The NICT JLE Corpus における名詞句構造の発達

小林 雄一郎　（法政大学 非常勤講師）

kobayashi0721@gmail.com

1. はじめに

近年、習熟度レベル別に分析可能な学習者コーパスが整備されつつあり、発表語彙における発達指標の特定が大きな関心を集めている。そして、高頻出語、品詞、単語 n-gram、品詞 n-gram、統語構造、談話標識、エラーなど、これまでも多くの発達指標が提案されてきた。

本研究は、日本人英語学習者による発話データにおける名詞句（NP）構造の発達に光を当てるものである。具体的には、The NICT JLE Corpus（和泉ほか 2004）に人手で構文情報を付与し NP 内のタグ連鎖をレベル別にクラスタリングする。

2. 先行研究

2.1. 母語話者（NS）発話における名詞句

Longman Spoken and Written English Corpus（LSWEC）に基づく文法書である Biber *et al.* によれば、NP は “determiner + premodifiers + head noun + postmodifiers”から成ると定義されている（1999: 574）。会話、小説、新聞、学術論文という 4 つのレジスター別に NP の修飾構造を分析すると、会話では、NP の約 85%は何の修飾も持たず、head noun のみで現れている。また、前置修飾を持つ NP と後置修飾を持つ NP の比率は同程度である（ibid: 578）。

2.2. 非母語話者（NNS）発話における名詞句

日本人英語学習者による発話コーパスを用いて、NP 構造の発達を統計的に調査した主な先行研究は 3 つある。

先ず、Kimura（2003）は、14 人分の NNS 発話データを用いて、NP 構造の発達を分析している。その分析にあたっては、Biber *et al.*（1999）に基づき、premodifier（4 タイプ）と postmodifier（6 タイプ）の分類法を提案している。

また、Kaneko（2006）は、4 段階にレベル分けされた NNS 発話データに NS 発話データを加えた計 5 つのサブコーパス（それぞれ約 5000 語）を用いて、NP の修飾構造（特に postmodifier）を調査している。

そして、三浦（2007）は、1281 人の NNS 発話データにおける品詞 n-gram を抽出した。このアプローチは、大規模データが解析できるという長所を持つ一方、n-gram 情報のみからでは詳細な句構造が分からないという短所もある。

3. 研究目的

本稿の目的は、レベルが上がっていくと、以下の 3 点がどのように変化するのかを調査し、今後の本格的な NP 研究に向けた基礎データを得ることにある。

- NP の平均長
- NP における前置修飾や後置修飾の長さとタイプ
- NP におけるタグ連鎖

4. データ

4.1. コーパス

本研究で用いるデータは、日本人英語学習者 1281 人の話し言葉データである The NICT JLE Corpus である。このコーパスには、SST（Standard Speaking Test）に基づく 9 段階の習熟度レベル情報が付与されており、167 人分のデータにエラータグが付与されている。

本稿では、そのエラータグ付きデータのレベル 3～8 から約 2000 語ずつを無作為抽出した約 12000 語を分析対象とする。なお、抽出にあたっては、対話者による発話の影響を排除するために、モノローグのタスクである stage 2（イラスト描写）と stage 4（ストーリー作り）を母集団とする。表 1 は、対象データの記述統計結果である。

表 1　対象データの記述統計結果

	Lv. 3	Lv. 4	Lv. 5	Lv. 6	Lv. 7	Lv. 8
Tokens	1982	2205	1941	2233	2140	2095
Types	513	566	523	523	587	503
STTR	33.20	33.85	32.50	32.60	35.35	30.75
MLP	6.27	7.30	9.07	9.42	11.38	10.12

4.2. 情報付与

句構造の分析には構文解析済みのデータが不可欠であるが、自動解析器を用いて、「学習者」による「発話」データを適切

な精度で構文解析することは難しい。そこで本研究では、Biber *et al.* (1999) に基づき、手作業で構文情報を付与する。また、品詞・文法項目の情報を付与するにあたっては、Penn Treebank のタグセット (Marcus *et al.* 1993) を用いる。

