Finding Syntax in Human Encephalography with Beam Search

John Hale[♠,△] **Chris Dyer**[♠] **Adhiguna Kuncoro**[♠,♣] **Jonathan R. Brennan**[♢]

[♠]DeepMind, London, UK

[♣]Department of Computer Science, University of Oxford

[♢]Department of Linguistics, University of Michigan

[△]Department of Linguistics, Cornell University

{jthale,cdyer,akuncoro}@google.com jobrenn@umich.edu

ACL 2018 best paper

読む人: 能地宏

著者達の関連研究 (ACL 2018)

LSTMs Can Learn Syntax-Sensitive Dependencies Well, But Modeling Structure Makes Them Better

Adhiguna Kuncoro♠♣ Chris Dyer♠ John Hale♠♡
Dani Yogatama♠ Stephen Clark♠ Phil Blunsom♠♣

本論文

Finding Syntax in Human Encephalography with Beam Search

John Hale♠,△ Chris Dyer♠ Adhiguna Kuncoro♠,♣ Jonathan R. Brennan♢

神経言語学

- 二つの論文は思想が共通している
- 前者の論文と、関連研究との流れも踏まえつつ、本論文の貢献をまとめる

著者達が主張したいこと

- LSTM は文の構造を (自動で) 高精度に捉えられると言われているが…

- LSTM でなく、**明示的な句の階層構造**をモデル化することの利点が存在
  - (文生成における) **長距離依存関係の解決をより正確に行える**

    LSTMs Can Learn Syntax-Sensitive Dependencies Well, But …

    Kuncoro, Dyer, Hale, Yogatama, Clark, and Blunsom

本論文

- (文理解において) **句構造モデルの方が人間の認知モデルとして妥当**であることを人の脳波との相関を見ることで検証

  Hale, Dyer, Kuncoro, and Brennan

⇒ LSTM は強力だが、人間の認知モデルとして妥当とは言えない。人間は明示的に文の構造を認識している。

背景1: LSTM の文法識別能力

LSTMs can Learn Syntax-Sensitive Dependencies Well

- Agreement task による LSTM の文法識別能力の評価が (一部) 盛ん
  - Linzen et al. (2016), Enguehard et al., (2017), Gulordava et al. (2018), etc.
  - LSTM は**線形モデル**だが、単語間の**長距離依存関係**をそこそこ正確に捉えることが可能 ⇒ **言語モデルに明示的な階層構造は必要ないのでは？**

昨年の松林さんのスライド

単数／複数

The keys to the cabinet _____

Tal Linzen

- 脳波の予測において PCFG より RNN の方が高精度 (Frank et al., 2015)

⇒ **人の文処理は明示的な階層構造の認識を伴わないのではないか？**

二つの論文を通しどちらもそうとは言えないことを主張

背景2: 計算心理言語学

Computational Psycholinguistics

- 従来の心理言語学: 実験室で特定のパターンの文に対する反応を観察
  - e.g., 書記が代議士が首相がうたた寝したと抗議したと報告した
  - 問題点: 特定の現象に対する理解が得られても、**人間の文理解に対する統一的なモデル (= broad-coverage model)** は得られない (**理論の頑健性**)
- 計算言語学との融合 (2000~):
  - 目標: **人が日常の文章を見聞きする際の反応を一つのモデルで**統一的に説明
  - 研究者: Brian Roark, Frank Keller, Vera Demberg, William Schuler, etc.
- 評価方法:
  - 過去の研究: **モデル (parser) の動作による人の読み時間の推定**が主流
  - 本研究: **parser の動作による人の脳波の動きの推定 (より直接的シグナル)**

