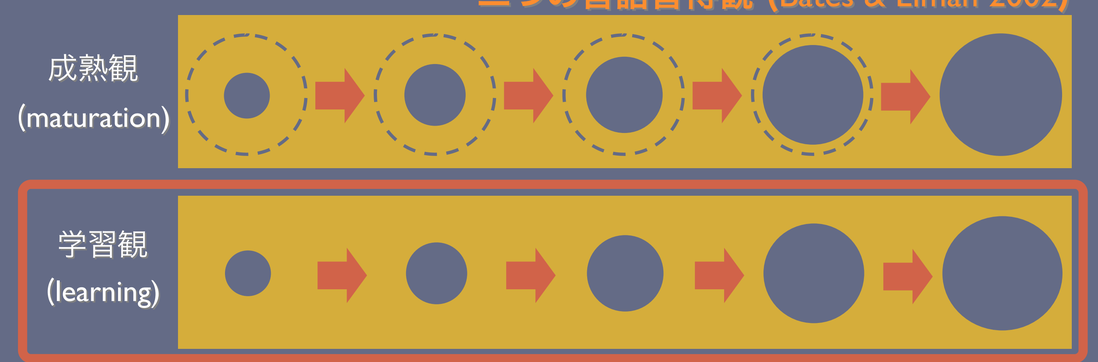
吉川 正人
machayoshikawa@dream.com
慶應義塾大学大学院
日本学術振興会特別研究員

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

1 <研究の背景と目的>

二つの言語習得観 (Bates & Elman 2002)



課題

刻一刻と変化する知識状態をどう記述し段階的な変化をどう計測するか

(憂うべき)現状

ほとんどこの検証は行われていない! (例外: Borenzstajn et al. 2009)

そこで...

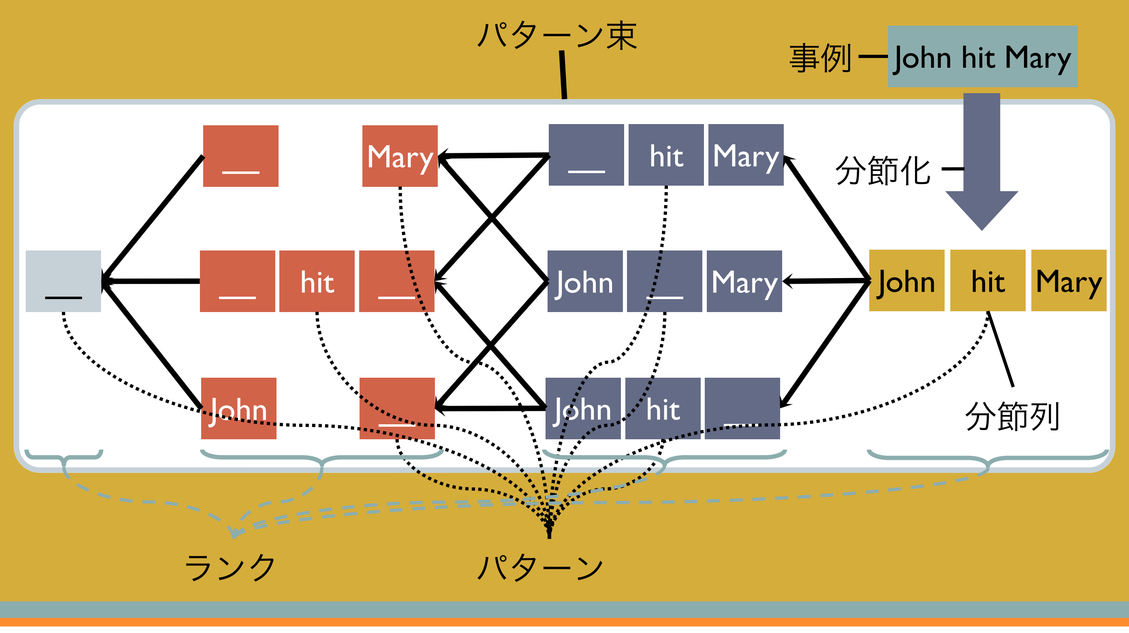
ならばやってみよう!

