はじめに

今まで学習者コーパスをもとに出来る第2言語習得研究や外国語教育への示唆をさまざまな角度から見てきたわけだが、今回は学習者コーパスを使った中間言語の文法構造解明の試みを紹介しよう。今回のテーマは英語指導にすぐに使えるというような practical なものではないが、このような研究が進めば、英語学習者がたどる文法発達の道筋がより明らかになり、結果的にシラバスや教授法などに大きな変更や助言が出来る可能性があることを念頭に置いて読んでいただきたい。

品詞の連鎖パタンの研究

さて、中間言語の文法構造をどのように研究するのだろうか？ アプローチの仕方はいろいろあるが、ここではコーパスを用いて出来るだけコンピューターによる自動化を念頭に置いた研究方法を紹介してみよう。その1つが、コーパスに付与されている品詞タグの連鎖パタンの研究である(自然言語処理では、2個組 (bigram)、3個組 (trigram) などと称して、文字・形態素・品詞などさまざまなレベルの言語統計情報を算出するのによく用いられる概念だ)。

コーパス・データに各単語の品詞情報をつけることはもはや常識になっている。British National Corpus, Bank of English などの1億語以上の大規模コーパスも各単語の品詞情報を持っている。これらの品詞タグは現在ではほとんどが自動品詞タグ付与のプログラムで付けられている。代表的なものは、Lancaster 大学の CLAWS, Nijmegen 大学の TOSCA, London 大学の ICE などである(どのタグ付与プログラムも有料で利用することが出来る)。

この10年ほどで文体計量学 (stylometry) の分野で、品詞のタグの連鎖パタンを抽出することにより作家特定 (authorship attribution) をする研究が盛んに行なわれるようになった。この手法を応用して、学習者の書いた作文などの品詞連鎖パタンを大規模に抽出することにより、学習者の中間言語の特徴的な構造パタンを取り出そうというのが、今回紹介する試みである。

Native speaker のパタンは？

学習者コーパスで特徴を探る前に、ネイティブ・スピーカーがどのような構造をよく使うのかを見てみることにしよう。表1はアメリカ人大学生の書いたエッセイ20万語のデータから抽出したものである(データの詳細は Aarts & Granger 1998 参照)。

表1 Native speaker の品詞連鎖パタン Top 20

	Pattern	Example
1	PUNC # #	
2	PREP ART N	
3	N PUNC #	
4	ART N PREP	
5	N PREP N	
6	N PREP ART	
7	ART ADJ N	
8	V ART N	
9	ADJ N PUNC	
10	PREP N PUNC	
11	ADJ N PREP	
12	# # PRON	
13	ART N PUNC	
14	PREP PRON N	
15	# # N	
16	N PREP PRON	
17	N AUX V	
18	PREP ART ADJ	
19	V PREP ART	
20	PRON AUX V	

注: # = sentence break; PUNC = punctuation

これは3個組 (trigram) と言って、品詞タグの3個分の連鎖を抽出して頻度順に並べたものである。品詞の連鎖なので、NP, PP, VP などの高次の統語範疇にはきちんと区切れていないが、品詞がお互いに確率的にどのような品詞と共起するかがよくわかる。

第1位に来ている PUNC # # は文末のピリオドのことで、あまり情報としては有益ではないが、全体を見て気づくのは、上位にはかなり名詞と前置詞に関する組み合わせが多く、10位台後半に動詞がらみの連鎖が多く現れていることである。これは文単位で考えれば、1つの動詞を中心に名詞句や前置詞句が複数接続する文構造を持ち、動詞部分よりもその周辺の名詞句、前置詞句の構造が複雑であることを示している。

表2 Native speaker と non-native speaker の品詞連鎖パタンの比較 (Aarts & Granger 1998: 135)

	ネイティブ Top 10	オランダ人	フィンランド人	フランス人
1	PREP ART N	−	−	−
2	ART N PREP	−	−	−
3	N PREP N	−	−	−
4	N PREP ART	−	−	−
5	ART ADJ N	～	～	+
6	V ART N	−	−	−
7	ADJ N PREP	−	～	−
8	## PRON	+	+	+
9	PREP PRON N	−	−	−
10	## N	−	−	−

(注) +: 過剰使用; −: 使用不足; ～: 有意差なし

Aarts & Granger (1998) では、分析をさらに進めて学習者データによる比較をしている。各々約15万語からなる3つの学習者コーパス(フランス人、オランダ人、フィンランド人大学生)で同様のtrigramの頻度分析を行ない、それをネイティブ・スピーカーの頻度と比較して、特に頻度順位に著しく差のあったものを抽出したのが表2である。

プラス(+)のマークが過剰使用 (overuse)、マイナス(−)が使用不足 (underuse)、～のマークはネイティブとの差がなかった項目を示している。これを見ると興味深い点がいくつかある。1つは、3つの学習者コーパスの傾向が非常によく似ていることである。第4位までは一貫して使用不足であり、名詞句のパタン (ART ADJ N) のみ、ネイティブと同傾向を示している。

さらに面白いのは、ネイティブの使用頻度に比べて、特に学習者が苦手とする構造は前置詞がらみの構造だという点だ。Aarts & Granger (1998) でもこの点は指摘されている。大学生レベルの上級学習者であっても、ネイティブ・スピーカーに比べると前置詞を含む trigram の連鎖が有意に少ない。これは何を表わしているのだろうか？

