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
Syntactic development as increasing productivity of patterns:
Investigating the acquisition process of language based on Pattern Lattice Model

吉川 正人[†,‡]
Masato YOSHIKAWA

[†]慶應義塾大学大学院，[‡]日本学術振興会特別研究員
Keio University, JSPS Research Fellow

machayoshikawa@dream.com

Abstract

This paper presents a quantitative research in which the gradual process of language acquisition assumed by the *usage-based* theory (e.g., [9]) is verified under an exemplar-based formal model known as *Pattern Lattice Model* (PLM). Specifically, using transcribed speech data produced by three infants included in a corpus[3] in CHILDES database[8], i) (syntactic) patterns were generated, ii) productivity of each pattern was computed, and iii) for each of three infants, the averages of productivity were compared among the three developmental periods. As a result, productivity of patterns is found to increase gradually with age, which verifies the assumption of usage-based view.

Keywords — Pattern Lattice Model (PLM), language acquisition, productivity, Usage-based Model of language

1. はじめに

「用法基盤 (Usage-based)」の習得モデル(e.g., [9])では幼児は具体的な一語文から始まり徐々に抽象性・一般性を獲得していく段階的な文法獲得プロセスを経ると考えられる．このプロセスは，例えばBorensztajnらによる「データ指向構文解析 (Data-Oriented Parsing, DOP)」(e.g., [1])の枠組みを用いたCHILDES データベース[8]内のコーパス分析[2] などで実証的・定量的な検証が行われている．

本稿では，このような検証の一つとして，「パターン束モデル (Pattern Lattice Model, PLM)」[4, 6, 7, 10]を用いた幼児発話の分析を提示する．具体的には，PLM の定義する「パターン」の生産性を測定し，年齢を経るに従って生産性が増加することを示す．本稿の目的は，1) 用法基盤の統語発達プロセスの定量的・統計的検証を強化すること，及び，2) この検証を通してPLMの妥当性を示すことである．

2. パターン束モデル (PLM)

パターン束モデル(PLM)は，[7]によって提案されたヒトの言語知識のモデルである．PLMでは，ヒトの言語知識は具体的な言語事例とその「索引 (indices)」の集積であると看做され，言語形式に関する知識 (≈ 統語知識)は以下で定義するパターンの体系であると考える．パターンは言語事例に対する形式的な索引とされる．

PLM では，具体的な言語事例e (e.g., (1a))の適切な分節化$T(e)$ (e.g., (1b))に対し単一の分節を変項Xで変項化したものをパターンと定義し，この変項化を網羅的に行い得られたパターンのべき集合 (e.g., (1c))を事例eのパターン集合$P(e)$ と定義する [7, pp. 670-671].

(1) a. I am a boy.

b. [I, am, a, boy]

c. {(I, am, a, boy), (_, am, a, boy), (I, am, a, _), (I, am, _, boy), (I, _, a, boy),(_, am, a, _), (_, am, _, boy), (_, _, a, boy), (I, am, _, _), (I, _, a, _), (I, _, _, boy), (_, am, _, _), (_, _, a, _), (_, _, _, boy), (I, _, _, _), (_, _, _, _) }

分節化のモデルは独立の基準で評価される必要があるが，本稿では，「単語分節 (word-segmentation)」と同一視する．

$P(e)$は任意の二つのパターン$p_i, p_j \in P(e)$において，(2)の関係を満たす場合に$[p_i \texttt{ is-a } p_j]$となる半順序集合，「パターン束(Pattern Lattice)」$L(e)$を構成する([7, p. 671], [4, p. 412]):

