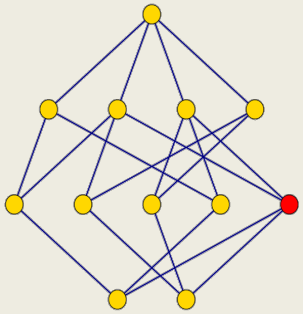
2011年7月9日
JAECS東支部課題別シンポジウム
「文を超えたコーパス研究」@慶應三田



幼児発話からの生産的統語パターンの獲得

文法発達の計算理論に向けて

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
http://www.yoshikawacademia.com

1. はじめに

概要

- 用法基盤モデル (Usage-based Model) の想定
  - 文法知識は「漸進的に (gradual)」獲得される
    ⇔ 「連続仮説 (Continuity Assumption)」
    - Cf. Tomasello 2003: Ch.1, "we can/can't get there from here"
  - 幼児の文法知識は刻一刻と変化していくということ
    ➔ データ駆動の柔軟な構造記述法が求められる
- 提案
  - 時点 t の発話 u_t の構造を記述するために以下を利用
    - 時点 t-1までの発話の履歴 $U_{t-} = [u_1, u_2, \ldots, u_{t-1}]$
  - そのための方法論の紹介と記述結果の提示を行う
    - パターン束モデルを用いたBrown コーパス in CHILDES の分析

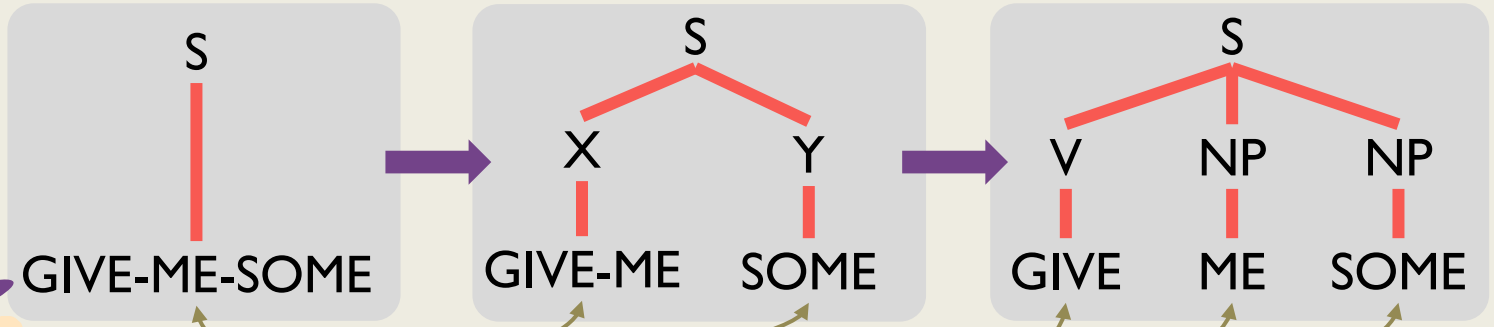
構成

- 2節: 背景
  - 問題の所在の明確化
  - 先行研究とその問題点の提示
- 3節: パターン束モデルの導入
  - パターン束モデルの概要
- 4節: 分析
  - データの紹介・方法の詳述
- 5節: 結果と考察
- 6節: まとめ・課題

2. 背景

刻一刻と変化する統語知識

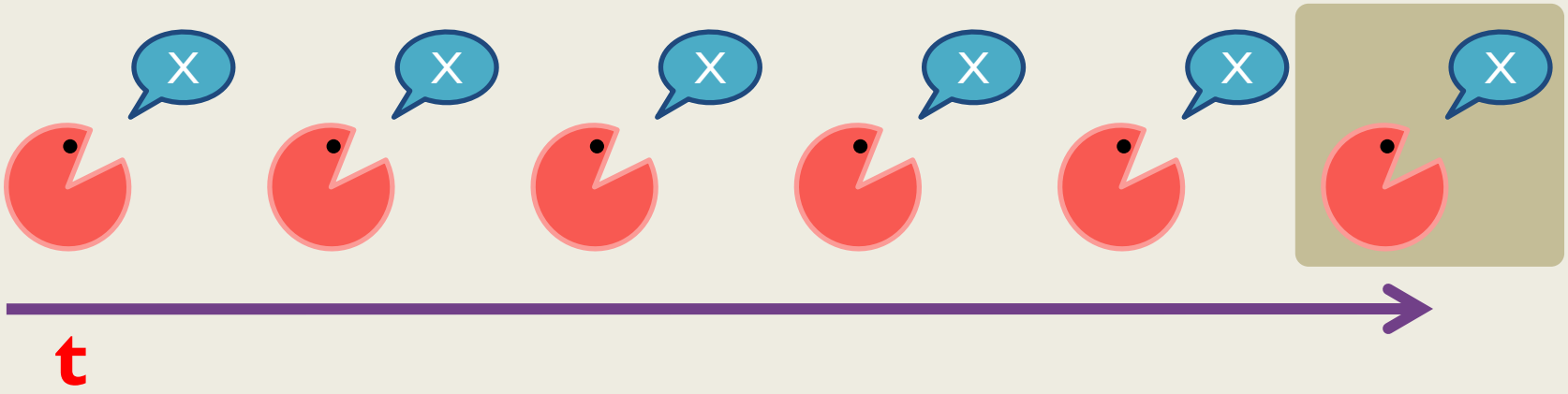
- 統語発達の記述の難しさ
  - 何が発達しているのか =「統語知識」の内実が不明
  - 統語知識は刻一刻と変化 ➔ 先駆的な表示は使用不可
  - 「発達度合い」をどう図るか
    - ▶ 獲得したかどうかの判断 / 発達度合いの「指標」



Give me some.

「文脈」の位置づけ

☼ 文脈 = 発話の履歴



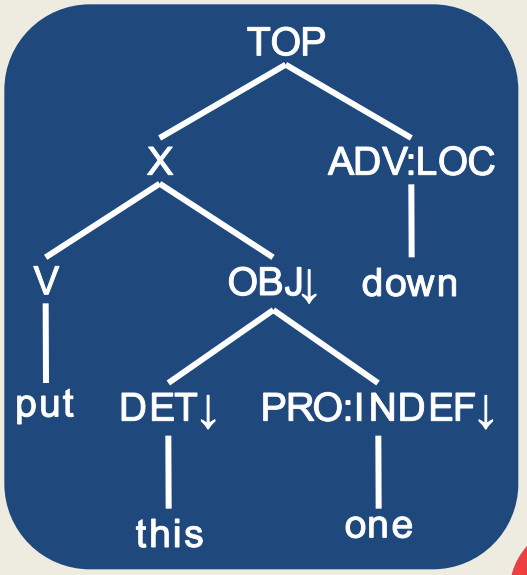
- 「同じ」発話も履歴が異なれば「位置づけ」が異なる
  ➔ 発話を履歴に相対的な形で解析

☼ ただし: 文法知識発達の「(局所) 離散性」を仮定

- 知識状態は一定の時間が経過しないと変化しない
  - ▶ 一度睡眠を取らないと変化しない = 1日の内には変化しない!

