生産性の漸増が語る統語知識の発達:
パターン束モデルに基づく段階的発達プロセスの計算的実証

吉川 正人 (慶應義塾大学 [院]/日本学術振興会特別研究員)

1. はじめに

用法基盤モデル (e.g., Langacker 1987) に基づく大規模な言語習得研究としては Tomasello (2003) が著名だが，そこで行われているのは行動データや実験結果に基づく「発達的事実の記述」であり，幼児が文法を習得しているという時，知識として何を習得しているのか，つまり，「知識の表示」に関する (直接的な) 研究は行われていない．Tomasello (2003) がこの表示の候補として想定しているのは「構文文法 (Construction Grammar: e.g., Croft 2001; Goldberg 1995)」の言うところの「構文 (constructions)」であるが，それがいかなる実体であるかというのは構文文法流の表示を用いた簡易的な提示があるのみで，定式化は成されていない．

本発表では，Tomasello (2003) の言うような，具体的な一語文 (Holophrases) から始まりスロットを持つ二語文・多語文 (e.g., 軸スキーマ)，そして抽象的な構文 (e.g., 二重目的語構文) と段階的に生産性・抽象性を獲得していく統語発達の「プロセス」に焦点を当て，幼児の産出データから可能な統語知識の「内部表現 (internal representation)」を計算的に構築することによってこのプロセスの実証を目指す．

具体的には，CHILDES データ (MacWhinney 2000) 内の Brown コーパス (Brown 1973) を用い，幼児の発話から年齢経過に従って漸増的にパターンを生成し，逐次得られたパターンの生産性を算定することによって，徐々に統語知識の生産性が上昇していくことを示す．

2. 前提

2.1. パターン束理論によるパターンの定義

パターン束理論 (e.g., 黒田・長谷部 2009, 以降 PLT) では，記憶の単位である任意の事例 e (e.g., 文: *John hit Mary.*) を任意の分節モデル T (e.g., 単語分節) によって分節化し分節列 $T(e)$ (e.g., [John, hit, Mary]) を得，$T(e)$ の分節を全ての分節が変項になるまで一つずつ再帰的に変項化を行い，得られた集合 P を e のパターン集合 $P(e)$，その構成要素をパターン $p \in P$ と定義する．

パターンは部分一致に基づく継承関係 (`is-a`) の階層構造を持つ．このような継承関係の既定されたパターン集合 P は半順序集合パターン束 (Pattern Lattice) L を構成する．尚，任意の事例 e から得られたパターン束を $L(e)$ と表記することとする．継承関係は推移律を満たすので，冗長な関係を除いた継承関係をハッセ図により図示すると Fig. 1 のようになる．Fig. 1 では，簡略化のため連続する変項を単一の変項に縮約するという処理を行っている．

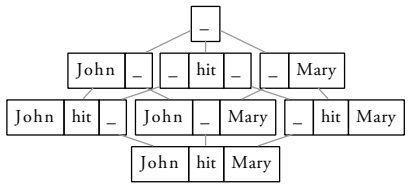

Fig. 1 e = *John hit Mary* の $L(e)$
(Pattern Lattice Builder (黒田・長谷部 2009) で作成)

2.2. パターンの発達としての統語発達

PLT の定義するパターンがヒトの統語知識の実体であるならば，幼児の統語発達とは，このパターンの体系の発達に他ならない．「パターンの体系の発達」には，以下の二つの側面があろう:

(1) a. 利用可能なパターンの総数の増加
b. 生産性や構造の複雑性など，個々のパターンの性質の成長

構文文法，特に急進的構文文法 (Radical Construction Grammar: e.g., Croft 2001) の想定によれば，言語知識は全て大小様々な規模からなる構文であるということになる．また，Tomasello (2003) が正しければ，幼児は少なくとも発達の初期段階では高度に抽象的な "Subj Verb Obj_1 Obj_2" のような構文を知識としては持ち合わせておらず，より具体的な，*I-wanna-do-it* といった一語文，*More* _ と言った二語文，*Put* _ *in* _ と言った多語文を経て抽象的な構文に行きつくパターンを辿ることになる．