1つは、前置詞句を使いこなせているかどうかが、ネイティブと外国人学習者を分ける1つの決め手になると言う点だ。前置詞句は、関係節などに比べて、重層的に繰り返しても文の容認度が落ちない。コーパスからも、はっきりと前置詞句を多用するネイティブの傾向が窺える。それに対して、学習者の英語では、複雑な前置詞句そのものの頻度もそうであるが、内部構造の複雑な前置詞句がなかなか作れないのである。

第2に、この前置詞句の使用不足は、指導方法にも関連があるかもしれない。たとえば、学校で教える英語の中での前置詞句の取り扱いは非常に周辺的なものだ。後置修飾の中では関係節や分詞はよく取り上げられ、文法書などでも解説がよくされるのであるが、前置詞句の取り扱いは語彙的な扱いが強くなり、往々にして文法の教科書でも最後のほうにおまけ程度に解説してあるくらいでマイナーな地位に甘んじている。

しかし、学習者コーパスを見る限り、かなり上級の学習者であっても、前置詞句を十分に使いこなせていない。一貫して見られる使用不足は、学習者の中間言語の特徴を示していると言え、後置修飾の指導法に示唆を与える興味深い事実だと言えよう。

日本人学習者の特徴は？

さて、Aarts & Granger (1998) の研究に触発されて、日本人学習者データでどの程度同様のアプローチが可能なのかを検討してみよう。日本人英語学習者コーパスのうち、タグ付与と後処理の関係で実験的に3万語(中学生)を抽出して、それを

英語指導 ……(11) Interlanguage grammar
words 投野由紀夫

もとに Aarts & Granger と比較を行なってみた。

まず、ランカスター大学の開発した CLAWS4 で学習者コーパスに自動処理で品詞をつけてみる。この場合、本来ならば postediting で誤ったタグなどの特定を行なうべきなのだが、今回はこの記事のための予備的な調査としてこの段階はカットしてしまった(しかし、ざっとテキストを見渡した感じでは CLAWS4 は内蔵の辞書によって基本的な単語についてはかなりの品詞特定を正確にしてくれるので、文法構造自体が多少乱れていても品詞は比較的正確についている印象を受けた)。CLAWS4 はタグが 146 種類あり (C7 tagset)、これでは細かすぎるので、Aarts & Granger を参考にして大きな品詞区分にタグを自動変換する(このへんは gawk などのツールで処理する)。その後、タグの n-gram の処理のため、元ファイルからタグの連鎖だけを抽出して別ファイルを作る。それを WordSmith にかけて、クラスター頻度を算出する、といった手順である(この手順および分析の詳細は Tono (forthcoming) 参照)。

表3　日本人英語学習者(中学生)の品詞連鎖パタン(約3万語)

PUNC PRON V	1740
PRON AUX V	1026
CONNECT PRON V	906
ART N PUNC	705
V ART N	687
V N PUNC	609
V PREP V	501
PRON V ADV	498
PREP ART N	492
PRON V PREP	474
V ART N	474
PRON V N	450
V PRON N	441
AUX NOT V	399
V ADV ADJ	396
V ADV PUNC	360
PRON V PREP	342
V PUNC PRON	336
PRON ADV V	318

表3が日本人の中学生レベルでの trigram の頻度リストである。ここでお断りしておかねばならないが、Aarts & Granger の分析では文区切りの記号 (#) を分析に含めていたが、今回の私の集計では文区切りの記号はあまり情報量がないので含めていない。

この表3からどのような日本人英語学習者の特徴がわかるだろうか？ まず注目されるのは、ネイティブのトップ20では動詞関連の連鎖が20位中4つだけであったのに対して、表3では16と圧倒的に動詞関連のパタンが多いことである。これは非常に面白い事実であって、中間言語としてこの時期の学習者が文を作る際に、動詞を中心とした単語の連鎖に頼って文を生成しているということを示している。それが、上級学習者になるほど、1つの動詞に対してその主語や目的語などとして取る名詞句などの構造が複雑になり、前置詞句も含めた名詞句・副詞句関連の連鎖が充実してくるということが言えそうである。

第2に、ネイティブ・スピーカーや Aarts & Granger の学習者コーパスの対象であった上級英語学習者などに比べて、中学生くらいの段階では極端に簡略した構造を多用している。例えば連鎖パタンに頻出する PRON V は He likes . . . などのパタンであるが、このような連鎖はネイティブや上級英語学習者のパタンの上位にはあまり現れない。彼らは通常 PRON AUX V のパタン、すなわち助動詞を伴った形を使う場合が圧倒的に多い。そういった意味で、最初はほとんど「主語＋本動詞」という単純パタンで出発しても、徐々に中間言語が発達して行くに連れて、法助動詞を用いた微妙なニュアンスの違いや時制や相の変化をきちんと表現するようになっていくということが言えよう。

中間言語の記述と習得理論との関連

以上の分析はまだまだコーパスの規模や品詞タグ付与の後処理などの精密化が必要であるので、

(p. 67 へ続く)

学習者コーパスと英語指導