先行研究とその問題

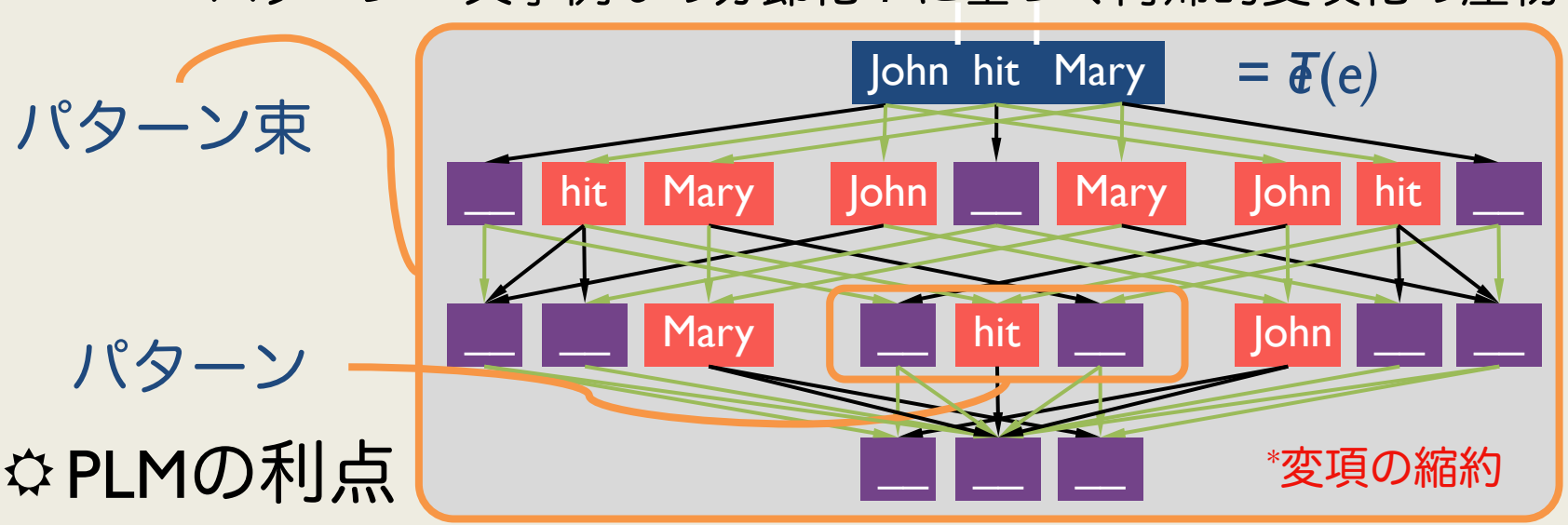
- 先行研究
  - Borensztajn et al. (2009)
    - Data-oriented Parsing (DOP; e.g., Bod 2006)の枠組み
    - ⬇ But
    - “right representation” (Borensztajn et al. 2009: 177)??
      - 二股枝分かれ,ダミーノード(“X”)
- 従って
  - 「入れ知恵」最小・データ駆動の表示と計算の理論が必要
- ではどうする?
  - 「パターン束モデル」の導入



3. パターン束モデル

パターン束モデルとは

- パターン束モデル (Pattern Lattice Model, PLM)
  - e.g., 黒田・長谷部 (2009), Kuroda (2009), 吉川 (2010)
- パターンとその継承関係のネットワークモデル
  - パターン = 文事例 e の分節化 T に基づく再帰的変項化の産物



- PLMの利点
  - 自動かつ網羅的な(超語彙)パターン生成
    - 入れ知恵 = 分節モデルのみ / 完全データ駆動

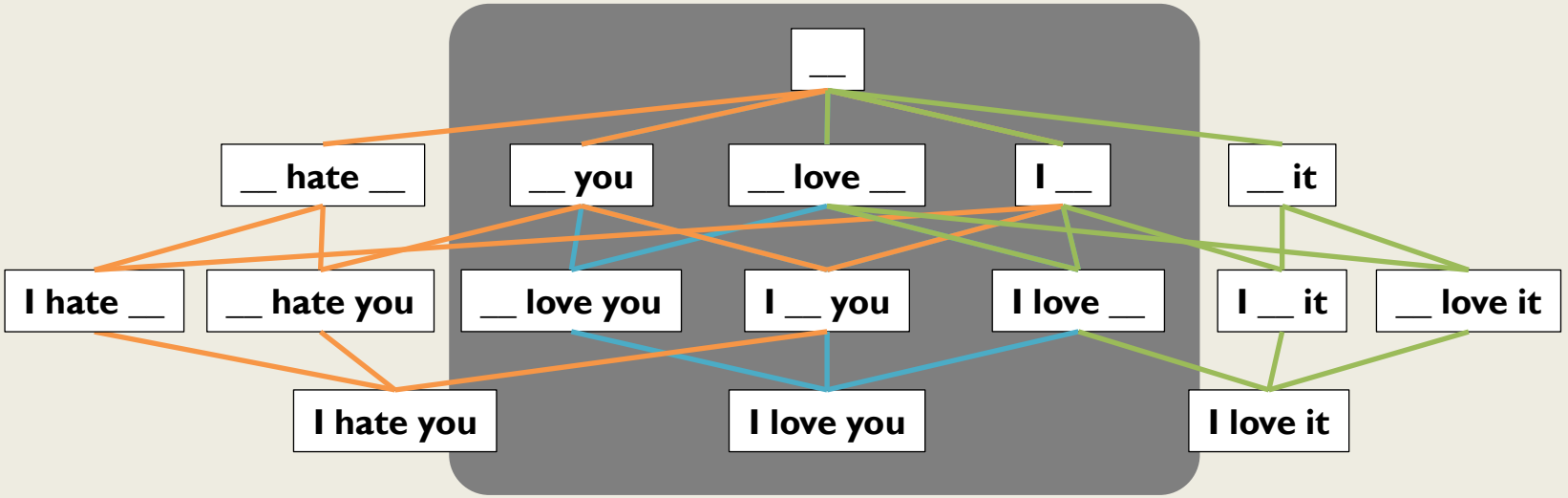
PLMの利点と問題

- データ駆動性
  - 品詞ラベルや限定された構造表示は用いない
  - 網羅性 (exhaustiveness)・貪欲性 (greediness)
- ただし: 柔軟性…？
  - 問題: 網羅的過ぎる
    - 与えられたデータに応じ得られる構造を変化させられない

  - 複数事例のパターン束を合成することで解決
    - 同時に与えるデータによって同形の発話でも解析結果が変化

パターン束の最適化

- 複数の事例から得られたパターン束の結合



- 冗長なパターンを削除する最適化処理

I love you

  - 1) 下位パターンに頻度が同一のものがあったら削除
  - 2) 頻度が 1 だったら削除

(事例個別のパターン束にのみ適用)