従って，幼児の発話からその時点での統語知識を逆算して考える際には，成人と同様な抽象的な構文の存在を前提とせず，より具体性の高い，語彙項目に依存した形での構文を想定するべきである．さらに，軸スキーマ (Pivot Schemas)，語結合 (Word Combinations)，語彙依存構文 (Item-based Constructions) と言った Tomasello (2003) の分類は，異なる発達段階で発現する質的に異なった構文であると考えられるが，その発現時期・順序には個人差があり，また一方が他方に取って替わられるような性質のものでもない．それ故，そのような構文の質的分類も論点先取にならないように留意する必要があり，「平等」にパターンを生成し，そこから何らかの選定作業や重みづけ等を行うことで，派生的に種々の構文的性

質を導きだせるようにするのが望ましい．

以上を考慮して，本研究では，「構文」の認定に外在的な評価基準を持ち込まず，幼児の発話から網羅的にパターンを生成し，後に述べる単純な頻度の基準のみで選定したパターンに対し分析を実施する．

また，パターンの発達の指標としては，(1a) に挙げたパターンの総数ではなく，(1b) に挙げた個々のパターンの性質の変化を見る．具体的には，後に述べる方法で算定した，パターンの「生産性 (productivity)」を利用する．

3. 調査

3.1. データ

データには，CHILDES データ (MacWhinney 2000) 内の Brown コーパス (Brown 1973) を利用した.[1)] このコーパスには Adam，Eve，Sarah という 3 幼児のデータが収録されている．

Table. 1　Brown コーパスの詳細

Child	**Age Range**	**Files**
Adam	2;3–3;4[4;10]	28/55
Eve	1;6–2;3	20
Sarah	2;3–3;8[5;1]	71/139

尚、今回はデータ量の多い Adam と Sarah に関しては約半数のデータのみを利用している．これは，データ量の多さが計算量の増加を招き非常に処理負荷がかかるという技術的な要因と，あまり長期間に渡るデータの場合変化が劇的で長期スパンでみた場合に有意な変化傾向が出ることが当然視されるため，比較的短い期間でデータを見ることが有意義であると判断したことによる．データの詳細は表 1 に示す通りである (より詳細な情報は http://childes.psy.cmu.edu/manuals/02englishusa.doc を参照)．

3.2. 方法

上記データから，i) 幼児の発話のみを抜き出し，ii) 言いさしや重複，ポーズの含まれる発話を除外したものを入力データとして用意した．表 2 に，このような前処理を行った後の Eve のデータの詳細を提示する．ファイルの id は時系列順に付与されており，数値が低いほど年齢が低く，大きいほど高い．

以上のように前処理を行ったデータに対し，以下のような手順で生産性の算定を行った:

Table. 2　前処理後の Eve のデータ

file	#sent	Mean Length	vocabulary	age
1	547	1.99	160	1;6
2	338	1.73	151	1;6
3	155	2.02	97	1;7
4	359	1.77	162	1;7
5	347	2.07	158	1;8
6	267	2.83	204	1;9
7	356	2.47	211	1;9
8	483	2.8	290	1;9
9	277	3.04	213	1;10
10	310	3.07	212	1;10
11	245	2.93	191	1;11
12	350	3.45	253	1;11
13	288	3.53	241	1;12
14	257	3.17	215	2;0
15	398	3.54	293	2;1
16	356	3.65	292	2;1
17	469	3.67	396	2;2
18	470	3.47	343	2;2
19	417	3.51	333	2;3
20	499	2.43	274	2;3

(#sent: 発話数; Mean Length: 平均文長 (語数))

(2)　a. 1 ファイル目のデータから PLT の定義に従ってパターンを生成する;
　　b. (2a) の結果から，頻度 (= 対応する事例の総数)・事例のバリエーション (= 対応する事例の形式の異なり数) 双方が 2 以上のパターンのみに対し，3.3 節に示す方法で生産性を算定し，その平均を求める;
　　c. 2 ファイル目のデータからパターンを生成し，(2a) の結果と結合する;
　　d. (2b) 同様，(2c) で得られたパターンの生産性を算定し，平均を求める;
　　e. 以上の工程を最後のファイルまで繰り返す

