英語学習者コーパスにおける名詞句発達分析

金子恵美子

会津大学

E-mail: kaneko@u-aizu.ac.jp

あらまし 本研究では、学習者英語発話コーパス（NICT JLE コーパス）のデータを利用し、日本人英語学習者の名詞句の発達を調査した。予備的調査ではあるが、冠詞も含めた限定詞は初級学習者でも中上級者と同等の頻度で使用され、形容詞の使用は学習者においては運用能力とともに増加するが、NS では使用頻度が下がる。一方、後置修飾語句の使用は NS も含め運用能力とともに、使用が増加した。本研究はエラーを含む学習者コーパスを使用する将来の研究に示唆を与える。

1. はじめに

第二言語習得や外国語教授法、言語テストの分野において、コーパスが利用されるようになって久しい。特に、同質性が高く（母語が同一、学習環境が近似、など）、かつ習熟度分布に広がりがある学習者のデータで構築した学習者コーパスは、疑似時系列データ（quasi-longitudinal data）を提供し、様々なレベルの学習者の特徴を比較することで、中間言語の発達過程の調査が可能となる[1]。学習者コーパスのこの有用性のため、1990 年代に、Cambridge ESOL が学習者の英作文データ（のちに発話データも含む）の収集を開始し、Cambridge Learner Corpus (CLC)を作成する。そして 2000 年代にはアメリカの Educational Testing Services (ETS)も Test of English as a Foreign Language (TOEFL)のライティング、スピーキング問題の解答を収集し、TOEFL 2000 Spoken and Written Academic Language Corpus (T2K-SWAL)[2]を構築する。そして日本国内では対面式の英語スピーキングテスト、Standard Speaking Test (SST)の発話データを利用し、National Institute of Information and Communications Technology (NICT) Japanese Learners of English (JLE) コーパスが官民学共同事業により作成され[3]、インターネットからダウンロードが可能（https://alaginrc.nict.go.jp/nict_jle/）である。本稿では NICT JLE コーパスを利用し、日本人英語学習者の名詞句の発達過程に関する予備的研究を報告したい。

2. 使用データ

2.1. NICT JLE コーパス

本コーパスは Standard Speaking Test (SST, (株) アルク)というインタビュー形式の英語口頭運用能力テストによって引き出された発話を元データとして構築され、運用能力は初級の下から上級までの 9 段階に分類されている。SST インタビューは約 15 分、試験官と被験者が 1 対 1 の対面で行い、被験者が絵などを描写する monologue task と、試験官と会話を行う dialogue task に分かれる。

2.2. SST レベルと CEFR-J レベル

SST は全米外国語教育協会(American Council on the Teaching of Foreign Languages、ACTFL)が作成した、ACTFL Proficiency Guidelines - Speaking に基づいて評価される。一方、ヨーロッパでは欧州評議会によりヨーロッパ言語共通参照枠 (Common European Framework of Reference for Languages, CEFR)が外国語運用能力の各国共通の指標として制定された。その存在感は日に日に高まっており、日本でも日本版 CEFR-J が作成された[4]。本研究は、CEFR-J のレベル描写記述にコーパスに基づいた基準特性(criterial features) を割り当てる研究の一環として行われているため、本来の SST レベルを CEFR-J のレベルに置き換えて議論する。各レベルの対応は以下の通りである。

表 1　レベル対応

CEFR-J	SST	
PreA1	1	Novice Low
A1.1	2	Novice Mid
A1.2/3	3	Novice High
A2.1	4	Intermediate Low
A2.2	5	Intermediate Low+
B1.1	6/7	Intermediate Mid
B1.2	8	Intermediate High
B2.1-C2	9	Advanced

本研究では、NICT JLE コーパスの A2.1 (Intermediate Low) 以上の被験者の monologue task（一枚の絵の描写）の発話を利用した。また本コーパスには同様のインタビューの英語のネイティブ・スピーカー（NS）による発話データも収録されており、学習者との比較のために本研究で使用した。

2.3. 剪定データ

学習者の発話において、言い直し、繰り返し、沈黙などの、非流暢性は大きくコミュニケーションに影響し、故に非常に重要な口頭運用能力の指標となる。一方、高いレベルでは強調や相手の理解を促すため戦略的に繰り返し、言い直しを利用するため、一概にコミュニケーションを妨害するとは言い切れない。名詞の使用を調査する本研究において、言い直したり繰り返した名詞句も学習者の発した名詞句として分析するか、無用なものとして取り除いて分析するか、検討する必要があった。そのため、言い直し、繰り返し、つなぎ言葉を取り除いた剪定(pruned)データを作成し、剪定の影響を調べた。コーパスサイズ、名詞句の数を表 2 にまとめた。名詞句数はコーパスサイズ 5000 語で正規化してある。コーパスサイズ、名詞句数ともに、レベルが低いほど剪定による影響が大きく、特に A2 レベルでは名詞句数は半分以下になる。NS の変化率、-5%を基準にすると、A2 レベルの繰り返しや言い直しは効率的なコミュニケーションのために使われたのではなく、不十分な口頭運用能力に起因すると予想され、このレベルの学習者が発する名詞句の半数が繰り返しや言い直しによるものと推測される。