    - 具体的なパターン優先でバリエーションのないものを削除

4. 分析

Brown コーパス in CHILDES

☼ データ

- Brown コーパス (Brown 1973)
  - CHILDES (MacWhinney 2000) 内のコーパスの一つ
  - 三幼児 (Adam, Eve, Sarah) と周囲の大人との対話のデータ
- データの概要
  - Adam: 2;3 ~ 5;2 (55 files)
    Eve: 1;6 ~ 2;3 (20 files)
    Sarah: 2;3 ~ 5;1 (139 files)

☼ 前処理

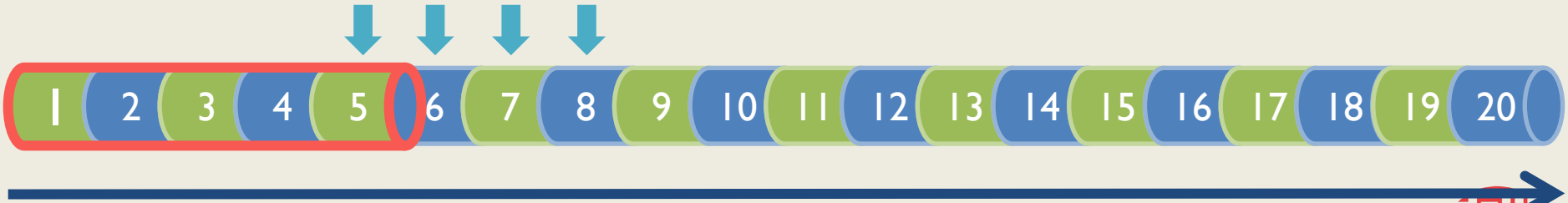
- 幼児の発話のみの抜き出し (= 大人の発話を除去)
- 重複・言いさし・ポーズの含まれる発話を除去
  - コーパス上でアノテートされている

Adam の詳細

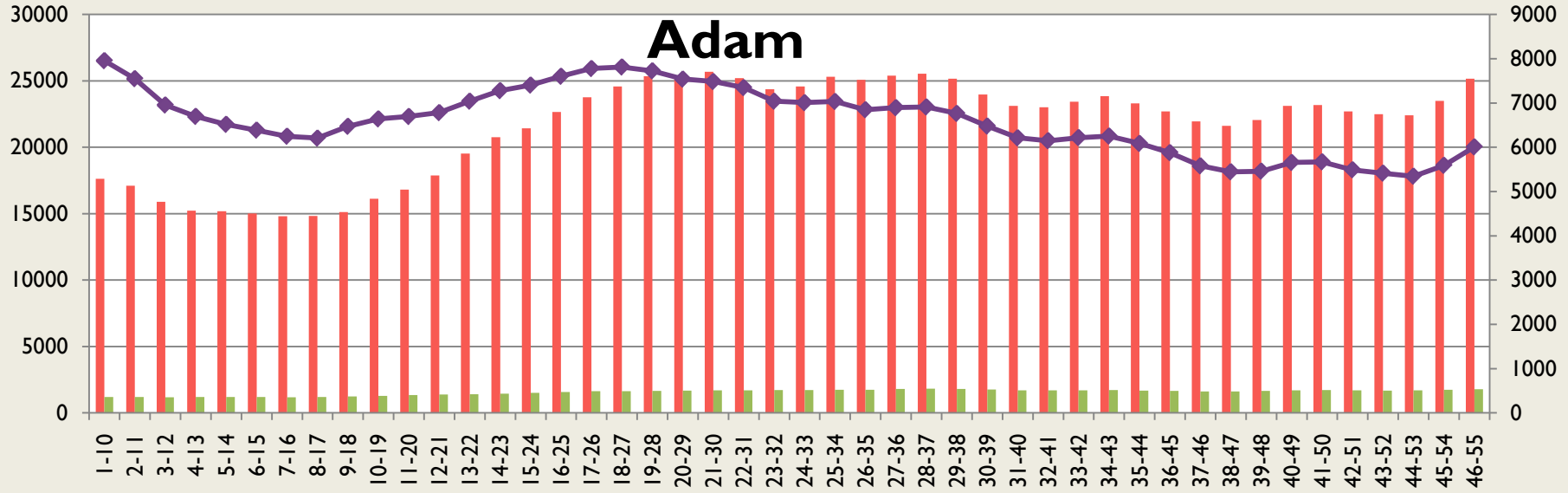
File	Age	File	Age	File	Age	File	Age
01	2;03.04	15	2;10.02	29	3;04.18	43	4;01.15
02	2;03.18	16	2;10.16	30	3;05.01	44	4;02.17
03	2;04.03	17	2;10.30	31	3;05.15	45	4;03.09
04	2;04.15	18	2;11.13	32	3;05.29	46	4;04.01
05	2;04.30	19	2;11.28	33	3;06.09	47	4;04.13
06	2;05.12	20	3;00.11	34	3;07.07	48	4;05.11
07	2;06.03	21	3;00.25	35	3;08.01	49	4;06.24
08	2;06.17	22	3;01.09	36	3;08.14	50	4;07.01
09	2;07.01	23	3;01.26	37	3;08.26	51	4;07.29
10	2;07.14	24	3;02.09	38	3;09.16	52	4;09.02
11	2;08.01	25	3;02.21	39	3;10.15	53	4;10.02
12	2;08.16	26	3;03.04	40	3;11.01	54	4;10.23
13	2;09.04	27	3;03.18	41	3;11.14	55	5;02.12
14	2;09.18	28	3;04.01	42	4;00.14		

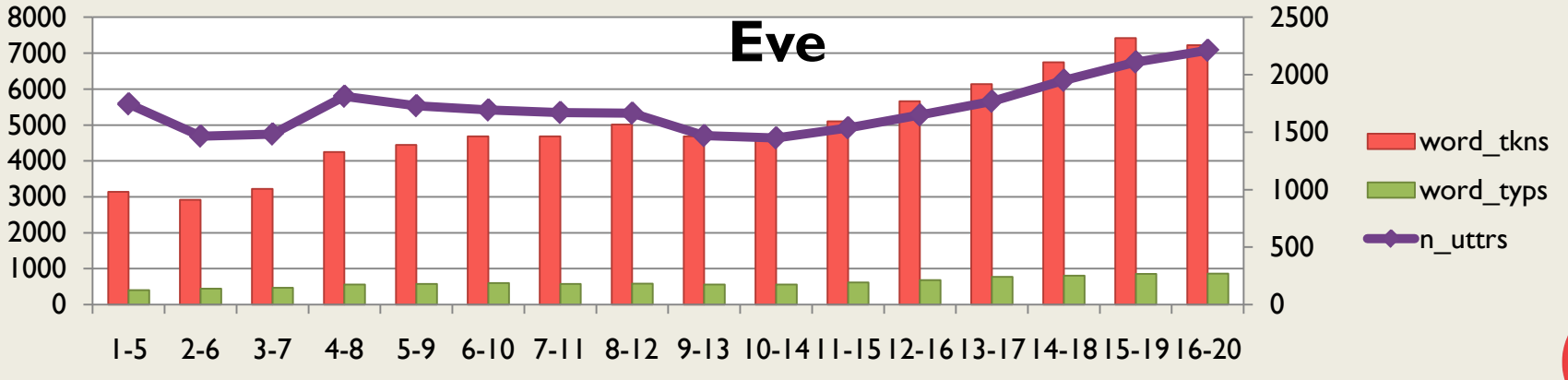
方法

- 各幼児 1 ファイル毎に時系列に沿って分析
  - 入力データ: 当該ファイル + 過去 *n* - 1ファイル
    - Adam: *n* = 10; Eve: *n* = 5; Sarah: *n* = 20
      - 要するにファイルの *n* グラムを作成
  - *n* ファイルからなる発話群からパターン束を生成
  - 得られたパターン束から末ファイル中の各発話を解析
- 解析は全て自作のスクリプトによって実行
  - スクリプト言語 Python (Ver. 2.6.5;Windows版) を使用