概要

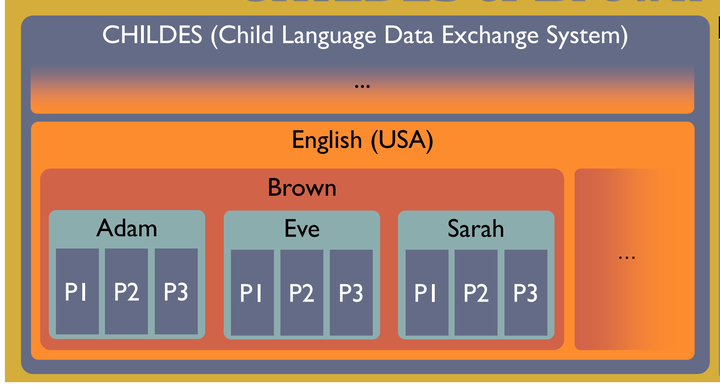
- パターン束モデル (Pattern Lattice Model: e.g., 黒田・長谷部 2009)に基づき
- Brown コーパス(Brown 1973) in CHILDES (MacWhinney 2000)を使用し
- 幼児の発話から「パターン」= 統語知識の候補を生成し
- その「生産性 (Productivity)」をシャノンのエントロピーで算定し
- 年齢を経る毎に生産性が高まっていく段階的な発達プロセスを検証する

2 <パターン束モデル (Pattern Lattice Model)>

- 黒田・長谷部 (2009)の提案したヒトの言語知識のモデル
  - ◎ (現時点では)主に**言語形式の組織化**のモデル
  - ◎ ヒトの統語知識は(ツリーではなく)**「パターン」の集合**であると考える
- パターン束
  - ◎ **事例** e の任意の**分節モデル** *T* による**分節化** *T*(e) に対する**再帰的変項化**の産物
  - ◎ 内実はパターンの「集合」 = **パターン集合** *P*(e)
  - ◎ 部分一致に基づく**「継承関係」** (is-a) の規定された**半順序集合** = **パターン束** *L*(e)
- 簡単な例
  - ◎ e = *John hit Mary.*　*T* = 単語分節　*T*(e) = [John, hit, Mary]
  - ◎ *P*(e) = {(John, hit, __), (John, __, Mary), (__, hit, Mary), (John, __), (__, hit, __), (__, Mary), (__)}
  - ◎ 補足: **連続する変項を単一の変項に縮約** ((John, __, __) → (John, __))



CHILDES & Brown's Corpus



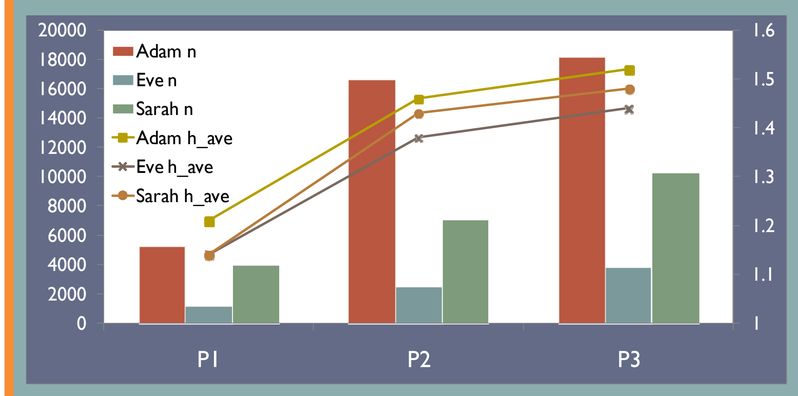
	Files	Age	#sent.	vocab.	t/t
Adam					
P1	1-16	2:3-2:11	11,184	1,407	0.056
P2	17-32	2:11-3:6	11,578	2,010	0.053
P3	33-48	3:6-4:5	9,071	2,006	0.055
Eve					
P1	1-7	1:6-1:9	3,485	669	0.102
P2	8-14	1:9-2:0	3,395	785	0.083
P3	15-20	2:1-2:3	3,535	958	0.087
Sarah					
P1	1-45	2:3-3:2	11,693	1,389	0.063
P2	46-90	3:2-4:1	8,384	1,706	0.075
P3	91-135	4:1-5:0	8,525	1,944	0.071