(2) n個の分節を持つ$p_i = [s_{i1}, s_{i2}, \ldots, s_{in}]$，$p_j = [s_{j1}, s_{j2}, \ldots, s_{jn}]$の$k$番目の分節をそれぞれ$s_{ik}$, s_{jk}として，全ての$k\,(1 \leq k \leq n)$において，

a. $s_{ik} = s_{jk}$

b. s_{jk} = 変項

のいずれかが当てはまる場合

パターン束の階層*rank*は非変項＝定項の数で定義される．束の頂点は*rank*＝0のパターン，底は事例である．

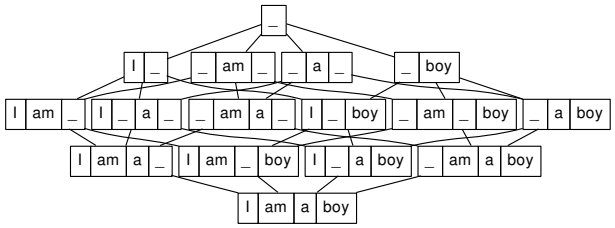


図 1 *e* = *I am a boy*の場合の$L(e)$
(Pattern Lattice Builder [7]で作成)

n個の事例の集合$E = \{e_1, e_2, \ldots, e_n\}$に対する$L(E)$は個々の事例$e_i$に対する$L(e_i)$を結合したものであるが，その際には，長さ＝分節数の異なる事例間のLには互換性がないため，連続する変項を単一の変項に縮約するという処理を行う.[1] このようにして得られたパターンの集合から，何らかの尺度でパターンの「有用性」を測定し，ある閾値を超えたものを「良いパターン(good patterns)」と認定する．

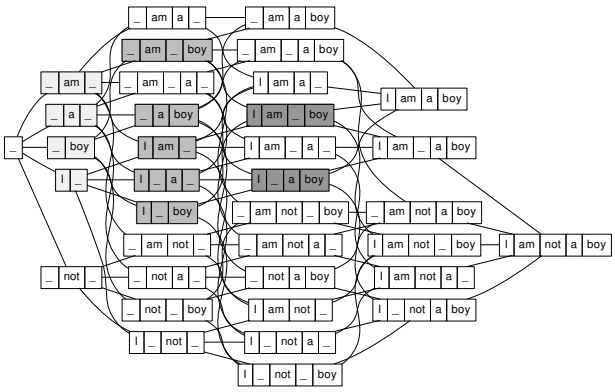


図 2 E ={*I am a boy*, *I am not a boy*}の場合の$L(E)$
(Pattern Lattice Builder [7]で作成)

ここで言うパターンの「良さ(goodness)」とは，そのパターンの(資源としての)「再利用可能性(recyclability)」のことである．多くの事例に共有されるパターンは，それだけi) 既に再利用の実績を持ち，ii)それ故に以後も再利用されるポテンシャルが高いパターンということになる.[2]

3. 仮説

幼児の言語発達の実態を鑑みると，PLMの想定から，以下の仮説が導かれる:

(3) a. 幼児の統語発達とは，PLMの定義する「パターン」の発達である．
b. 補足: パターンの発達とは，
i) 利用可能なパターンの総数が増え，
ii) 個々のパターンがより生産的に複数の表現に流用できるようになる
ことである．

例えば幼児が「一語文(Holophrases)」など変項を含まない具体的で短い表現から産出を始めるという事実は，上に述べた「良いパターン」と認定される条件から，よく使われる決まり文句(e.g., *Hello*, *I wanna do it*) や語彙レベルのパターン(e.g., *Phone*, *Make*)[9]がいち早く流用可能となるということから説明可能である．

この仮説からは，次のことが予測される:

(4) PLMのパターン生成アルゴリズムを用いて幼児の発話からパターンを生成した場合，その総数及び生産性は年齢を経る毎に上昇していくはずである．

以下で，このことを調査によって検証する．

4. 調査

4.1 データ

データには，Borensztajnら[2]と同様，CHILDESデータ内のBrown コーパス[3]における幼児本人(Adam, Eve, Sarah)の発話のみから，言い差し・重複(“+...,”や“+/.”，及び“+,”でマークされる)や休止(“#” でマークされる)の含まれる発話を取り除き，データ量の観点のみから3分割したものを利用した.[3] 表1に各データの概要を提示する．