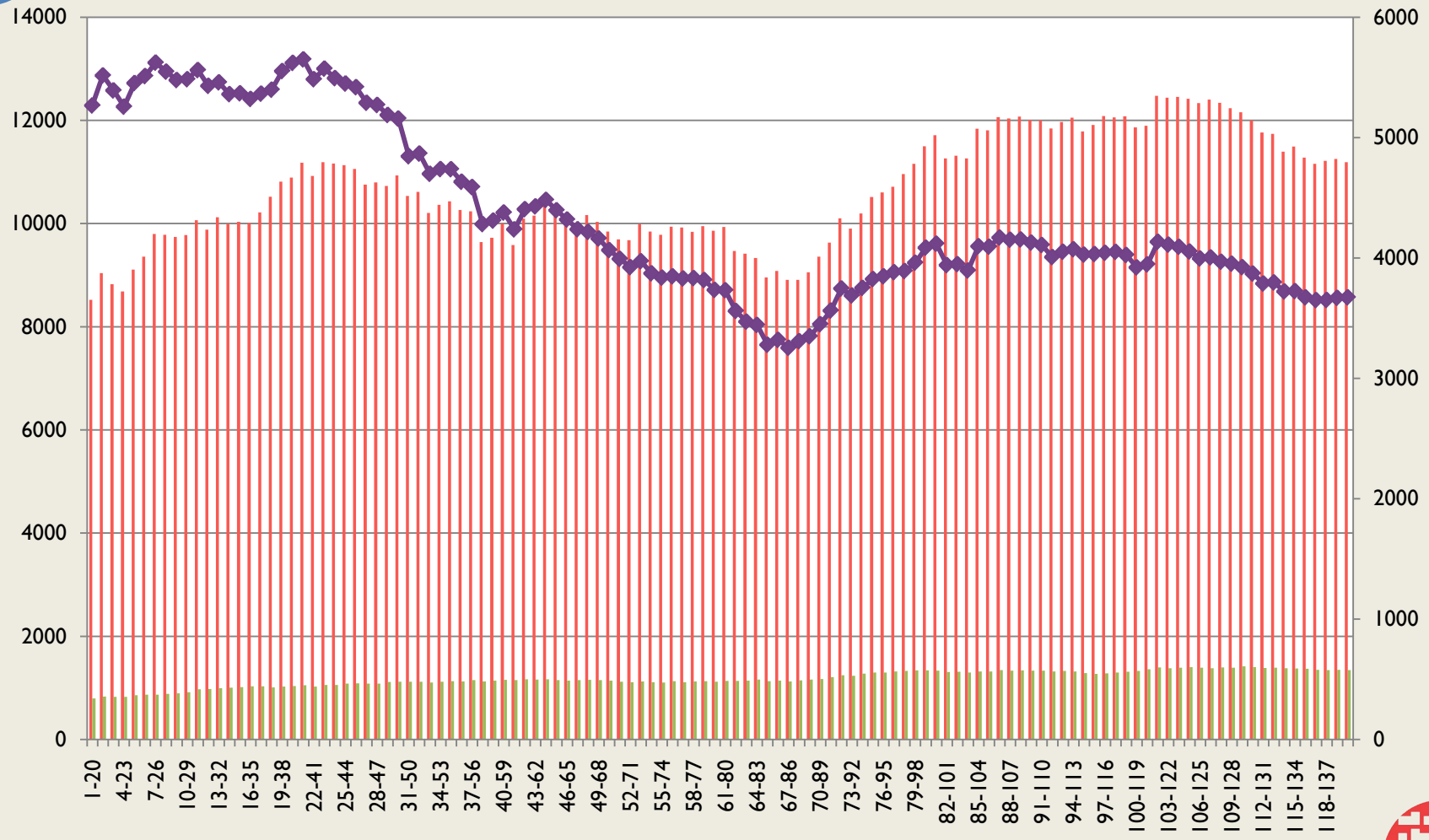


Adam, Eve のデータ





Sarah のデータ


14000
12000
10000
8000
6000
4000
2000
0
6000
5000
4000
3000
2000
1000
0
1-20
4-23
7-26
10-29
13-32
16-35
19-38
22-41
25-44
28-47
31-50
34-53
37-56
40-59
43-62
46-65
49-68
52-71
55-74
58-77
61-80
64-83
67-86
70-89
73-92
76-95
79-98
82-101
85-104
88-107
91-110
94-113
97-116
100-119
103-122
106-125
109-128
112-131
115-134
118-137

どう料理するか

- 分析 [1]
  - 生起ファイル数の多い発話形式の構造の変化を見る
    - ➔ 同一発話形式の構造変化プロセスを捕捉
      - 分析対象は次スライド (e.g., *I can't do it*, *Let me see*)
  - 補助的に
    - パターンのバリエーションの豊かさを「生産性」として算定
      - 計算の詳細は 吉川 (掲載予定) 及び [配布資料] 参照
    - ➔ 構成パターンの生産性の合計を発話の「情報量」として算定
- 分析 [2]
  - 一発話当たりの平均パターン数の推移を見る
    - 発話当たりのパターン数は発話の「複雑さ」の指標
    - ただし: パターン数は発話長に比例する
      → 発話長毎に平均化したものの平均の推移を見る