3 <調査>

- データ
  - ◎ CHILDES (MacWhinney 2000) 内の**Brownコーパス**(Brown 1973)
    - ▸ **幼児** = {Adam, Eve, Sarah} の発話のみを抜き出し
    - ▸ **重複・言いさし、ポーズ**の含まれる発話を除外
  - ◎ 3幼児それぞれのデータを**データ量が揃うように3等分** = {P1, P2, P3}
- 方法
  - ◎ 3幼児 × 3データに対しそれぞれPLMのアルゴリズムによる**パターン生成** … 1
  - ◎ **頻度 ≥ 2** のパターンを**選抜** (**「良いパターン(Good Patterns)」**の選定) … 2
  - ◎ 良いパターンから**「パターン削減 (Pattern Reduction)」**によりさらに選抜 … 3
    - ▸ 削減1: ランクが一つ上 = **階層が一つ下のパターンの異なりが一つ**なら**削除**
    - ▸ 削減2: ランクが一つ上 = **階層が一つ下のパターンが全て良いパターン**なら**削除**
  - ◎ 選抜パターンに対し**シャノンのエントロピー**を用いて**生産性**を算定 … 4
- 生産性の算定
  - ◎ 算定方法は**別紙 (Appendix 1)** 参照
  - ◎ P1-P2間, P2-P3間, P1-P3間で生産性の**平均の差の検定** (**Wilcoxonの順位和検定**) … 5

4 <結果と考察>

- **パターン数・生産性の平均値**: 右下図・下図の通り



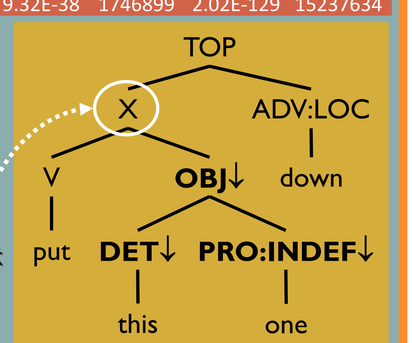
得られたパターンの例
(Adam, 頻度上位20位)

id	pattern	freq	h	rank
p3748	(what, __)	609	4.75	1
p258	(__, dat)	581	3.35	1
p60	(__, it)	413	6.94	1
p211	(__, a, __)	360	1.21	1
p83	(where, __)	333	6.39	1
p51	(I, __)	312	7.59	1
p47	(yeah)	304	0.00	1
p5398	(what, dat)	260	0.00	2
p81	(__, go)	239	6.09	1
p37	(__, dere)	223	5.35	1
p71	(no)	223	0.00	1
p29	(put, __)	216	6.99	1
p10	(who, __)	201	1.78	1
p119	(__, in, __)	200	1.59	1
p42	(Adam, __)	186	6.92	1
p591	(__, there)	185	6.35	1
p2798	(dat, __)	170	6.79	1
p2144	(__, dat, __)	161	1.51	1
p594	(__, in, there)	153	6.35	2
p1008	(__, it, __)	150	1.52	1

