# アテンションが記憶想起の認知モデルたりうるならば、記憶の表現としては何が妥当か？

吉田遼 [1]　磯野真之介 [1]　梶川康平 [1]　染谷大河 [1]
杉本侑嗣 [2]　大関洋平 [1]
[1] 東京大学　[2] 大阪大学
{yoshiryo0617, isono-shinnosuke, kohei-kajikawa,
taiga98-0809, oseki}@g.ecc.u-tokyo.ac.jp
sugimoto.yushi.hmt@osaka-u.ac.jp

## 概要

近年の計算心理言語学では、アテンションの人間の記憶想起のモデルとしての妥当性が検証されている。しかし、Transformer のアテンションが扱う表現の単位はトークンであるのに対し、伝統的に計算心理言語学では人間の文処理は統語構造の構築を伴うとされてきた。本研究では、統語構造を表現単位として扱う Transformer (Transformer Grammar, TG) のアテンションが、人間の記憶想起のモデルとして妥当なのかを検証する。結果、TG のアテンションは Transformer を上回る読み時間の説明力を達成し、アテンションの記憶想起アルゴリズムとしての妥当性を裏付けるとともに、統語構造を表現単位として想定することの重要性を示した。

## 1　はじめに

「自然言語処理で用いられる言語モデルは、人間の文処理モデルとしても妥当であるか」は、計算心理言語学の一大トピックである。ここ 10 年ほど、この問いは主に「言語モデルの次単語予測は人間の予測処理のモデルとして妥当なのか」という側面から取り組まれてきた（例： [1]）。具体的には、サプライザル [2, 3] を橋渡し仮説とした、言語モデルの算出確率と人間の読み時間などとの対照である。しかし、より近年、自然言語処理での Transformer [4] の活躍を受け、新たな側面からの検証が活発になり始めている。それは、「アテンションは人間の**記憶想起**のモデルたりうるか」という側面からの検証である [5, 6]。ここでは、アテンションのクエリ・キーによる先行トークンの重みづけを、人間の「手がかりに基づく記憶からの取り出し (cue-based retrieval, [7])」のモデルとみなし、例えばアテンション重みのエントロピーにより取り出しの困難さ（干渉度合い）を定量化することが試みられている。

前述の言語モデルによる予測処理のモデリングの多くが Marr の 3 つのレベル [8] における計算理論のレベル付近に位置付けられるのに対し、アテンションによる記憶想起のモデリングは、より具体的な表現とアルゴリズムのレベルに明確に位置付けられる。このレベルにおいては、アテンションというアルゴリズムとそれが操作の対象とする**表現（つまり、トークン列）**は不可分であり、その総体がモデリングの対象となる [9]。一方、伝統的な（計算）心理言語学では、人間の文処理は線形な単語列の処理のみには還元できず、逐次的な**統語構造**の構築を伴うことが示されてきた。例えば、予測処理のモデリングでは、統語構造を陽に構築する言語モデルは RNN や Transformer よりも人間の脳活動の説明力が高いことが示されているし [10, 11]、心理言語学における記憶減衰のモデリングでも、統語理論に基づく記憶単位の妥当性が主張されている [12]。

これらを踏まえると、アテンションが人間の文処理における記憶想起一般のアルゴリズムたりうるのであれば、トークン列を表現単位としたモデルだけでなく、統語構造を表現単位としたモデルについての妥当性も期待できるであろう。本研究では、統語構造を表現単位として扱う Transformer (Transformer Grammar, TG) のアテンションが、人間の記憶想起のモデルとして妥当なのかを検証する。結果、TG のアテンションは Transformer を上回る読み時間の説明力を達成し、アテンションの記憶想起アルゴリズムとしての妥当性が裏付けると共に、統語構造を表現単位として想定することの重要性を示した。