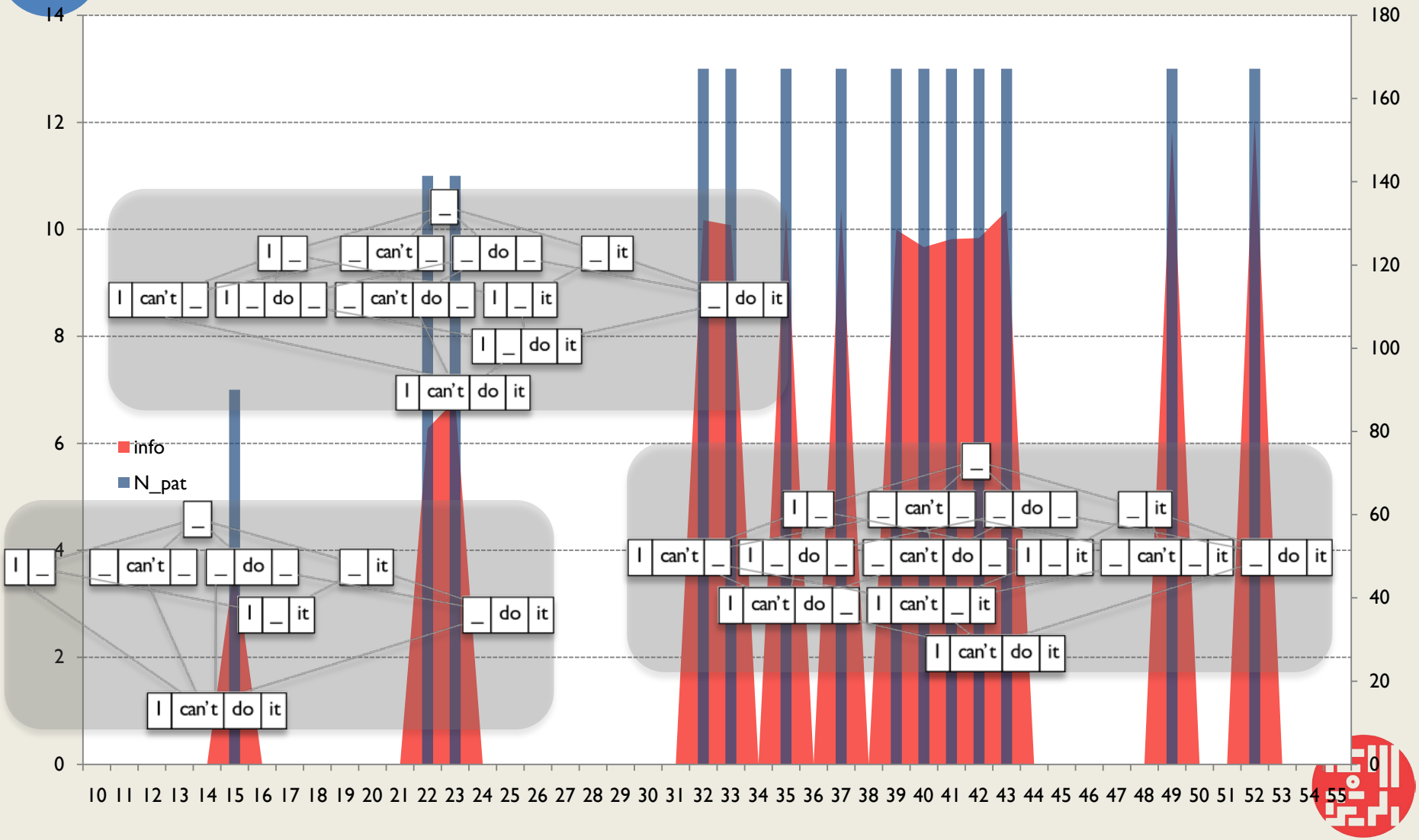
[1] の分析対象

- Adam
  - 1) *I can't do it*
    2) *I don't know*
    3) *I got it*
- Eve
  - 1) *let me have it*
    2) *I want some more*
    3) *I want some more tapioca*
- Sarah
  - 1) *let me see*
    2) *I know it*
    3) *I don't know*

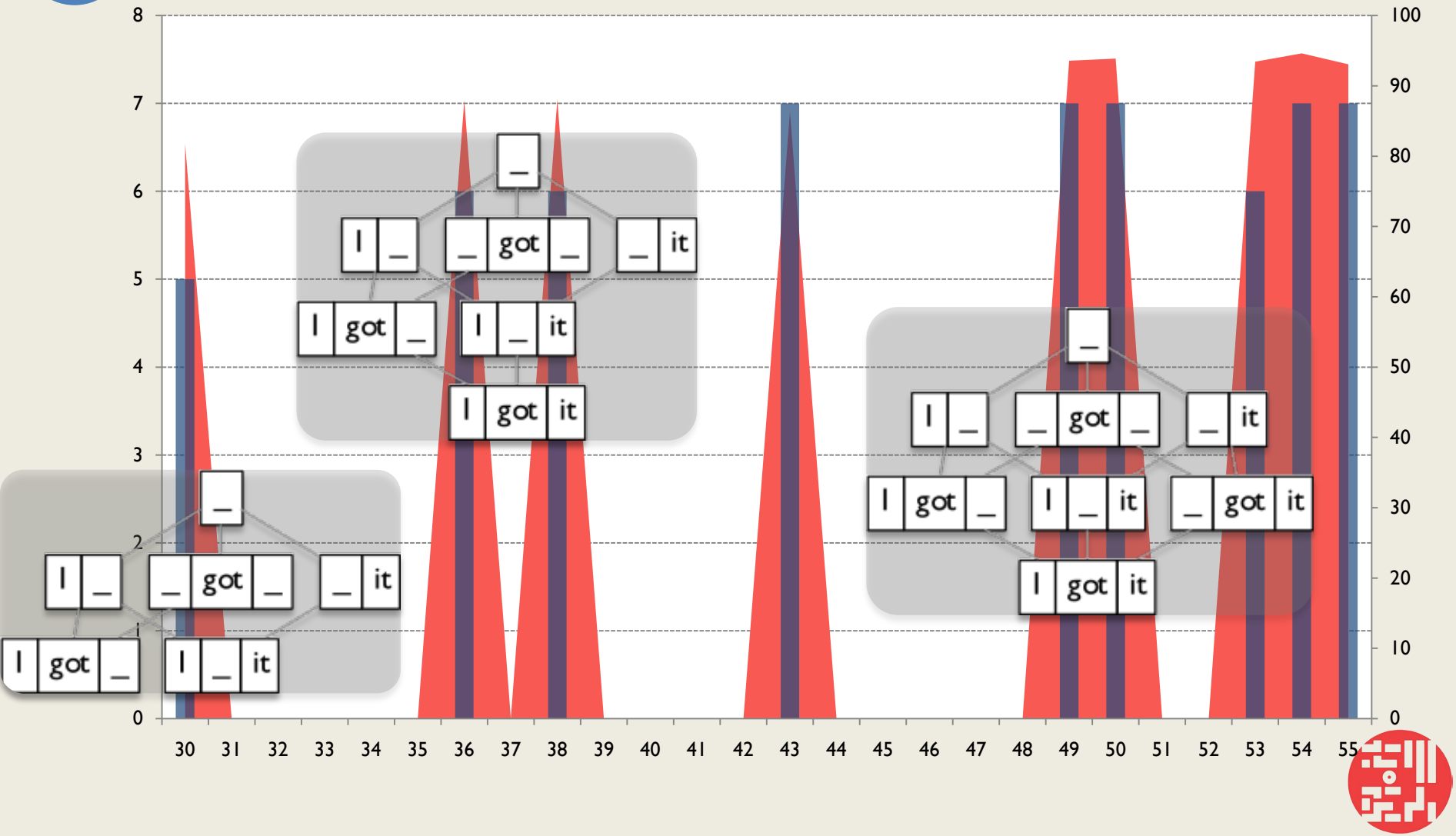
(3語以上の発話に限定)