表 2　データ剪定結果

	NS	B2	B1.2	B1.1	A2
総語数	4363	4061	4857	4020	3890
剪定後	3666	3072	3867	2786	2273
変化率	-16%	-24%	-20%	-31%	-42%
名詞句数	701	567	557	468	844
剪定後	666	523	498	410	368
変化率	-5%	-8%	-11%	-12%	-56%

本研究の目的が母語話者ではなく学習者の、そして口頭運用能力一般ではなく名詞句の発達分析であることを考慮し、剪定データを利用することにした。

3. 分析手法

NICT JLE コーパスから対象レベルの絵の描写の発話を約 5000 語ランダムに抽出し、CLAWS part-of-speech tagger[5] により POS タグ付を行った。結果は 5000 語で正規化し、名詞の種類の分析では様々な名詞句がどのような比率で発生しているのかを明らかにするため、コーパスサイズで正規化した結果を名詞句総数 1000 語で再度正規化した。分析した名詞句の種類は以下のものである。

1. 冠詞+名詞(*a book*)
2. 指示形容詞+名詞（*this book*)
3. 数詞+名詞（*three students*)
4. 所有格＋名詞（*my car*)
5. 形容詞＋名詞（*beautiful music*)
6. 名詞+（副詞）+前置詞句(*the cat on the car*)
7. 名詞+（副詞）+現在/過去分詞（*the dog sleeping on the floor*)
8. 名詞+節(*the man I met*)

このうち 1 から 5 までは日本語の語順と同様に前置修飾語句を伴う名詞句、5 から 8 は後置修飾語を伴う名詞句で、日本人学習者にはより難易度が高い。これらを antconc[6]で検索し、後置修飾語句を伴う名詞句の結果は目視で不正確な結果を取り除いた。学習者の運用能力に特徴的にみられる名詞句を調査するため、カイ二乗検定を行った。またカテゴリーデータにおける post hoc テストとして分割カイ二乗 (partitioning Chi-square)[7]を行った。分割カイ二乗検定は、研究者の仮説に基づ

きデータを組み合わせて新しいカテゴリーを作成するときに使用する。

4. 結果

NS と学習者のそれぞれの名詞句の数は下記の通りである。

表 3　名詞句分布(名詞 1000 語で正規化 [1])

名詞句種類	NS	B2	B1.2	B1.1	A2
1	410	373	377	351	368
2	62	57	66	66	60
3	44	86	68	76	95
4	27	61	28	49	49
5	119	147	145	127	119
6	84	67	42	33	8
7	29	31	16	7	5
8	27	36	7	21	8

[7]のやり方に従い、冠詞+名詞(名詞句種類 1)と指示形容詞+名詞（種類 2）の間に有意差がないと仮定し、カイ二乗検定を実施した結果、有意差は認められなかった（χ^2 (4, N = 472)= .76)。次に、数詞+名詞(種類 3)と所有格＋名詞(種類 4)の分布に有意差がないと仮定し、カイ二乗検定を実施した結果、有意差は認められなかった(χ^2 (4, N = 583)= .33)。この結果「冠詞/指示形容詞＋名詞（1+2)」「数詞/所有格＋名詞（3+4)」という新カテゴリーを作成した。次に、後置修飾語＋名詞という新しいカテゴリーを作成できるか調べるため、後置修飾語句を伴う名詞句 3 種類（6,7,8）に対しそれぞれの組み合わせでカイ二乗検定を α =.017 に調整して 3 度行った。その結果、分詞と関係詞を伴う名詞句（7 と 8）の分布には有意差がなかった（χ^2 (4, N = 187)= .026)ため 1 つのカテゴリーとした。修正後の名詞句分布は表 4 のようになった。

5. 考察

表 4 が示す通り、学習者のみに限ると、形容詞＋名詞（名詞句種類 5）はレベルが

[1] 各レベルの名詞句合計が 1000 に満たないのは、名詞の前後に修飾語句が付随していない名詞が存在するためと推測される。

表 4　修正後名詞句分布

名詞句種類	NS	B2	B1.2	B1.1	A2
1+2	472	430	443	417	408
3+4	71	147	96	125	144
5	119	147	145	127	119
6	84	67	42	33	8
7+8	56	67	23	28	13

上がるにつれて頻度が増すが、NS の使用頻度は A2 レベルと同じである。また冠詞を使用するのは日本人英語学習者には難しい印象があるが、冠詞＋名詞(種類 1)と指示形容詞＋名詞（種類 2）の間に有意差はなく、この二種の名詞句から作成した新カテゴリー（1+2）の分布は NS も含め大差がない。以上から、正確さはさておき比較的学習段階初期から冠詞、指示形容詞、数詞、所有格など限定詞(determiners)の使用は見られ、形容詞の使用はレベルが上がるにつれて使用が増す。NS は後置修飾を使用して形容詞よりも更に詳細な修飾が可能であるため、形容詞の使用頻度は減ったと考えられる。