# 2 背景

## 2.1 Normalized Attention Entropy (NAE)

人間の文処理は記憶想起を伴い、ある単語（例：動詞）を入力とした際、様々な手がかりに基づき作業記憶から要素（例：その項）が取り出されるとされている。その証拠として、目的とする要素に類似した要素が文中に存在すると取り出しが干渉され、処理負荷が増大する（例：読み時間が長くなる）ことが示されている [7]。[5] は、アテンションのクエリ・キーによる先行トークンの重みづけをこの取り出しのモデルと捉え、アテンション重みのエントロピー (Attention Entropy, AE) で干渉度合いを定量化することを提案した。

AE は特定の刺激文における干渉効果のモデリングのために提案されたものだが、[6] はこれを自然な刺激文一般へと拡張するため、(i) トークン数に応じた最大エントロピーによる標準化及び (ii) 和が 1 となるための標準化を施した (**Normalized AE, NAE**) :[1)]

$$\mathrm{NAE}_{l,h,i} = \frac{\mathbf{a}_{l,h,i[1:i-1]}^{\top}}{\log_2 (i-1)\mathbf{1}^{\top}\mathbf{a}_{l,h,i[1:i-1]}} (\log_2 \frac{\mathbf{a}_{l,h,i[1:i-1]}}{\mathbf{1}^{\top}\mathbf{a}_{l,h,i[1:i-1]}}) \quad (1)$$

ここで、$\mathbf{a}_{l,h,i}$ は $l$ 層目・$h$ 番目のヘッドにおける $i$ 番目のトークンをクエリとした際のアテンション重みのベクトルである。[2)] 本研究では、この NAE を橋渡し仮説として採用し、アテンションによる人間の記憶想起のモデリングを行う。[3)]

## 2.2 Transformer Grammar (TG)

**Transformer Grammar (TG; [14])** は、単語列と対応する統語構造（例：`(S (NP The blue bird NP) (VP sings VP) S)`）の生成モデルであり、以下の 3 つのアクション列に対する確率分布をモデル化する：

- `(X`: 非終端記号（句ノード）`(X` を生成する
- `w`: 終端記号（単語）`w` を生成する
- `X)`: `X)` を生成し、直近の開いた句ノードを閉じる

ここで、TG において最も主要な特徴は、`X)` の生成直後に、閉じられた句ノードを親とする部分木のベクトル表現を取得し (図 1 COMPOSE)、以降の次アクション予測 (図 1 STACK) ではその表現のみを参照するという点である。すなわち、TG のアテンションでは、統語構造（における部分木）を表現の単位とする。

先行研究では、TG の算出確率は、普通の Transformer よりも人間の文法判断との一致率が高いこと [14] や、Transformer に説明できない脳活動を説明できること [11] が示されている。本研究では、TG のアテンションが人間の読み時間に対して説明力を持つかどうかを検証する。

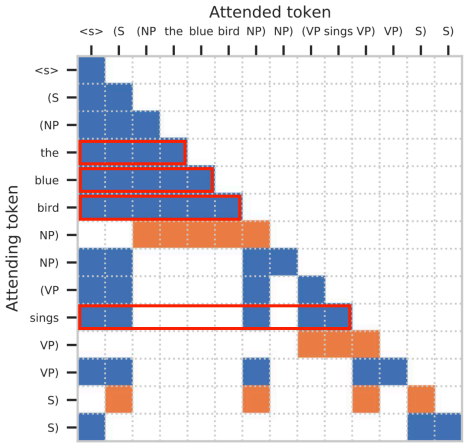


**図 1** TG のアテンション。[14] より引用・改変。COMPOSE で統語構造における部分木の表現を得て、以降の STACK では部分木表現を参照する。各単語の NAE の算出に用いるアテンションを赤枠で示した。

# 3 TG の NAE の算出

TG で NAE を算出するに当たっては、以下の 2 つについての仮定を置く必要がある：

1. 推論時には文のみが与えられるが、その背後の統語構造としてどのようなものを仮定するか？
2. 単語以外にも (X や X) をクエリとしたアテンションが存在するが、これらの負荷をどのように単語に割り当てるか？