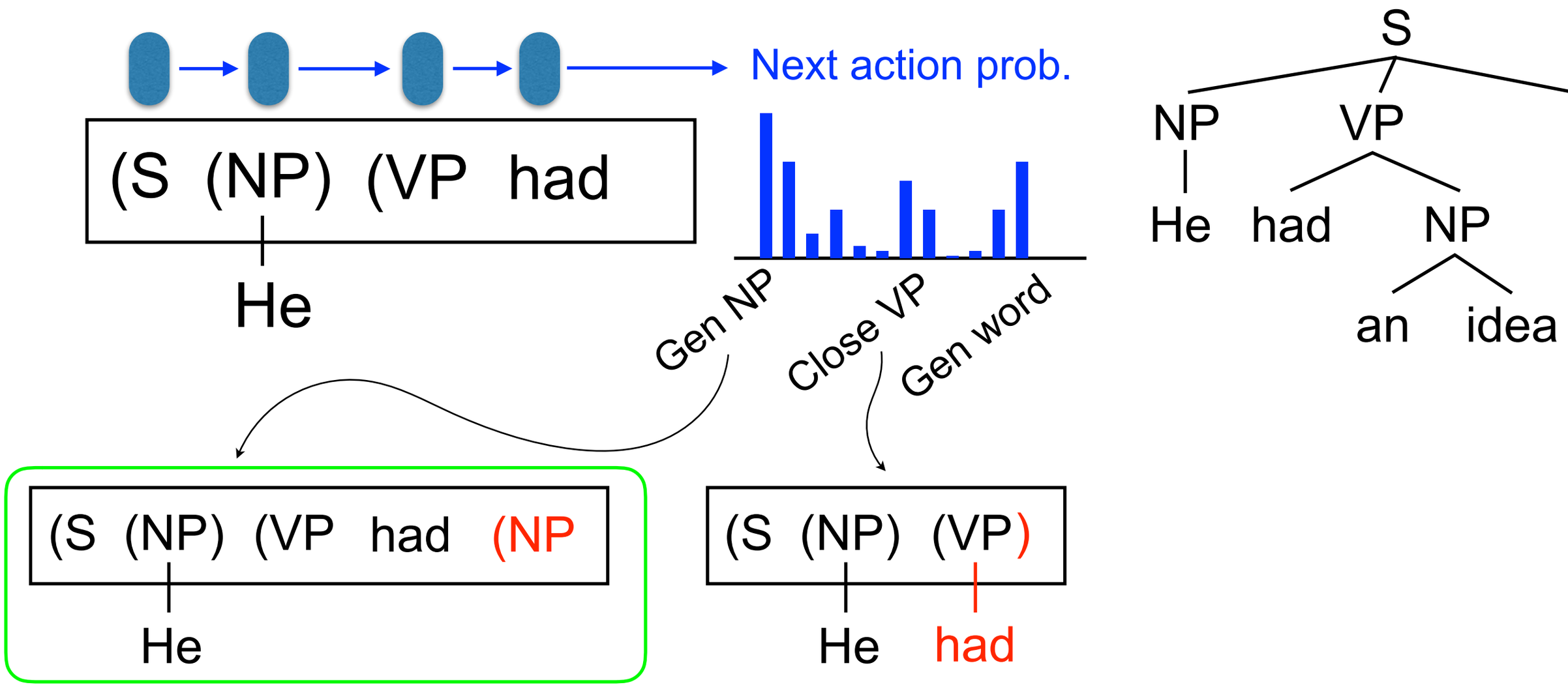
研究を一言で: 彼らの parser (RNNG) の認知的妥当性を脳波により検証

RNNG in 1 minute

Recurrent neural network grammars

(Dyer et al., 2016; Kuncoro et al., 2017; Choe & Charniak, 2016; Stern et al., 2017)

LSTM を用いたトップダウンなニューラル句構造生成モデル



- **スタックLSTM: スタック中の部分木のベクトルを入力とする LSTM**

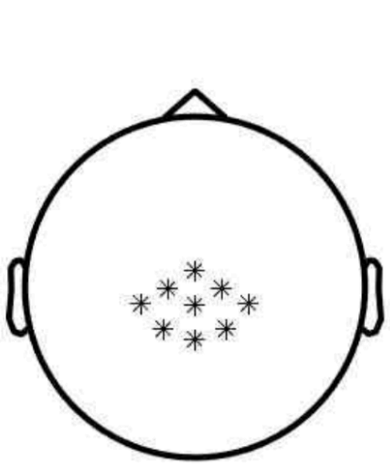
脳波のモデル化とは？

脳関係はアヤシイ略語が多い (EEG, MEG, ERP, P600, etc.)