この3分割は今述べたようにデータ量が3分の1ずつになるように行ったものであり，発達の段階や幼児間の年齢の対応は考慮されていない．

4.2 PLMの適用

3幼児×3データ計9種類のデータに対し，先に提示した方法でパターン集合を生成した.[4] ただ

[1] これは「変項の再帰的単純化[7, p. 671] (RSV)」と呼ばれる処理の簡略版である．本来のRSVは，連続するl個の変項列Xと$l-1$個の変項列X'に対して[X' is-a X]とする処理である．RSVは，従って，ラティス上で同一のランク内にis-a関係を持つパターン対が生じることを許す処理となっている．このことは束論上は問題ない(=問題なく上限と下限を定義できる)が，パターン束の構造上はやや問題を孕む．このことに理論的な整合性を付けるには，ラティスに「奥行き」という概念を導入し，他次元の構造を考える必要が生じる(黒田航, p.c.)

[2] ただし，上述の通りPLMは言語記憶の実体は「事例(exemplars)」であると考えるので，正確にはここでの「再利用可能性」とはパターンの再利用可能性ではなく，そのパターンを(共)有する事例の再利用可能性を意味する．

[3] CHILDESは大人(主に養育者)と幼児との自然な対話のデータベースであり，共通のフォーマットで記述された複数の言語の様々なコーパスが含まれる．中でも[2]や本研究で使用したBrownコーパス[3]は，データ量及び調査期間の長さの観点から比較的大きな規模を持つため，様々な研究で利用されている．

[4] パターンの生成及び集計にはPython (ver. 2.6.5; Windows版)を用い独自に作成したプログラムを利用した．

表 1 データ概要[2, Table1 を改変]

	Files	Age	#sent.	MLU	vocab.	t/t
Adam						
P1	1-16	2:3-2:11	11,184	1.83-2.90	1,407	.056
P2	17-32	2:11-3:6	11,578	2.44-4.06	2,010	.053
P3	33-48	3:6-4:5	9,071	3.63-4.97	2,006	.055
Eve						
P1	1-7	1:6-1:9	3,485	1.53-2.28	669	.102
P2	8-14	1:9-2:0	3,395	2.51-3.22	785	.083
P3	15-20	2:1-2:3	3,535	2.60-3.41	958	.087
Sarah						
P1	1-45	2:3-3:2	11,693	1.48-2.70	1,389	.063
P2	46-90	3:2-4:1	8,384	2.23-3.70	1,706	.075
P3	91-135	4:1-5:0	8,525	2.98-4.86	1,944	.071

(Files:ファイルのID; #sent: 文の数; MLU: ファイル毎の語数の文平均; vocab: 語の異なり数; t/t:語のタイプ-トークン比)

し，PLMのパターン生成アルゴリズムでは，n個の分節を持つ事例からはn^2個のパターンが生成されるため，分節数が多くなる(≈文が長くなる)とパターンの数が膨大になり組み合わせ爆発が起こる．これを回避するため，以下の調整を行っている[10]:

(5) a. 語数mが閾値l(e.g., 7語)を超える文$s = \{w_1, w_2, \ldots, w_m\}$に対して，

b. w_1からw_{l-1}の連続の末尾に変項を一つ追加したパターン$I_{\text{init}} = [w_1, w_2, \ldots, w_{l-1}, _]$を作成する;

c. w_{m-l+1}からw_mまでの連続の先頭に変項を一つ追加したパターン$I_{\text{end}} = [_, w_{m-l+1}, w_{m-l+2}, \ldots, w_m]$を作成する;

d. w_2からw_{m-1}の連続に対し，$n = l-2$となるn-gramを作成し，その先頭と末尾に変項を一つずつ追加したパターン(群)I_{mid}を作成する;

e. $I_{\text{init}}, I_{\text{mid}}, I_{\text{end}}$それぞれに対しパターン集合$P(I_{\text{init}}), P(I_{\text{mid}}), P(I_{\text{end}})$を生成する;

f. sのパターン集合$P(s)$を$P(I_{\text{init}})$, $P(I_{\text{mid}})$, $P(I_{\text{end}})$の和集合$= P(I_{\text{init}}) \cup P(I_{\text{mid}}) \cup P(I_{\text{end}})$とする