5. 結果と考察

5.1. NP の平均長

レベルが上がっていくと、NP の長さは発達していくのであろうか。以下の図 1 は、6 段階のレベル別データにおける NP の平均長を示したものである。

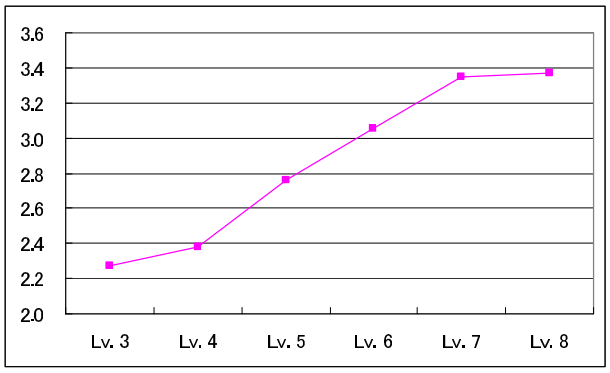

図 1　NP の平均長

図 1 を見ると、NP の平均長は、レベルが上がるにつれて長くなっていく。特に、レベル 4〜7 にかけて顕著に発達している。Kruskal-Wallis test の結果、NP の平均長には、レベル間で 0.1%水準の有意差が見られた (chi-square = 90.246, df. = 5, ***)。また、Steel-Dwass の方法による多重比較の結果、隣接するレベル間では、レベル 4 と 5 の間に 5%水準の有意差が見られた。このことから、NP の平均長がレベルを推定する指標となり得ることが分かる。

5.2. 前置修飾 vs. 後置修飾

前節では、レベルが上がるにつれて、NP の平均長が発達していくことが明らかにされた。では、どのような修飾構造の発達によって、NP は長くなっていくのであろうか。

以下の図 2 は、各レベルにおける ① 修飾を持たない NP (Head Only)、② 前置修飾のみを持つ NP (Pre Only)、③ 後置修飾のみを持つ NP (Post Only)、④ 前置修飾と後置修飾の両方を持つ NP (Pre + Post) の 4 種類の比率を集計し、100%積み上げ棒グラフに示したものである。

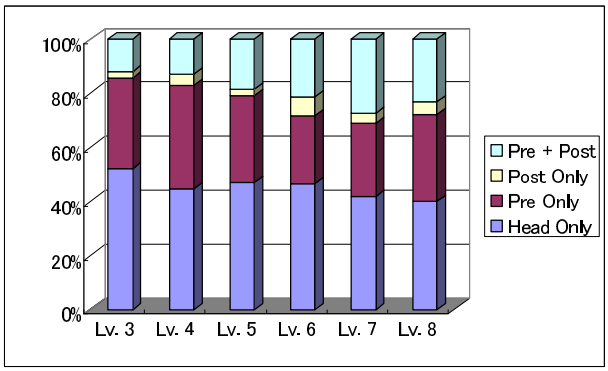

図 2　NP における修飾構造のタイプ

図 2 を見ると、NS 発話の場合 (Biber *et al.* 1999: 578) と同じく、何の修飾も持たない NP は、全てのレベルにおいて約半数を占めている。そして、レベルが上がるにつれて、前置修飾と後置修飾の両方を持つ NP の比率が上昇していく。後置修飾を持たない日本語を背景とする NNS は、英語を用いる際も後置修飾を避ける傾向があると言われる (e.g. Kimura 2003)。しかしながら、今回の結果を見る限り、日本語を母語とする NNS も、NS と同程度の比率で後置修飾を使用していることが分かる。

5.3. 前置修飾と後置修飾の平均長

では、レベルが上がっていくと、前置修飾 (Pre) や後置修飾 (Post) の長さは発達していくのであろうか。以下の図 3 は、その結果である。

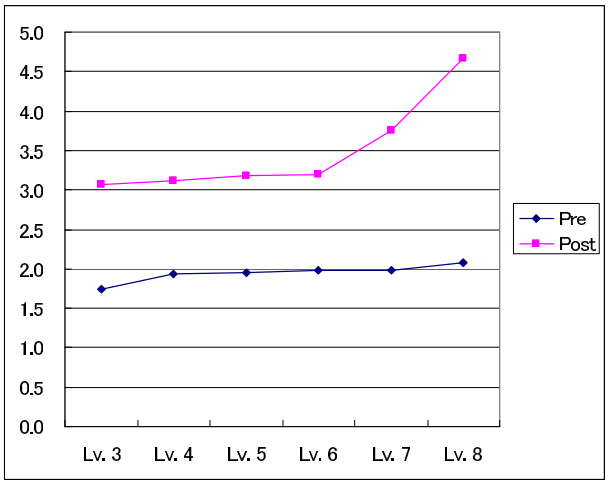

図 3　前置修飾と後置修飾の平均長

図 3 を見ると、前置修飾の長さはわずかながら発達するものの、Kruskal-Wallis test の結果、有意差は見られなかった。その一方、後置修飾の長さは、レベル 6〜8 にかけて顕著に発達する。Kruskal-Wallis test の結果、後置修飾の平均長には、