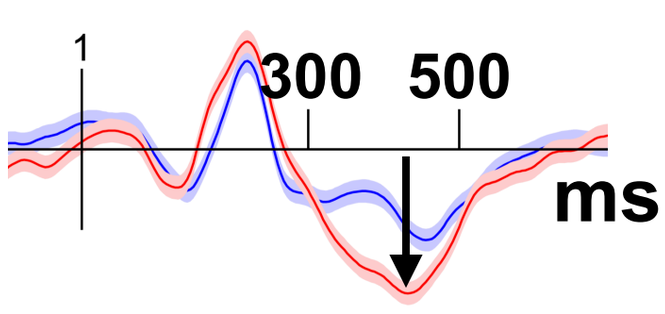
本研究で重要なのは ERP (event related potential)

⇒ 脳が特定の処理をする時、特定の部位に特定のタイミングで生じる電位

部屋に年老いた夏休みが…



N400
300–500 ms



Frank et al. (2015)

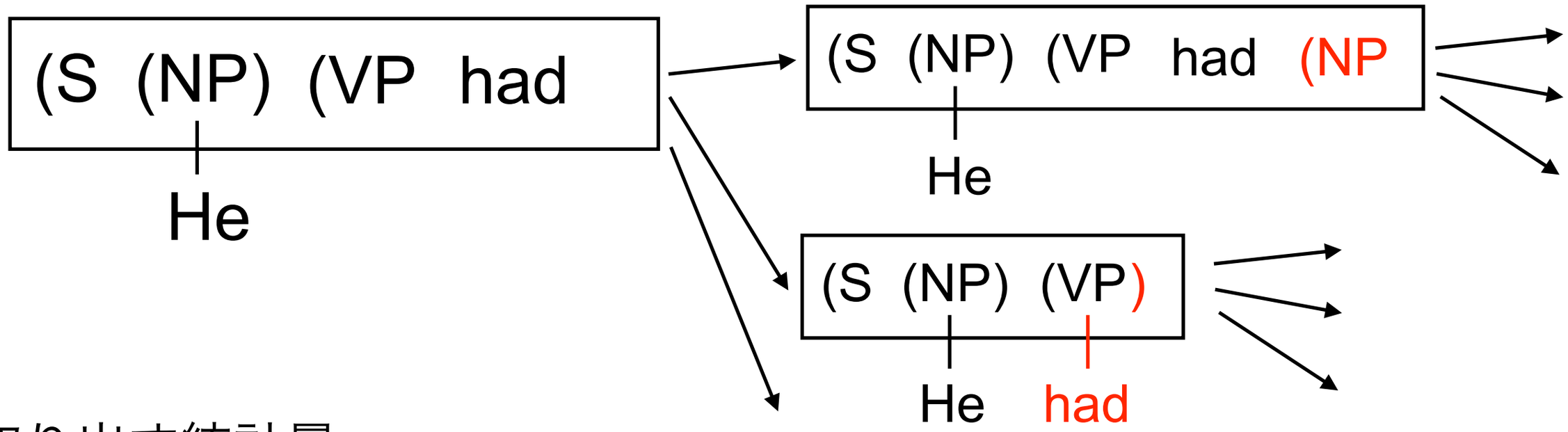
N400 (と呼ばれる ERP):

不自然な語を聞いて~400msに中央部で生じる負の電位

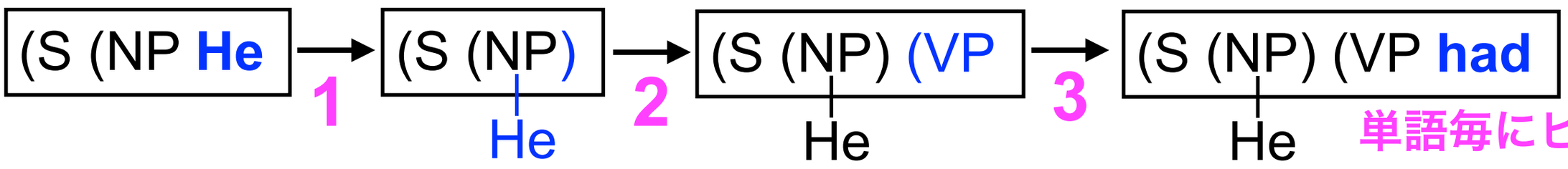
アプローチ: RNNG (or LSTM) が文を処理する際の統計量によって同じ文を人が聞いた際生じる ERP を回帰できるか、評価する

RNNG からの**単語の統計量**の抽出

- 探索法: beam search + ヒューリスティクス (Stern et al., 2017)



- 取り出す統計量
  - **Distance:** 前の単語と次の単語の間に発生したアクション数



  - **Surprisal:** 次の単語の予測確率の対数 (LSTM を使って説明)