今回は$l = 7$に設定した．

得られたパターン集合に対し，頻度＝対応する事例の個数fが2を上回るものを「良いパターン」とし，この良いパターンのみを選定した．得られたパターン集合をP，良いパターンの集合を$P_G := \{p \in P | f(p) \geq 2\}$ (ただし$f(p)$はpの頻度)とする．

さらに，明らかに有用性の無い(かもしくは限りなく無いに等しい)と言えるパターンを削減する処理，「パターン削減(Pattern Reduction)」を二段階で行った．パターン削減の詳細を以下に記す:

(6) a. 任意のパターン$p_i \in P$に対して，$[p_j$ `is-a` $p_i]$，つまり$[p_j < p_i]$となるようなパターンp_jの集合を$P_{<i} := \{p_j \in P | p_j < p_i\}$とする;

b. パターンpの*rank*を$r(p)$として，$[r(p_k) = r(p_i)+1]$となるパターン$p_k \in P_{<i}$の集合を$P^{+1}_{<i} := \{p_k \in P_{<i} | r(p_k) = r(p_i)+1\}$とする;

c. 削減1: もし$|P^{+1}_{<i}| = 1$ならば，p_iを削減する;

d. 削減2: もし$P^{+1}_{<i} \subset P_G$ならば，p_iを削減する

(6c)は，事例から分節を一つのみ変項化したようなパターンの一部に当てはまる．もし複数のタイプの事例にそのパターンが共有されていなかった場合，(6c)により削減される．(6d)は，*rank*の一つ上のパターンが全て「良いパターン」とされている場合に起きる削減である．一つ*rank*の上のパターンが全て有用であるならば，そのパターンの振る舞いは一つランクが上のパターンの振る舞いに還元される．例えば

(7) a. (_, like, _)

b. (I, like, _), (you, like, _), (_, like, it), (_, like, this)

というパターンがあったとして，(7b)の4パターン全てが「良いパターン」となる場合，(7a)の振る舞いはこの4者に還元されると看做され，削減される(図3参照).

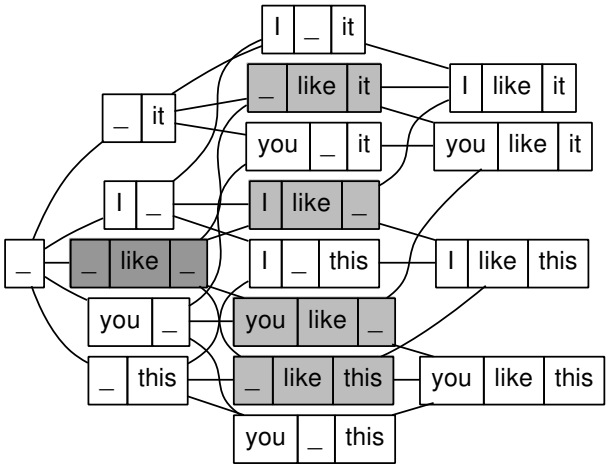


図 3 削減2で削減されうるパターンの階層関係
(Pattern Lattice Builder [7]で作成)