5. 結果と考察

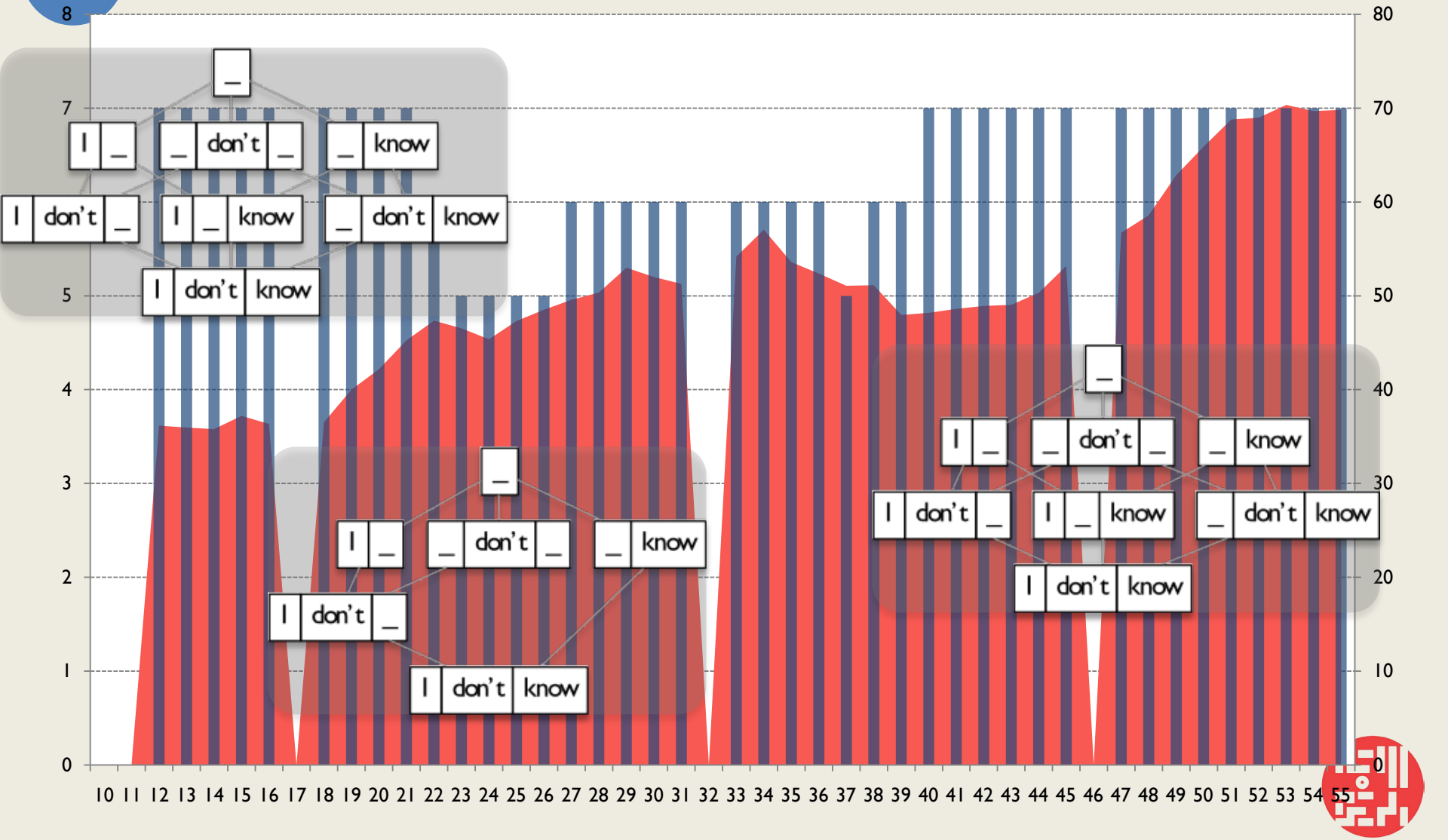
結果[1a] (Adam: I can't do it)



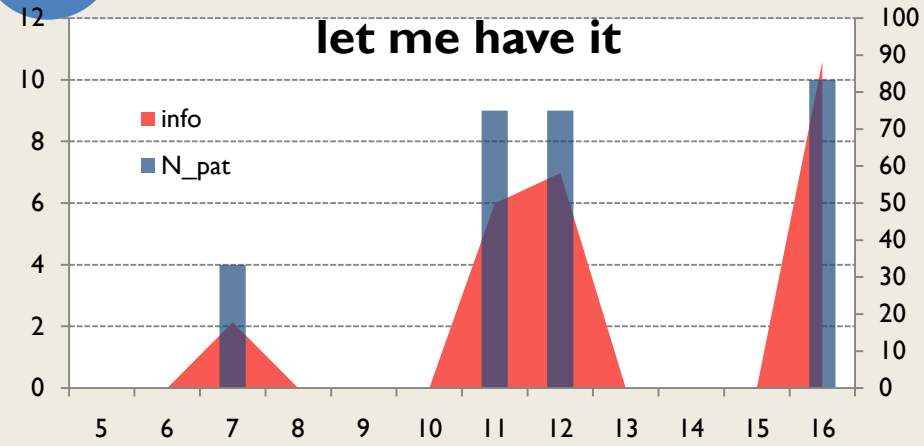
結果[1b] (Adam: I got it)