この作業によって，パターンの生産性が年齢を経る毎に増加するプロセスを定量的に検証することが可能となる.[2)] (2b) について一点追記しておくと，事例のバリエーションが 1(つまり対応する事例の異なりが 1) でも，変項を全く含まないようなパターン (e.g., “where go”) は除外せずに残してある．

ただし，この方法で生産性を計算し結果数値がファイルを追加する毎に増加していることが確かめられたとしても，その増加は単に生産性を計算するのに利用したデータの量が増加したために生じたものであって，パターンの生産性が増加したために生じたわけではないという可能性が否定できない．実際，データの量が増えればそれだけ表現のバラエティが増えることは十分に予測でき，表現のバラエティの増加はそこから生成されたパターンの生産性の上昇に直結する．

この問題を解決するには，上に示した時系列に

[1)] CHILDES とは Child Language Data Exchange System の略で，幼児と周囲の大人 (主に養育者) の自然な対話のデータベースである．共通のフォーマットで記録された，様々な言語の様々なコーパスが含まれる (参考: http://childes.psy.cmu.edu/)．中でも今回利用した Brown コーパスは，データ量や収録期間の長さから，様々な研究で利用されている．

[2)] 以上の工程及び後に述べる生産性の計算は全てスクリプト言語 Python (ver. 2.6.5) を用いて作成した独自のプログラムを使用して行った．

沿った漸増的なパターン生成に加えて，ファイルの順序をランダム化したデータを複数用意し，時系列に無関係な単純なデータ量の増加によってではパターンの生産性は上昇しないこと，もしくは，時系列に沿ってデータを漸増させた場合とは異なる増加の傾向を示す，ということを示す必要がある．そこで本調査では，時系列に沿った順序になっているデータに加え，順序をランダム化したデータを 50 パターン用意した．

3.3. 生産性の算定

パターンの生産性の指標としては，シャノンのエントロピー (H) を利用した.[3] 具体的には，変項ごとにエントロピーを計算し，複数の変項を持つパターン (e.g., “put _ in _”) に関しては i) 各変項のエントロピーを足し合わせ，ii) その後変項間の共変動率を見積り，iii) 最後にその分を調整することでパターン p 全体のエントロピー H(p) を計算している.[4]

4. 結果と考察

4.1. 結果概要

調査の結果，どの幼児も年齢を経る毎に (= 入力ファイル数の増加に従って) ほぼ線形に生産性の平均が上昇していくことが確かめられた (Fig. 2，参照; 縦軸 (H_ave) が生産性の平均，横軸 (Scale) がその時点での入力データ量 (発話数)).

Table. 3　各時点の生産性上位 10 パターン (Adam)

rank	1	H	14	H	28	H
1	Adam _	4.96	I _	7.21	I _	10.01
2	my _	4.61	xxx _	6.96	you _	8.13
3	I _	4.52	put _	6.94	dat _	8.01
4	_ paper	3.81	Adam _	6.91	_ it	8.01
5	get _	3.73	_ it	6.81	xxx _	7.80
6	_ go	3.63	dat _	6.65	I going _	7.68
7	no _	3.59	where _	6.41	do _	7.66
8	Mommy _	3.59	_ there	6.29	a _	7.61
9	_ there	3.59	_ in there	6.29	it's _	7.39
10	go _	3.57	let _	6.09	put _	7.38

(1, 14, 28 はそれぞれ追加ファイル数)

増加の傾向を確かめるにあたって，本調査では線形回帰を利用した．線形回帰は，各データ追加時点でのパターンのエントロピーの平均値を目的変数，その時点でのデータ量 (= 発話数) を説明変数として行った.[5] Fig. 2 から明らかなように，どの幼児も非常に高い線形の傾向を見せており，回帰の当てはまり度合いを示す決定係数 (R^2) も高い値を示している (Adam: R^2 = 0.96; Eve: R^2 = 0.95; Sarah: R^2 = 0.98; R^2 が 1 に近いほど回帰の当てはまり度合いが高いことになる).