レベル間で 0.1%水準の有意差が見られた (chi-square = 35.051, df. = 5, ***)。

5.4. 前置修飾のタイプ

レベルが上がっていくにつれて、head noun を前方から修飾する要素は変化するのであろうか。以下の表 2 は、その結果である。

表2　前置修飾のタイプ

	Lv. 3	Lv. 4	Lv. 5	Lv. 6	Lv. 7	Lv. 8
Adjective phrase	55 42.97%	44 30.99%	55 40.44%	42 50.60%	48 47.52%	42 39.25%
Noun phrase	27 21.09%	31 21.83%	47 34.56%	15 18.07%	18 17.82%	34 31.78%
Genitive form	45 35.16%	62 43.66%	33 24.26%	25 30.12%	32 31.68%	29 27.10%
Particle clause	1 0.78%	5 3.52%	1 0.74%	1 1.20%	3 2.97%	2 1.87%
Total	128 100.00%	142 100.00%	136 100.00%	83 100.00%	101 100.00%	107 100.00%

表 2 を見る限り、前方修飾に関して、レベル間の差異を見出すのは難しい。

5.5. 後置修飾のタイプ

レベルが上がっていくにつれて、head noun を後方から修飾する要素は変化するのであろうか。以下の表 3 は、その結果である。

表3　後置修飾のタイプ

	Lv. 3	Lv. 4	Lv. 5	Lv. 6	Lv. 7	Lv. 8
Appositive noun	1 3.57%	1 2.13%	1 1.79%	0 0.00%	4 5.26%	4 5.97%
Prepositional phrase	23 82.14%	38 80.85%	48 85.71%	58 86.57%	57 75.00%	41 61.19%
To-clause	0 0.00%	2 4.26%	0 0.00%	0 0.00%	9 11.84%	3 4.48%
Particle clause	2 7.14%	1 2.13%	3 5.36%	2 2.99%	3 3.95%	10 14.93%
Relative clause	2 7.14%	4 8.51%	3 5.36%	6 8.96%	3 3.95%	8 11.94%
Adjective phrase	0 0.00%	1 2.13%	1 1.79%	1 1.49%	0 0.00%	1 1.49%
Total	28 100.00%	47 100.00%	56 100.00%	67 100.00%	76 100.00%	67 100.00%

表 3 を見ると、NS の発話と同様に、後置修飾の大半は前置詞句である。それ以外に注目すべき点は、レベル 3 から、一般に日本人英語が使用を避けると言われる関係代名詞や分詞による後置修飾が見られる点である。

5.6. タグ連鎖

NP 構造の発達を詳細にみるためには、品詞・文法項目タグの連鎖を見なければならない。しかしながら、単純な n-gram 分析では、NP の一部にしか光を当てることができない。そこで本研究では、n の数を設定せず、実際の NP の長さを保存したままの連鎖パターンを抽出した (e.g. “a cat” ⇒ “DT-NN”, “a lady getting on her chair” ⇒ “DT-NN-VBG-IN-PRP$-NN”)。以下、各レベルにおける連鎖パターンの上位 20 タイプ (本稿末尾の付録) を質的に分析する。

先ず、初級の学習者 (レベル 3～4) の場合、何の修飾も持たない“NN”の頻度が最も高い。だが、レベル 5 では“DT-NN”の頻度が頻度 1 位となり、そのあと“NN”の順位は徐々に下がっていく。このことは、日本人英語学習者による冠詞の習得過程 (e.g. 和泉ほか 2004: 131-139) を反映したものである (冠詞の頻度とレベルの順位相関係数は 0.886)。

次に、レベルが上がっていくにつれて、後置修飾を持つ NP が増加し、IN を含む連鎖パターンが増えていく (前置詞の頻度とレベルの順位相関係数は 0.826)。また、レベル 5～6 あたりから、前置詞句内に前置詞句が埋め込まれるパターン (e.g. “DT-NN-IN-DT-NN-IN-DT-NN”) も目立つようになる (前置詞句の発達については小林・葉田野 (2007) を参照)。そして、レベル 8 における IN を含む連鎖の減少は、関係代名詞や分詞による後置修飾の頻度の増加の影響を受けている。その他、タグ連鎖の頻度表には、初級での“bad fluency” (Kaneko 2006) とも言われる名詞の羅列 (e.g. “NN-NN-NN”) や、中級以上での double determiner の使用 (e.g. “PDT-DT-JJ-NN”) などが散見される。

6. 今後の課題

今後の課題としては、① データの改良と拡大、② 修飾構造の詳細な分析 (e.g. 埋め込みの深さ、位置)、③ 単語レベルの分析 (e.g. of を用いた後置修飾)、④ SLA 理論の検証、⑤ NS データとの比較、⑥ 教育的示唆などが挙げられる。

参考文献

[1] Biber, D., S. Johansson, G. Leech, S. Conrad, and E. Finegan (1999) *Longman Grammar of Spoken and Written English.* London: Longman.

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

[3] Kaneko, E. (2006) “Corpus-Based Research on the Development of Nominal Modifiers in L2.” Paper Given

at American Association for Applied Corpus Linguistics. October 21th, 2006. Arizona: Northern Arizona University.