4.3 生産性の算出

選定したパターンそれぞれに対し，生産性を各変項に対するシャノンのエントロピー(H: 以下，単にエントロピー)で定義した．変項vのエントロ

ピー$H(v)$はvの値の集合$\{w_1, w_2, \ldots, w_m\}$に対し，それぞれの値$w_i$の全体に占める割合を$p(w_i)$として，

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

で算出される．

変項が複数ある場合(e.g., (put, _, in, _)) は各変項のエントロピーを合計したものをそのパターンのエントロピーとした．即ち，パターンpのi番目の変項をv_i, そのエントロピーを$H(v_i)$として，pのエントロピー$H(p)$を以下のように求めた:

$$H(p) = \sum_{i=1}^{n} H(v_i) \quad (2)$$

ただし，複数の変項の振る舞いが互いに独立でなく，共変動が生じている場合，単純なエントロピーの合計値は実際の生産性より高くなるため，この計算方法ではパターンの生産性を過大評価している恐れがある．これを避けるため以下のような補正を行った:[5)]

(8) a. n個の変項を持つパターンpのi番目の変項において，その実現値の値(典型的には語)の集合を$v_i = \{w_{i1}, w_{i2}, \ldots, w_{im}\}$とする;

b. n個の変項の値の直積集合を

$$V^*(p) = v_1 \times v_2 \times \ldots \times v_n = \prod_{i=1}^{n} v_i$$

とする;

c. pにおけるj番目の実現値のn組を$t_j = (a_{j1}, a_{j2}, \ldots, a_{jn})$として，その集合を$A(p) = \{t_1, t_2, \ldots, t_l\}$とする;[6)]

d. pにおける変項間の独立度$d(p)$を

$$d(p) = \left(\frac{|A(p) \cap V^*(p)|}{|V^*(p)|} \right)^{\frac{1}{\alpha}}$$

とする(ただし$|x|$は集合xの元の数);

e. pのエントロピー$H(p)$を$d(p)$を補正係数として以下のよう補正する

$$H(p) = d(p) \sum_{i=1}^{n} H(v_i)$$

もし複数の変項が完全に独立しているならば，その値の組み合わせとして可能なものは全て出現し得ることを意味する．従ってこの場合，上の(8d)で定義した独立度$d(p)$は1となり，式(2)の値がそのままそのパターンのエントロピーとなる．

しかし，変項の値が独立でなく，共変動がある場合，その度合いに応じて$d(p)$が減少する．即ち，組み合わせに「抜け」が生じる．完全に共変動している場合は，$A(p)$は$V^*(p)$の$1/n$しかないこととなり，この比率をそのまま$d(p)$とするならば，$H(p)$はn個の各変項のエントロピーの平均値となる．

しかし，このような「抜け」が共変動によるものなのか，単に「データの抜け(sparseness)」によるものなのかは判断できない．ここで，(言語)データには必ず抜けがあるものだとすると，完全に独立の場合を除き，共変動率を少なく(= 独立度を多く)見積もるように修正をかける必要が生じる．(8d)における指数$\frac{1}{\alpha}$のαはこの修正度合いを調整するパラメータである．本調査では$\alpha = 2$に設定した.[7)]

5. 結果と考察

5.1 結果概要

3幼児$\times$3データ計9データから得られたパターンの総計(P),「良いパターン」の数(G), 削減1で削減されたパターン数(R1), 削減2で削減されたパターン数(R2), 最終的に残ったパターン数(S)をそれぞれ以下に提示する．表2から，どの幼児もパターンの総数が増加していっていることが分かる．

表2 得られたパターンの総数

	P	G	R1	R2	S
Adam					
P1	36795	6606	722	627	5257
P2	145972	21313	1467	3152	16694
P3	195858	21876	1006	2613	18257
Eve					
P1	8602	1389	139	60	1190
P2	21785	3379	343	542	2494
P3	43544	4998	302	817	3879
Sarah					
P1	24530	5012	594	442	3976
P2	71412	8787	579	1113	7095
P3	117711	13647	884	2432	10331

最終的に残ったパターンに対して算出されたエントロピーの集計結果(平均，最大値等)を表3に提示する．表3に見る通り，どの幼児もエントロピーの平均が年齢を経る毎に増大していることがわかる．