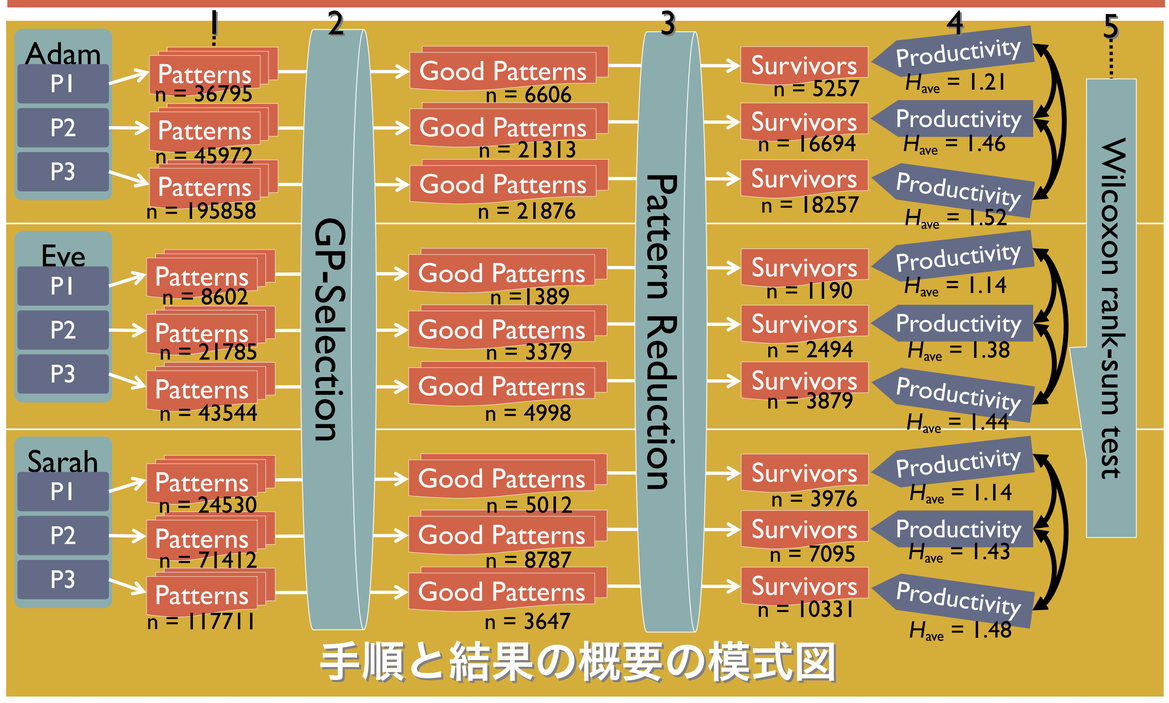
- 生産性の差の検定結果 ⇒
  - ◎ 軒並み低いp値

	Adam		Eve		Sarah	
	p-value	W	p-value	W	p-value	W
P1-P2	2.57E-117	34759997	7.74E-20	1212357	8.95E-83	11034700
P2-P3	4.16E-22	1.43E+08	5.55E-06	4515610	9.06E-09	34798423
P1-P3	2.36E-187	35466067	9.32E-38	1746899	2.02E-129	15237634

- エントロピー上昇は何の証左か?
  - ◎ 幼児の発話における「**体系的ばらつき**」の増加
    = 適度な**部分一致**を含む**バラエティ豊かな発話**への変化
- 先行研究 (Borenzstajn et al. 2009)との簡単な比較
  - ◎ "Right representation"? : Tree Substitution Grammar = **木**表示
    - ▸ 「木」より**「束」**の方が**「弱い」**かつ**「網羅的」**
    - ▸ 「抽象度」を**過大評価**している節あり



Borenzstajn et al. (2009 : 183) より一部改変



手順と結果の概要の模式図

Keio University
1858 CALAMVS GLADIO FORTIOR

<謝辞>

黒田 航氏　(情報通信研究機構)
長谷部 陽一郎氏　(同志社大学)
井上 逸兵教授　(慶應義塾大学)
中村 文紀氏　(慶應義塾大学大学院)

<主要参考文献>

- Bates, E., & Elman, J. 2002. Connectionism and the study of change. In Johnson, M., Munakata, Y., & Gilmore, R. O. (eds.) *Brain development and cognition: A reader* (2nd ed.). (pp. 420-440) Oxford: Blackwell Publishing.
- 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, 1 (1)*, 175–188.
- Brown, R. 1973. *A first language: The early stages*. Cambridge, MA.: Harvard University 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).
- MacWhinney, B. 2000. *The CHILDES project: Tools for analyzing talk*. Mahwah, New Jersey: Lawrence Erlbaum Associates.