次に後置修飾語句の分布を見てみる。名詞+前置詞句(例：*the woman with long hair*, 種類 6）の使用頻度は、レベルが上がるにつれてより頻繁にスピーキングで使用されることがわかる。名詞+前置詞の連続は、正規表現の検索式では修飾関係にあるものだけを取り出すことができず、目視で修飾関係にはないものを除外したが、その結果の減少率は半分から 12%程度と、誤差とは言えないものであった。名詞+現在/過去分詞（例：*the girl sending mail, a road called Station Way*, 種類 7）と名詞＋関係詞節（例：*a person who fell off the ski*, 種類 8）は意味的にも形式的にも似通っているため、分布に有意差なしと仮説を立てカイ二乗検定を行い、その結果カテゴリーを統一した。しかしながら、本研究では関係代名詞節を伴わない関係詞節（例：*the computer screen the girl is looking at*）を検索できていない。関係代名詞を省略できるのは、先行詞が関係詞節の目的語にあたるときで、このような目的語型の関係詞

（例 *the man Tom met*）は、先行詞が関係詞節の主語である主語型(例 *the man who met Tom*)よりも学習者には難易度が高いことが知られており[8, 9]、NS の名詞句 7+8 の使用が B2 より少ないのは、この検索方法に起因する可能性がある。すべての関係詞節が検索結果に反映されると、B2 や NS で名詞句種類 8 の数が増加することが予想され、カイ二乗検定の結果も違ったものになるかもしれない。

6. 結論

以上の結果をまとめると、1) 限定詞は学習の初期段階から使用され、頻度に関しては習熟度が上がってもあまり変化はない、2)形容詞は学習者レベルが上がると使用頻度が上がる、3)後置修飾語句は NS も含めて、運用能力が高いと使用頻度が上がる。

本研究は予備的研究であるが、今後の学習者発話コーパスを利用しての同様の研究に示唆を与える。学習者発話コーパスの場合、自動検索で検索できる言語単位には限界があった。本研究では前置修飾語句を伴うものと、後置修飾語句を伴うものを別々に検索したため、両方が発生している名詞句は二重でカウントされてしまう。限定詞や形容詞、後置修飾語句もない「裸」の名詞句を検索するのは困難で、名詞の総数から検索できたものを引いて算出しても、上記の理由により実際よりも少なく見積もられる。また考察のところで述べた通り、関係代名詞を伴わない関係詞節の検索ができなかった。これらは正規表現を工夫し、時間をかければ解決できるのかもしれないが、そもそもエラーを多く含む学習者発話コーパスの POS タグの正確さが確立できず、そのため検索の信頼性も危うい。またエラーに関して、後置修飾語句を伴う名詞句、特に関係詞など長い名詞句は正確に発話できたもののみをカウントするのか、不正確でもカウントするのか、またエラーにより名詞句とも文章とも判断がつかない場合はどう対処するのか、など、言語処理分野の範疇を出て、コーパスを読み判断する「人」による作業が不可欠であるように思われるが、その処理には膨大な時間と人力を要する。本研究の結果と目視確認から、直観的には B1.1 以上は機械分析を行ってもその不正確さは誤差として扱えるような印象だった。一方、A2 は日本人英語学習者の中に最も多く存在するレベルであり、特にこのレベル以下では機械のみに頼らず、人による分析が必要と思われる。

参考文献

[1] S. Granger, "A bird's-eye view of learner corpus research," in *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, S. Granger, *et al.*, Eds., Amsterdam: John Benjamins Publishing, 2002, pp. 3-33.

[2] D. Biber, et al. (2004, Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus. *TOEFL Monograph Series*. Available: http://www.ets.org/Media/Research/pdf/RM-04-03.pdf

[3] 和泉絵美他., 『日本人 1200 人の英語スピーキングコーパス』. 東京: アルク, 2004.

[4] 投野由紀夫, 『CAN‐DO リスト作成・活用 英語到達度指標 CEFR‐J ガイドブック』. 東京: 大修館書店, 2013.

[5] UCREL. (2010) CLAWS part-of-speech tagger for English ver. 6. Available: http://ucrel.lancs.ac.uk/claws/

[6] L. Anthony. (2005). Antconc ver, 3.4.3. Available: http://www.laurenceanthony.net/.

[7] D. Rindskopf, "Partitioning chi-square: Something old, something new, something borrowed, but nothing BLUE (just ML)," *Categorical variables in developmental research: Methods of analysis*, pp. 183-202, 1996.

[8] C. Doughty, "Second language instruction does make a difference: Evidence from an empirical study of SL relativization," *Studies in Second Language Acquisition*, vol. 13, pp. 431-469, 1991.

[9] K. Hyltenstam, "Typological markedness as a research tool in the study of second language acquisition," in *Current Trends in European Second Language Acquisition Research*, H. Decher, Ed., Clevedon: Multilingual Matters Ltd, 1990, pp. 23-36.