エントロピーの平均の増加が有意なものであるかどうかの検証にあたっては，Wilcoxon の順位和

[5)] この過大評価の問題は査読者の一人にご指摘頂いたことであり，また，同じ査読者に(8)に提示した補正方法の原案をご提示頂いた．この場を借りて謝意を表したい．

[6)] 例えば$p = (_, \text{got}, _)$として，*I got it*という実例が出現している場合，(I, it)という実現値組が得られる)

[7)] 言うまでもなくこの妥当性は独立の基準でたしかめる必要がある．今後の課題としたい．

表3 パターンの個数とエントロピー

	N	h_ave	h_max	h_min	h_std
Adam					
P1	5257	1.21	7.59	0.16	1.03
P2	16694	1.46	10.15	0.13	0.78
P3	18257	1.52	9.96	0.07	0.70
Eve					
P1	1190	1.14	6.69	0.37	0.93
P2	2494	1.38	7.68	0.26	0.81
P3	3879	1.44	8.29	0.55	0.72
Sarah					
P1	3976	1.14	9.14	0.44	1.05
P2	7095	1.43	9.36	0.15	0.81
P3	10331	1.48	9.09	0.28	0.76

(N: パターンの個数; h_x: エントロピーの平均(ave), 最大値(max), 最小値(min), 標準偏差(std))

検定を利用した.[8] これはPLMの定義するパターンの頻度分布が統計的にどのような性質を持ったものであるかは今のところ未知であるためである. 表4にWilcoxonの順位和検定の結果(p値及び統計量W)を提示する.

表4 エントロピーの平均の差の検定(Wilcoxon test)

	p-value	W
Adam		
P1-P2	2.57E-117	34759996.5
P2-P3	4.16E-22	143392303.5
P1-P3	2.36E-187	35466067
Eve		
P1-P2	7.74E-20	1212357
P2-P3	5.55E-06	4515609.5
P1-P3	9.32E-38	1746898.5
Sarah		
P1-P2	8.95E-83	11034700
P2-P3	9.06E-09	34798423
P1-P3	2.02E-129	15237634

表4から明らかなように, 全ての幼児の全ての組において有意差が認められた. また, 表5に, AdamのP1の結果のうち, 頻度が上位20位までのパターンを例示する.

5.2 考察

以上の結果は, PLMの定義するパターンが統語知識の何らかの実体を捉えているとすれば, 幼児の統語発達をうまく説明する. ただし, 以下の点には注意が必要である:

(9) 今回パターンの生成に使った入力データは幼児本人の発話＝産出データである

[8] Wilcoxonの順位和検定は, 統計言語RのPythonインターフェースである“RPy2 (R for Python 2)”モジュールを利用して行った.

表5 Adam (P1)の頻度上位20位のパターン

id	pattern	freq	h	rank
p3748	(what, _)	609	4.753605972	1
p258	(_, dat)	581	3.353254861	1
p60	(_, it)	413	6.938419458	1
p211	(_, a, _)	360	1.210920686	1
p83	(where, _)	333	6.385406368	1
p51	(I, _)	312	7.592606597	1
p47	(yeah)	304	0	1
p5398	(what, dat)	260	0	2
p81	(_, go)	239	6.094410938	1
p37	(_, dere)	223	5.345826243	1
p71	(no)	223	0	1
p29	(put, _)	216	6.991296312	1
p10	(who, _)	201	1.781998418	1
p119	(_, in, _)	200	1.59391624	1
p42	(Adam, _)	186	6.923208949	1
p591	(_, there)	185	6.348410713	1
p2798	(dat, _)	170	6.78865101	1
p2144	(_, dat, _)	161	1.512231192	1
p594	(_, in, there)	153	6.354518787	2
p1008	(_, it, _)	150	1.517097192	1

これはつまり, 今回得られたパターン群(の諸性質)が幼児の得た入力データ(≈養育者の発話データ)に基づく模擬的な「学習」の結果ではないということである. むしろ, 本調査で示したのは, 幼児の産出データから「逆算」して, その背後にある言語知識の可能な形態を「再現」したものであると言える.[9]