どちらも文中の複雑性の度合いを捉える統計量

Surprisal

$$\mathrm{Surprisal}(w_i) = -\log \boxed{P(w_i|\mathrm{context})}$$

LSTM の場合、単純に次の単語の予測確率

- 単純な言語モデルの方が理解しやすい

$$P(w_i|\mathrm{context}) = P(w_i|w_1 \cdots w_{i-1})$$

- 次の単語の確率値が小さいほど、大きくなる量 (次の単語に対する驚き)
- 予測できない単語が出現すると大きくなる

- RNNG (木構造モデル) の場合: 木構造を周辺化した量

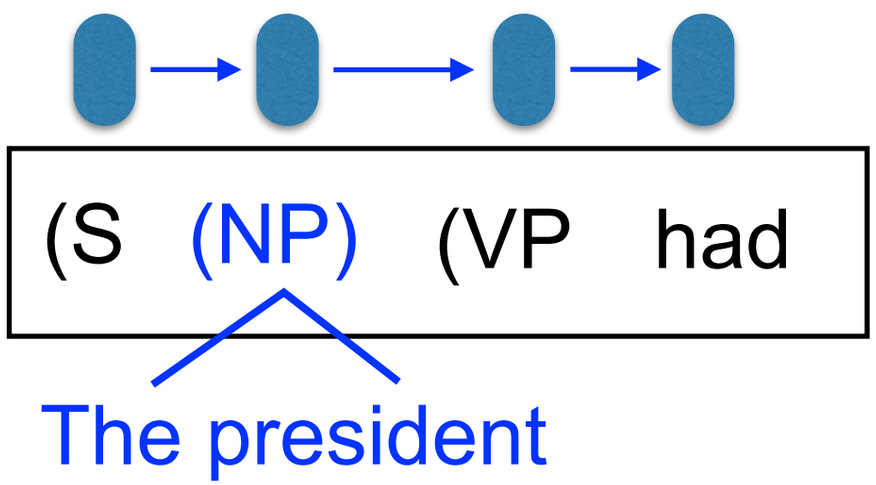
$$P(w_i|\mathrm{context}) = \sum_{t \in T} P(w_i|w_1 \cdots w_{i-1}, t) \approx \sum_{t \in \mathrm{Beam}} P(w_i|w_1 \cdots, w_{i-1}, t)$$

実験設定

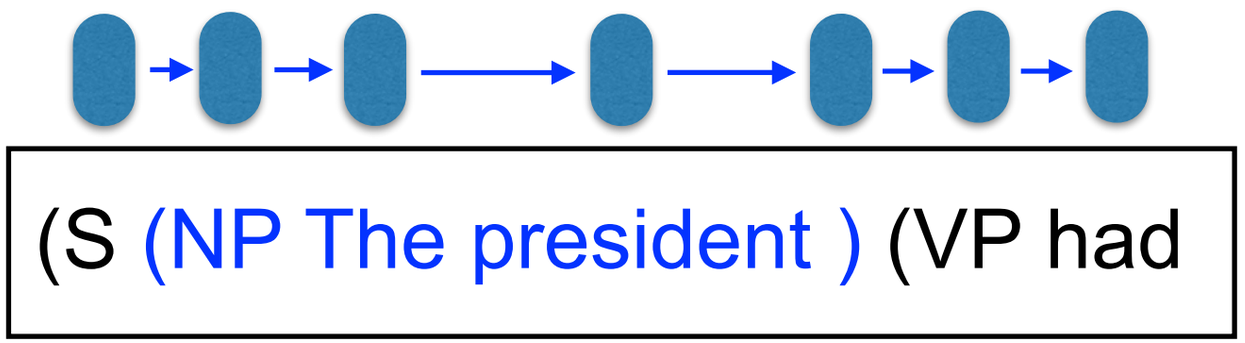
- 脳波の計測:
  - 被験者: ミシガン大 (Brennan) の native speaker 33人 (頭上に61個の電極)
  - データ: “アリスの冒険” の第1章の読み聞かせ
  - ノイズを削減するための様々な工夫 (フィルタ, ICA, etc.)
- 評価:
  - LSTM の **surprisal**, RNNG の **surprisal**, RNNG の **distance** などのうち、どれが最も特徴的な ERG (脳波の反応) を捉えるか
  - Linear regression の結果を尤度比検定により比較
  - どのモデルも単純な文の複雑性 (文長など) の要素は含める
  - LM, RNNG の学習: “アリスの冒険” の他の章を用いる
    - RNNG の訓練データ: Stanford parser が予測した木を正解とみなす