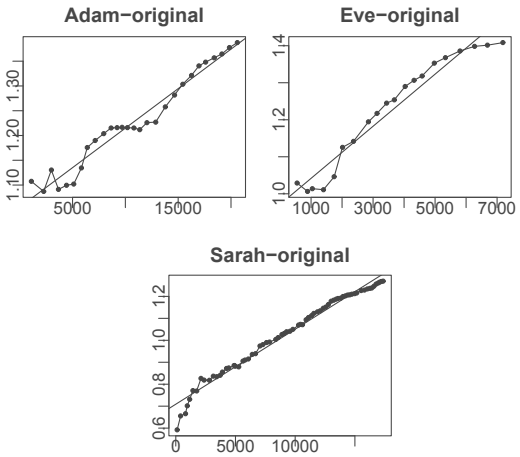

Fig. 2　3 幼児の元データの散布図
(直線は回帰直線)

時系列に沿ってデータを追加した元データ (以降 O) と順序をランダム化したデータ (以降 R) との比較にあたっては，R の諸データに対しても線形回帰を行い，回帰直線の傾きを求めた．ただし，回帰直線の傾きが大きかったとしても，回帰の当てはまり度合いが低かった場合，生産性増加のパターンが線形とは言えないということになる．従って，今回は $R^2 \geq 0.5$ のもののみを選定した．この傾きの一覧を Table 4 に示す．便宜上傾きは 100000 倍にしてある．

Table 4 から，O の傾きと R の傾きとの差は顕著であり，比較的傾きが大きいもの (e.g., Adam の random45) は決定係数がさほど大きくなく，結果的に O との差異は大きくなっていると言える (Fig. 3 も参照)．従って，R の諸データは O ほどの増加傾向は見られず，また増加のパターンも異なっていると結論付けられる．

最後に，O と R の傾きの差が有意であるかどうか，1 標本の t 検定で検証した．具体的には，O が特別な特性を持たない，ランダムな順序でデータを追加したものの一種であるという帰無仮説を想定し，R の傾きの平均と標準偏差に対し，O の傾きが有意に離れているかどうかを検定した．結果，どの

[3] シャノンのエントロピーとは，確率空間上の事象の生起確率に基づく情報量のことで，パターンの変項 v のエントロピー $H(v)$ は，v の i 番目の実現値を w_i，その全体に占める割合を $p(w_i)$ として，以下のように求められる:

$$H(v) = -\sum_{i=1}^{m} p(w_i) \log_2 p(w_i) \quad (1)$$

[4] 計算の詳細は吉川 (2010) を参照されたい．

[5] 線形回帰及び散布図のプロットは，統計言語 R の Python インターフェースである “rpy2” モジュール経由で R の関数を利用し実施した．

Table. 4　$R^2 \geq 0.5$ の回帰直線の傾き

Adam	slope	R^2	Sarah	slope	R^2	Eve	slope	R^2
org	1.59	0.96	org	7.04	0.94	org	3.42	0.98
r04	0.31	0.50	r05	3.30	0.64	r01	0.90	0.75
r05	0.98	0.77	r06	1.29	0.56	r02	0.41	0.55
r08	0.43	0.63	r07	3.36	0.59	r07	0.84	0.51
r10	0.72	0.57	r09	3.75	0.53	r12	0.78	0.61
r12	0.73	0.62	r12	2.05	0.67	r16	0.71	0.59
r13	0.99	0.65	r16	0.81	0.55	r26	0.67	0.83
r16	-0.25	0.60	r17	4.10	0.58	r27	1.19	0.79
r19	0.67	0.59	r20	1.92	0.56	r28	1.23	0.63
r20	-0.27	0.69	r25	3.44	0.57	r29	0.69	0.80
r21	-0.21	0.58	r29	1.83	0.73	r35	1.03	0.55
r28	0.93	0.74	r36	3.50	0.87	r36	0.44	0.57
r29	0.40	0.54	r39	1.76	0.59	r37	0.78	0.53
r33	0.92	0.82	r40	1.21	0.69	r39	0.85	0.53
r45	1.16	0.57	r44	3.39	0.53	r43	0.40	0.64
r47	0.52	0.52	r45	1.84	0.64	r44	1.06	0.56
			r49	1.32	0.60	r49	1.41	0.53
						r50	0.81	0.69