このことを踏まえると, 今回の結果の意味するところは以下のようなことであると言えよう:

(10) ヒトの統語知識をPLMの規定するパターンの体系であると看做す限りにおいては, 幼児は年齢を経るに従って, より生産性の高い, 数多くのパターンを使うことで可能となる発話を行うようになる[10] [11]

5.3 エントロピーの高さは何を意味しているか

本調査では生産性の指標としてシャノンのエントロピーを用いたが, ここで, エントロピー(の高さ)が一体何を意味しているのか, 本当に本調査で示したかった生産性の尺度となっているのか, ということを改めて考えてみたい.

起こりうる事象がn個ある場合エントロピー(H)は$[H \leq n]$となる. 従って, ある変項の実現値の集合v_iにおいては$[H \leq |v_i|]$となるということである. $[H = n]$となるのは, 各事象の生起確率が完全にラ

[9] これはBorensztajnらの研究[2]でも同じである

[10] ただし, 今回はパターンの数(＝異なり数)の増加が有意であることは示せていない.

[11] もちろん, 今回入力に用いたデータに見られる発話を可能にするには「少なくとも本調査で提示したパターンの使用が必要である」と言っているだけであり, それ以外のパターンを知識として持っている可能性は否定されない.

ンダムで均一な場合である．以上から，変項のエントロピーとは，その変項において現れうる実現値の数が多ければ多いほど高く，また，その生起確率が均一であればあるほど高いということになる．

この性質は「パターンの生産性」の指標として適切なものだろうか．すぐに分かるように，変項の可能な実現値の数が多いということは，その変項の多様性・流用性の高さを示していると言え，この点に関しては生産性の高さの指標として申し分ないと言える．では，各実現値の生起確率が均一であればあるほど高くなるという性質はどうだろうか．実現値の生起確率が均一であるということは，どの実現値を用いた場合でも等しく利用可能であることを意味し，それだけ「自由に」利用できる表現ということを意味する．従ってこの場合もやはり流用性の高さを示していると言え，生産性の指標として妥当であると結論付けられる．

ただし，今回示した年齢経過に伴うパターンの生産性の上昇は，統語知識の発達の一側面しか捉えられていない可能性があるということは付け加えておかねばならない．統語の発達には，明らかに，個々のパターンの生産性上昇とは別に，知っているパターンの総数＝異なり数の増加も影響している．脚注10) でも述べたが，本調査ではパターンの異なり数の増加が有意であることは示せていない．それでも表に見るように，明らかに異なり数は増加傾向を示しており，その点で今述べた問題は恐らく問題にはならないと思われるが，今後はパターンの異なり数の変化を取り込んだ(統計)指標によって，生産性の向上を数値化できるように努める必要があることは言うまでもない.[12)]

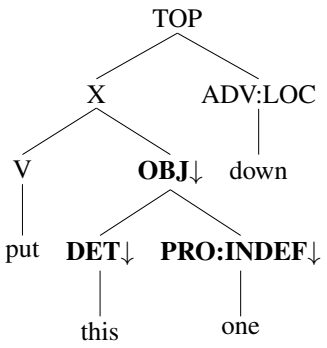
5.4 Borensztajnら[2]との比較

ここで簡単にではあるが，Borensztajnらの研究[2]との比較を試みる．本研究と[2]との決定的な違いは，統語知識の表示(法)である．彼らは「(確率的)木代入文法 ((Probabilistic) Tree-Substitution Grammar, TSG)」を理論的背景としており[2, p. 178]，従って統語知識の表示(の単位)は大小様々な規模からなる「木」である．また，明言はされていないが，論文中に登場するいくつかの木を見るからに，二股枝分かれしか許されないようである．一方本研究では，統語知識の表示に「パターン」を用いている．