3つのモデル

- LSTM: 言語モデル (256 units)
  - 統計量: Surprisal のみ
- RNNG: 逐次的な木構造生成モデル (170 units)
  - 統計量: Surprisal, Distance, Entropy (省略)
- RNNG-comp: RNNG の劣化版 (Choe & Charniak 2016; Stern et al., 2017)
  - スタック中の部分木を集約せず、単なる S 式の列に LSTM を適用

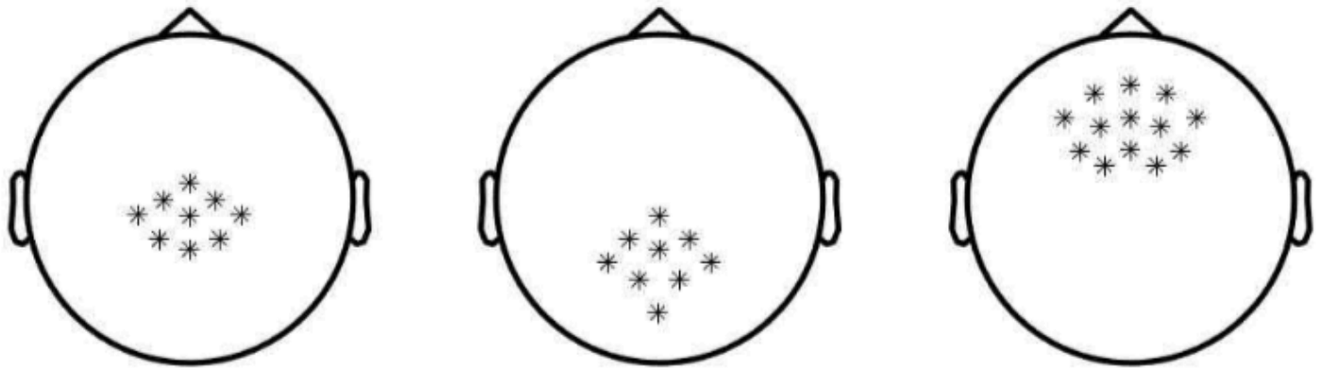


RNNG



RNNG-comp

尤度比検定: 主な結果



	$\text{RNNG}_{-\text{comp}} > \text{LSTM}$			$\text{RNNG} > \text{RNNG}_{-\text{comp}}$		
	χ^2	df	p	χ^2	df	p
DISTANCE, "P600" region						
ビーム幅 $k = 200$	13.409	1	**0.00025**	4.198	1	0.04047
ビーム幅 $k = 400$	15.842	1	**<0.0001**	3.853	1	0.04966
$k = 600$	13.955	1	**0.00019**	3.371	1	0.06635
SURPRISAL, "ANT" region						
ビーム幅 $k = 100$	3.671	1	0.05537	13.167	1	**0.00028**
ビーム幅 $k = 200$	3.993	1	0.04570	10.860	1	**0.00098**
$k = 400$	3.902	1	0.04824	10.189	1	**0.00141**