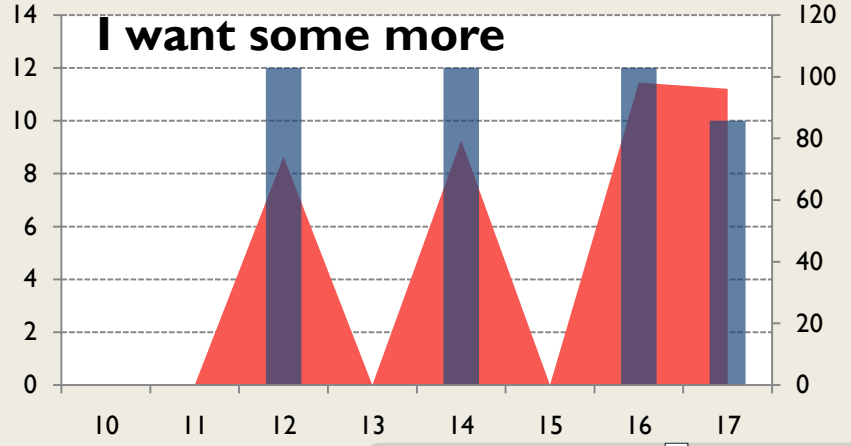


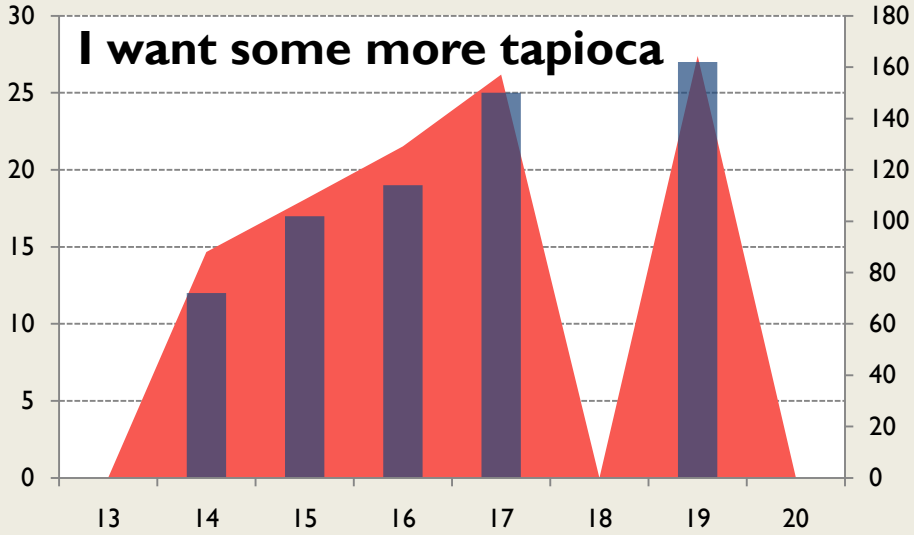
結果[1c] (Adam: I don't know)

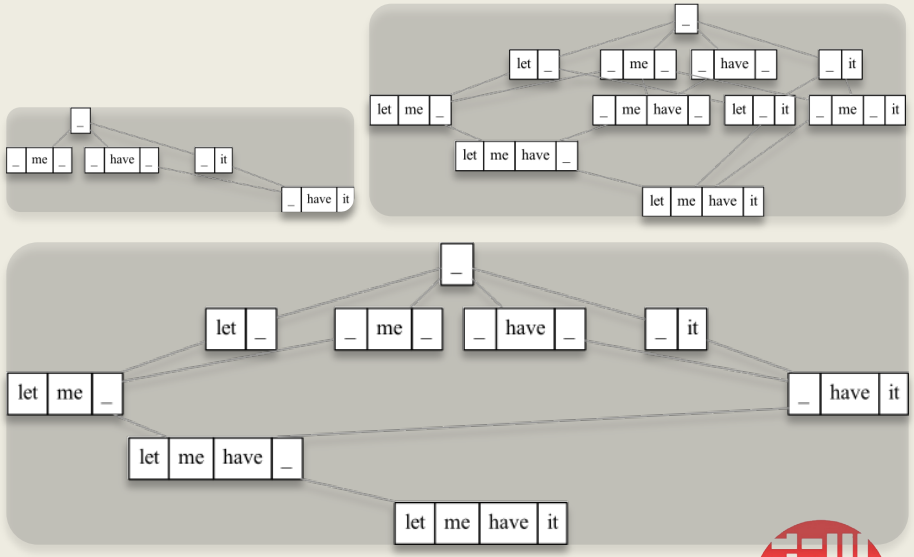


結果 [1d] (Eve)

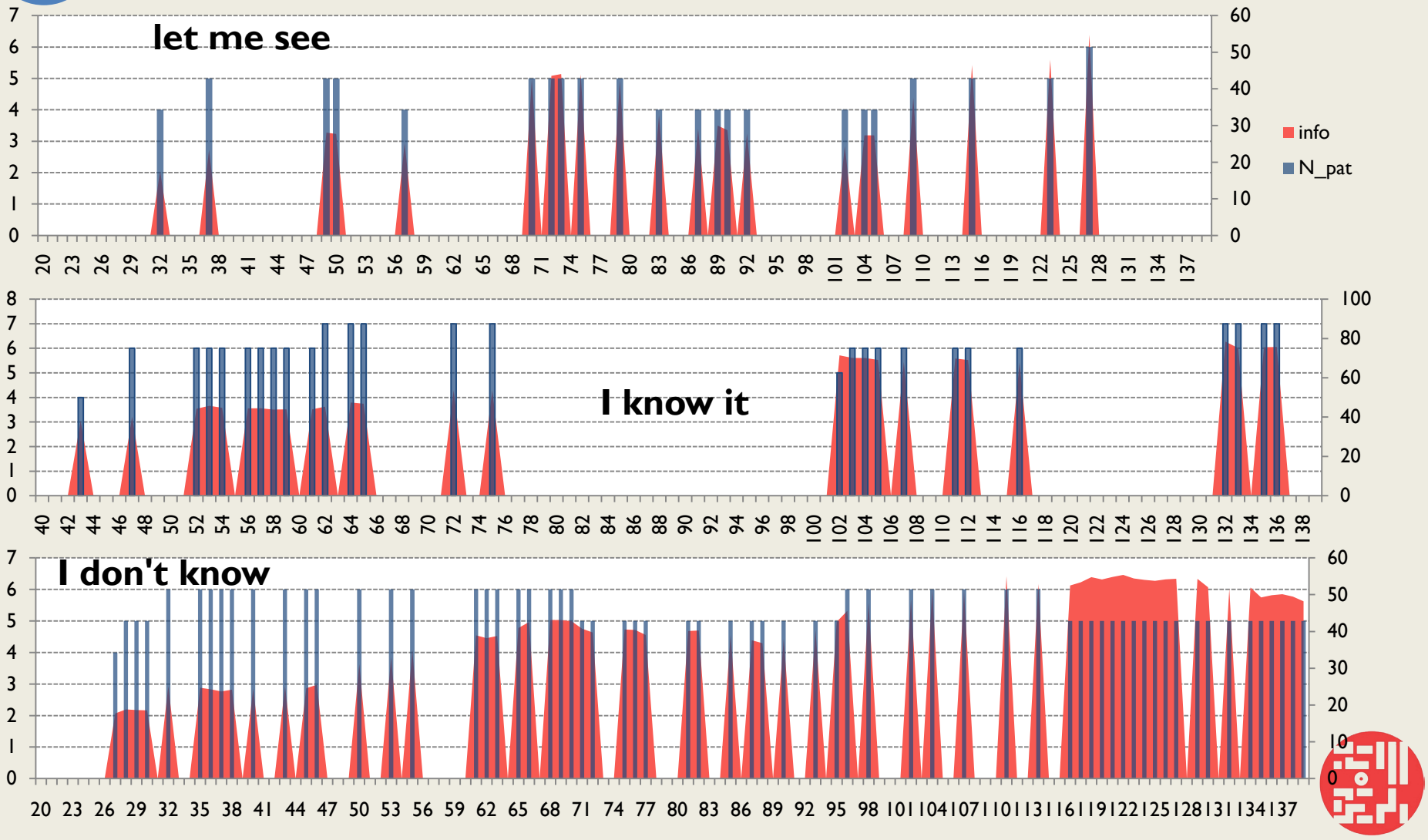

let me have it
info
N_pat


I want some more


I want some more tapioca


let me have it

結果 [1e] (Sarah)