[4] Kimura, M. (2003) "Japanese EFL Learners' Process of Noun Phrase Development: A Performance Analysis Using L2 Learners' Spoken Data." *Journal of Educational Research* 8: 61-67.

[5] 小林雄一郎・葉田野不二美 (2007) 「発話コーパスにおける前置詞句の発達」 『第 33 回全国英語教育学会大分研究大会予稿集Ⅱ』 pp. 135-138.

[6] Marcus, M. P., B. Santorini, and M. A. Marcinkiewicz (1993) "Building a Large Annotated Corpus of English: The Penn Treebank." *Computational Linguistics* 19: 313-330.

[7] 三浦愛香 (2007) 「英語学習段階と名詞句の内部構造発達」 英語コーパス学会 第 29 回大会 (2007 年 4 月 28 日, 同志社大学).

付録: NP における品詞・文法項目タグの連鎖

Rk	Lv. 3		Lv. 4		Lv. 5	
	Sequence	Freq.	Sequence	Freq.	Sequence	Freq.
1	NN	85	NN	61	DT-NN	81
2	DT-NN	81	DT-NN	56	NN	38
3	PRP$-NN	26	PRP$-NN	43	PRP$-NN	23
4	JJ-NN	23	NN-NN	11	DT-JJ-NN	11
5	NN-NN	12	JJ-NN	9	JJ-NN	11
6	NNP	9	DT-NNS	8	DT-NN-IN-DT-NN	9
7	DT-JJ-NN	8	NNS	8	DT-NNS	7
8	JJ-NNS	7	DT-JJ-NN	7	JJ-NNS	4
9	CD-NN	6	NNP	7	CD-IN-NN	3
10	CD-NNS	6	CD-NNS	6	CD-NNS	3
11	DT-NN-IN-DT-NN	5	DT-NN-NN	5	DT-JJ-NNS	3
12	DT-NNS	5	PRP$-NNS	5	DT-NN-CC-NN	3
13	CD-IN-NN	4	RB-JJ-NN	5	DT-RB-JJ-NN	3
14	DT-NN-NN	4	DT-NN-IN-NN	4	NN-IN-DT-NN	3
15	RB-JJ-NN	4	JJ-NNS	4	NN-NN	3
16	CD-CD-NN	3	NN-IN-NN	4	NNP	3
17	DT-NN-CC-DT-NN	3	NNP-NN	4	NNS	3
18	NN-CC-NN	3	DT-NN-IN-DT-NN	3	RB-JJ-NN	3
19	NNS	3	NN-POS-NN	3	CD-NN	2
20	PRP$-NN-NN	3	CD-IN-NN	2	DT-NN-IN-DT-JJ-NN	2

Rk	Lv. 6		Lv. 7		Lv. 8	
	Sequence	Freq.	Sequence	Freq.	Sequence	Freq.
1	DT-NN	74	DT-NN	70	DT-NN	77
2	NN	33	PRP$-NN	20	PRP$-NN	22
3	DT-JJ-NN	16	NN	19	DT-NN-NN	16
4	PRP$-NN	15	DT-JJ-NN	13	NN	15
5	NNS	11	JJ-NN	10	DT-JJ-NN	11
6	DT-NNS	6	DT-NNS	8	CD-NNS	5
7	DT-NN-IN-DT-NN	4	DT-NN-IN-NNS	4	DT-NNS	5
8	DT-NN-IN-PRP$-NN	4	DT-NN-NN	4	JJ-NNS	5
9	JJ-NN	4	NNS	4	PRP$-NN-NN	5
10	PRP$-NNS	4	DT-NN-IN-NN	3	NNS	4
11	CD-NNS	3	DT-NN-IN-NN-IN-NN	3	CD-NN	3
12	DT-NN-CC-NN	3	CD-IN-DT-NNS	2	DT-NN-IN-DT-NN	3
13	DT-NN-IN-NN	3	CD-NN	2	DT-NN-WP-VBD-VBG-DT-NN	3
14	JJ-NNS	3	DT-JJ-NN-NN	2	JJ-NN	3
15	NN-IN-DT-NN	3	DT-NN-IN-DT-NN	2	NN-NN	3
16	NN-IN-NN	3	DT-NN-TO-VB	2	CD-JJ-NNS	2
17	NNP-NN	3	DT-RB-JJ-NN	2	DT-JJ-NN-WP-VBD-VBG-DT-NN	2
18	CD-NN	2	NN-IN-DT-NN	2	DT-NN-NN-NN	2
19	CD-NNS-IN-DT-NN	2	PRP$-NN-IN-DT-NN	2	CD-CC-DT-NN-NNS	1
20	DT-CD-NN-NN	2	CD-IN-DT-NN-NNS	1	CD-CD-CD-NNS-VBG-IN-DT-NN	1