RNNG > RNNG-comp > LSTM の順で対象とする ERP を良く予測

論文の主張

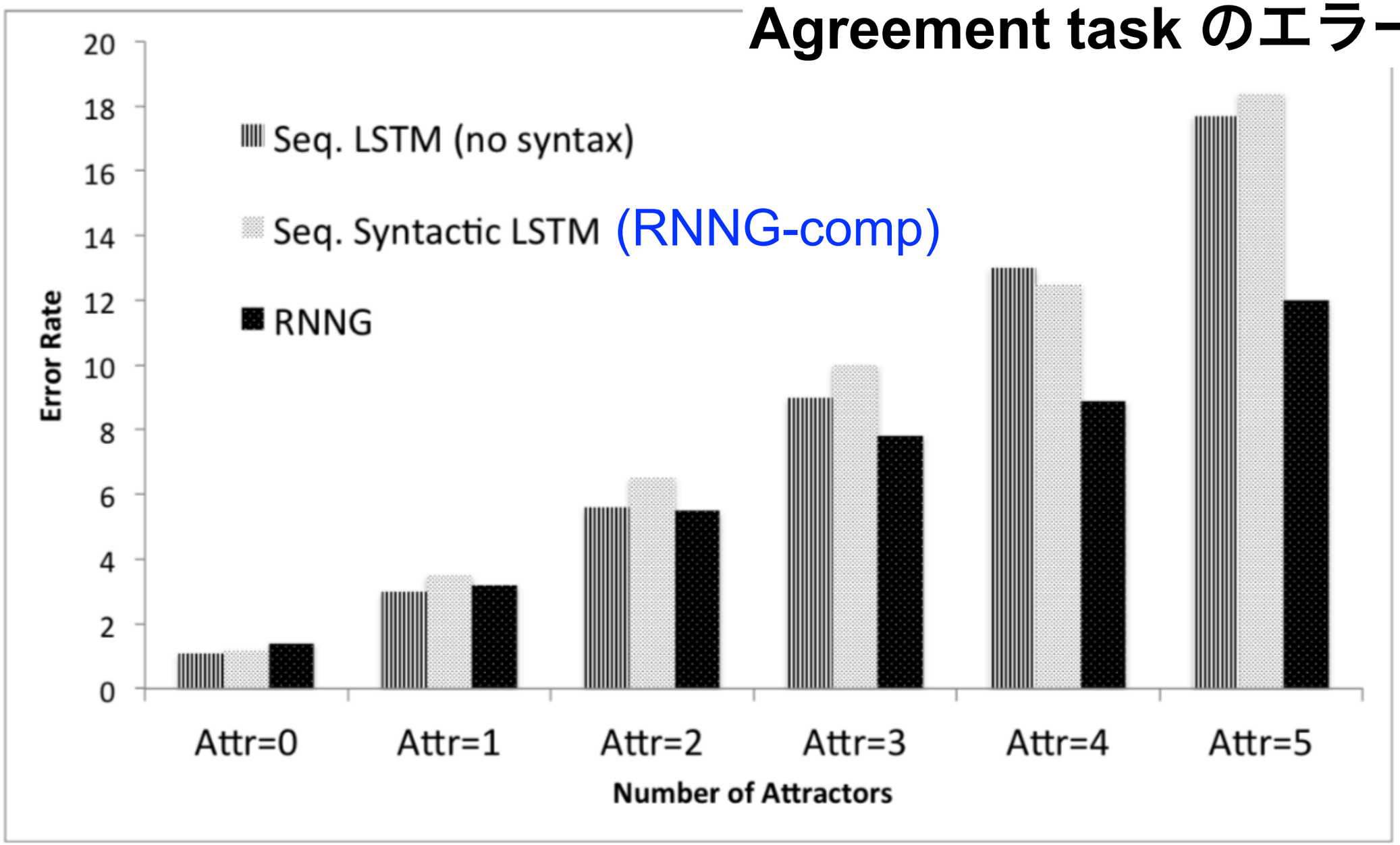
- RNNG > RNNG-comp ⇒ スタック中のシンボル合成操作が重要
  - RNNG-comp: P600 成分の検出能力は LSTM より上
  - しかし early peak (ANT) の検出は RNNG に劣る

- 同じモデルでも、異なる統計量 (複雑度) が異なる ERP を検出
  - P600 は **distance**, ANT は **surprisal** により検知される
  - 脳内の異なる現象が一つのモデルで説明できる、という点で、認知的に望ましいと言えるのでは

もう一つの論文について

LSTMs Can Learn Syntax-Sensitive Dependencies Well, But Modeling Structure Makes Them Better

Agreement task のエラー率



Agreement を予測する際も RNNG で合成操作を行うことが重要

所感

- なぜ best-paper か？
  - ものすごく革新的か、というと、そうではない
  - 人間の文処理を模倣するためのモデルは過去にも存在 (Demberg et al. など)
  - **神経言語学者と手を組んで NLP の最新の成果 (RNNG + 逐次的な探索) を丁寧に認知実験に適用**し成果を得た点が評価された (?)
- 突っ込みどころは多々存在
  - アリスだけで学習した LM は弱すぎるのでは？
  - Stanford parser の予測した結果から学習して良いのか？
  - “ANT” 領域に着目する理由が書かれていない (引用先が未出版)

本研究だけでは人間の文処理について確かなことは何も言えない

今後のベースラインとして重要: 今回の枠組みで、モデルを精緻化することでどこまで人間に近づけるか？言語学に feedback を与えられるか？