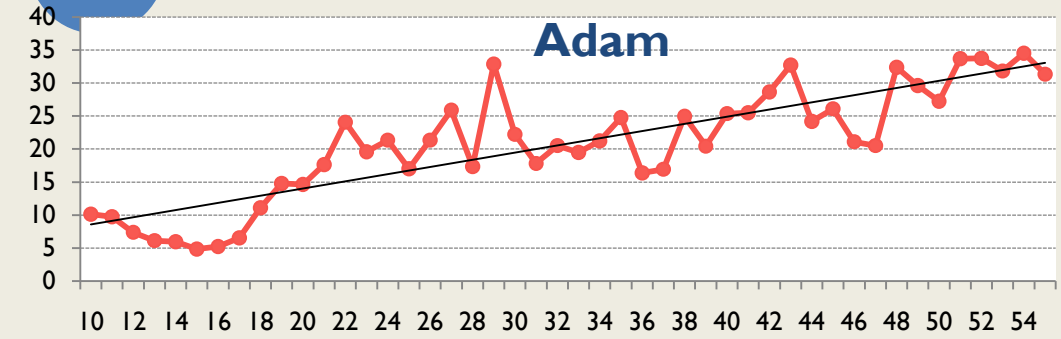


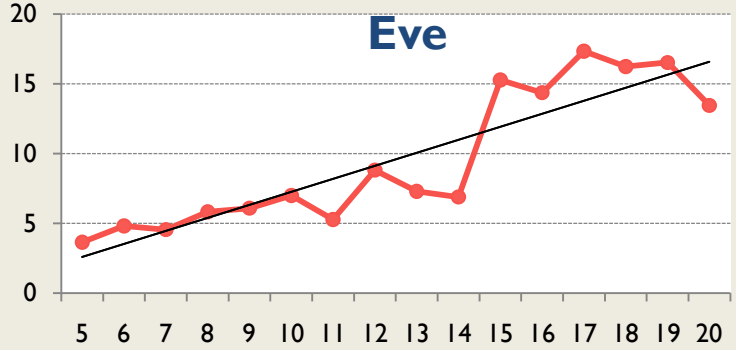
考察 [1]

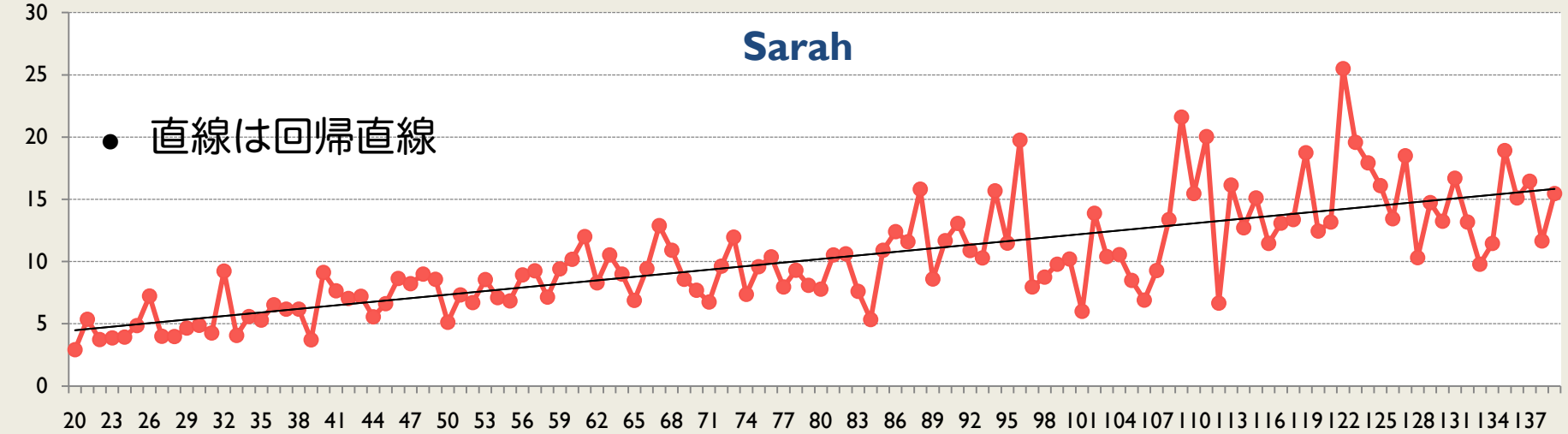
- 必ずしも構成パターンの数は増加しない
  - それでも個々のパターンの生産性は増加の傾向が顕著
    - 複雑さの上昇はある程度で打ち止め?
      ➔ パターンの組み合わせで生産的に発話?
- ただし: データの抜けの問題は影響しているはず
  - 年 (月) 齢が上がれば発話する頻度が増える
  - 収録されている時間以外の発話の割合が増える
    - 発話量はほぼ一定

    ➔ 収録データ自体は年齢経過に伴ってどんどん疎に
  - パターンのバリエーション現象 → 最適化で削減?

結果と考察 [2]







- 直線は回帰直線

☼全体の傾向としては発話の複雑度の上昇を確認

- Sarah の上昇傾向は恐らくファイルの分割数の影響

6. 結語

まとめと課題

- 本発表では
  - 統語発達を「漸進的」と考えた場合の記述問題を提起
  - 発話の履歴を「文脈」として利用することを提案
  - 文脈を利用した発話構造解析手法としてのPLMを導入
  - PLM を用いたBrown コーパスの解析結果を提示
  - 全体の傾向として発話構造の複雑化の軌跡を捕捉

  した
- 課題
  - 記述の断片性
  - 技術的な未成熟
    - 解析方法 / プログラミング技術

謝辞・参考文献

参考文献

- 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.
- Kuroda, K. 2009. Pattern lattice as a model for linguistic knowledge and performance. In *Proceedings of the 23rd pacific asia conference on language, information and computation* (pp. 278–287).
- 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.
- Tomasello, M. 2003. *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA.: Harvard University Press.
- 吉川正人. 2010. 「語」を越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』98, 221–207.
- 吉川正人. 掲載予定. スキーマの計算理論を求めて:漸進する統語発達過程の記述問題とその解法. 山梨正明(編) 『認知言語学論考 第10巻』東京: ひつじ書房.

謝辞

☼以下の方々にこの場を借りて謝意を表します

- シンポジウムメンバー
  - 大谷 直輝氏 (埼玉大学)
  - 鈴木 大介氏 (京都大学大学院 / 日本学術振興会)
  - 伊澤 宜仁氏 (慶應義塾大学大学院)

- メンバー以外 (五十音順)
  - 淺尾 仁彦氏 (SUNY Buffalo / 京都大学大学院)
  - 井上 逸兵氏 (慶應義塾大学)
  - 黒田 航氏 (京都大学非常勤 / 京都工芸繊維大学非常勤)
  - 佐治 伸郎氏 (日本学術振興会)
  - 長谷部 陽一郎氏 (同志社大学)

ご清聴有難うございました。