(org: 元データ; r: ランダム)

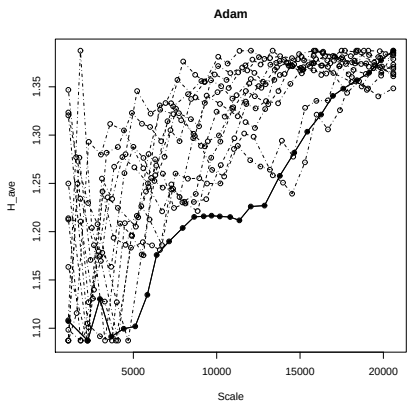

Fig. 3　$R^2 \geq 0.5$ かつ傾き ≥ 0 のデータ (Adam)
(実線は O，破線は R)

幼児も O の傾きは R の傾きから有意に離れており，生産性の上昇傾向が有意に異なっていることが確かめられた (Adam: t(49) = -20.4534, p < 2.2e-16; Eve: t(49) = -32.5039, p < 2.2e-16; Sarah: t(49) = -55.1863, p < 2.2e-16).

4.2.　考察

本来回帰分析というのは，ある目的変数 (今回の場合エントロピーの平均値の変化) がある説明変数 (今回の場合データ量の増加) によってどれくらい説明できるかを分析するものであり，時系列に沿ってデータを漸増させた元データが線形回帰において高い決定係数を示すということは，エントロピーの平均の増加がデータ量の増加によって説明できる，ということを意味する．

しかしながら，重要な点はデータの追加順序をランダム化した場合には決定係数が明らかに落ち込むということであり，エントロピーの平均の増加は単純なデータ量の増加によって引き起こされるものではないということである．従って，今回の回帰分析の示す結果は，データ量の増加の裏にある見えない尺度とエントロピーの平均の増加との強い相関である．データの追加順がランダムになると一見データ量は同じように増加しているように見えても，その裏にある「見えない尺度」の増加の傾向が崩れ，結果的に回帰直線の傾きと決定係数の低下が生じているということである．

以上から，本調査で示されたエントロピーの増加にみるパターンの生産性の漸増は，Tomasello (2003) 等が主張する幼児の段階的な統語発達のプロセスをうまく捉えているものと考えられる．

4.3.　本調査の意義

今回行った年齢経過に伴う生産性の漸増プロセスの検証は，統語発達の一側面を示したものに過ぎない．特に，意味に関連する情報やパターンの性質など，いくつかの重要な情報を考慮していない点は問題を孕むかもしれない．

しかしながら，今回の調査の意義として重要なのは，PLT によってもたらされるこのような計算手法を用いれば，発達度合いの指標を変えるだけで，様々な観点から計算的に統語の発達「プロセス」を実証できる，ということである．このことが，理論の予測力向上，様々な周辺分野 (e.g., 自然言語処理，計算言語学) との親和性向上など，認知言語学に数多くの利益をもたらすのは明らかである．

参考文献

Brown, R. 1973. *A first language: The early stages.* Cambridge, MA.: Harvard University Press.

Croft, W. 2001. *Radical construction grammar: Syntactic theory in typological perspective.* Oxford; New York: Oxford University Press.

Goldberg, A. 1995. *Constructions: A construction grammar approach to argument structure.* Chicago; London: University of Chicago Press.

黒田航・長谷部陽一郎. 2009. Pattern Lattice を使った (ヒトの) 言語知識と処理のモデル化. 言語処理学会第 15 回大会発表論文集 (pp. 670–673).

Langacker, R. 1987. *Foundations of cognitive grammar vol 1.: Theoretical prerequisites.* Stanford: Stanford University Press.

MacWhinney, B. 2000. *The CHILDES project: Tools for analyzing talk.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Tomasello, M. 2003. *Constructing a language: A usage-based theory of language acquisition.* Cambridge, MA.: Harvard University Press.

吉川正人. 2010. パターンの生産性に見る統語発達: パターン束モデルに基づく習得プロセスの検証. 日本認知科学会第 27 回大会発表論文集.