[12)] このことも前述の査読者にご指摘頂いた．残念ながらこの解決策は見いだせないままになってしまったが，重要な指摘に感謝したい．

この違いはそれ自体優劣をつけられるような性質のものではないが，ただ一点，その表示の性質が故に生じていると考えられる[2]の問題点を指摘しておきたい．今述べたように，[2]では二股枝分かれの木のみを表示の単位として認めているために，以下のような，明らかに構文の「深さ」＝階層の数を過大評価しているケースが存在する[2, p. 183]:

(11)



この場合もし*put*と*down*を結びつける不連続かつ三又のノードが許されれば，*X*のようなノードは不要になり，階層も一つ減ることになる．このことが意味するのは軽微ではないように思われる．

また，[2]では定量的な調査は行われており，構造木の「深さ」(=「階層数」)や「構文(constructions)」(= 深さが2以上の木)の数，末端・非末端のノードの数等，様々な項目で年齢による増加が見られたことが報告されている[2, pp. 183-184]が，統計的な検定などは行われていない．行えるにも関わらず実施していないのか，何らかの要因で原理的に不可能なため実施していないのかは分からないが，この点においては(その妥当性は別途検証する必要があるにせよ)統計的な裏付けのある本研究の方が優れていると言うことは可能であろう．

6. 結語

本稿では，パターン束モデル(PLM)を用いて，CHILDES内のコーパスにおける3幼児の産出データを元にパターンを生成し，その変項のエントロピーに基づく生産性の算定を行った．その結果，3幼児共に年齢を経るごとにパターンの生産性が有意に増加していることが認められた．この結果から，統語知識をPLMの定義するパターンの体系と看做した場合，徐々に生産性を増す統語発達のプロセスを定量的・統計的に捉えることに成功したと言える．

ただし明らかに一つ問題となるのは，今回行ったデータの3分割法である．これは発達の段階等を無視し，人為的に分割を行ったものであり，また，さらに細かく分割した場合に，十分な統計量が確保できず，今回認められたような有意差が認

められなくなる可能性があるという問題も挙げられる．

このような問題を回避するには，

(12) a. データとして幼児本人の産出データではなく，入力となった大人(主に養育者)の発話を用いて，

b. 漸増的にデータを与え徐々にパターンを獲得させるような動的な処理を行い，

c. 本人の産出データとの整合性によりその評価を行う

のが望ましい．今後はそのような方法論を模索しつつ，より妥当な統語発達のモデル構築を目指したい．

参考文献

[1] Bod, R., (2006) “Exemplar-based syntax: How to get productivity from examples” *Linguistic review*, Vol. 23, No. 3, pp. 291-320.

[2] Borensztajn, G., Zuidema,W., & Bod, R., (2009) “Children’s grammars grow more abstract with age: Evidence from an automatic procedure for identifying the productive units of language” *Topics in Cognitive Science*, Vol. 1, No. 1, pp. 175–188.

[3] Brown, R. W., (1973) *A first language: The early stages*, Cambridge, MA.: Harvard University Press.

[4] 長谷部陽一郎, (2009) “計算的手法を用いた構文習得研究の可能性” 言語文化, Vol. 12, No. 2, pp. 395–420.

[5] 黒田航, (2007) “徹底した用法基盤主義の下での文法獲得:「極端に豊かな事例記憶」の仮説で描く新しい筋書き” 言語, Vol. 36, No. 11, pp. 24–34.

[6] Kuroda, K., (2009) “Pattern lattice as a model for linguistic knowledge and performance” *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation*, pp. 278–287.

[7] 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.

[8] MacWhinney, B., (2000) *The CHILDES project: Tools for analyzing talk*, Mahwah; New Jersey: Lawrence Erlbaum Associates.

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

[10] 吉川正人, (2010) “「語」を超えた単位に基づくコーパス分析に向けて: パターンラティスモデル(PLM) とその有用性” 藝文研究, Vol. 98, pp. 50–64.