夫先生に大変ご迷惑をおかけいたしました。

ここに心よりお詫び申し上げますとともに，慎んで下記の通り訂正させていただきます。

w k 9 e t 001.xml
- 001：通し番号
- t：tagged
- e：トピック記号
- 9：学年(year 9)
- k：国立附属
- w：writing

STEP英語情報編集部