本研究では、そのそれぞれについて以下の仮定を設定する：

1-A. 人間は逐次的な文処理に際して部分並列的に統語構造構築を行っている [15, 10] との知見に

1) [6] は、AE では読み時間を説明する回帰モデルが収束しないことを示している。

2) [6] は、複数のアテンション重みの標準化手法を対象に探索的に NAE を算出しているが、本研究では、自己ペース読み時間コーパス (§4) で最も説明力が高かったノルムによる標準化 [13] を採用する。

3) [6] は、NAE 以外にも連続する時刻でのアテンション重みの距離に基づく指標を複数提案しているが、本研究では、(i) TG は時刻によって要素数が変化するため距離が自明に定義できないこと (§2.2)、(ii) [6] では自己ペース読み時間コーパス (§4) で NAE の説明力が高いことが示されていること、から NAE のみを用いる。

基づき、単語同期ビームサーチ [16] により導出された、複数の統語構造を仮定する。[4)]

2-A. 単語を入力とした際の記憶想起を対象とするため、各単語をクエリとした際のアテンションのみを対象とし、(X や X) をクエリとした際のアテンションについては考慮しない（図 1）。

これらの仮定のもと、TG の NAE を、(i) 各単語時点で同期されたビーム内の統語構造のそれぞれについて、当該単語をクエリとした際の NAE を算出し、(ii) それを各統語構造の確率で重み付けて足し合わせたもの、として定義する：

$$\mathrm{NAE\_TG}_{l,h,i} \coloneqq \frac{\sum_{t \in \mathrm{Beam}_i} p(t) \cdot \mathrm{NAE}^{t}_{l,h,i}}{\sum_{t \in \mathrm{Beam}_i} p(t)} \tag{2}$$

## 4 実験設定

**言語モデル** 252M の TG・Transformer を用いた。パラメタは全て [14, 11] にならった（付録 A）。NAE の算出に際しては、[6] にならい、最終層のみを対象とし、各ヘッドの NAE を足し合わせた。

**学習データ** BLLIP-LG（42M トークン・1.8M 文）を用いた。前処理等は全て [14, 11] にならった。TG の学習には [19] により句構造パーサ [20] でパースされた統語構造を用い、トークナイザには学習セットで学習された 32K の SentencePiece トークナイザ [21] を用いた。

**読み時間データ** Natural Stories コーパス [22] を用いた。10 の物語・485 文・10,245 単語からなり、各単語には 181 人の英語母語話者による自己ペース読み時間が付与されている。[22] によるフィルタリングに加え、[6] にならって文頭および文末の単語を取り除き、766,467 のデータポイントがモデリングの対象となった。

**評価手法** 読み時間の予測への各言語モデルの NAE の貢献度を評価した。具体的には、対数読み時間を従属変数とし、(i) テキストのナイーブな特徴量、(ii) 両者の言語モデルのサプライザル、(iii) 統語理論に基づく記憶減衰のメトリクス [12]、[5)] を説明変数およびランダム効果とした回帰モデル（式 3；ベースライン回帰モデル）に対し、各言語モデルの NAE（tg_nae または transformer_nae）を説明変数およびランダム効果に加えた際の、説明力の向上分（対数尤度の増加分、ΔLogLik）を評価した。説明変数の詳細は付録 B に示した。単語がサブワードに分割される場合、[6] にならい単語内のサブワードの NAE については足し合わせた。

$$\begin{aligned}\log(\mathrm{RT}) \sim\ & \mathrm{zone} + \mathrm{position} + \mathrm{wordlen} \\ & + \mathrm{unigram} + \mathrm{bigram} + \mathrm{trigram} \\ & + \mathrm{transformer\_surp} + \mathrm{tg\_surp} + \mathrm{clt} \\ & + (1 + \mathrm{zone} + \mathrm{position} + \mathrm{wordlen} \\ & \quad + \mathrm{unigram} + \mathrm{bigram} + \mathrm{trigram} \\ & \quad + \mathrm{transformer\_surp} + \mathrm{tg\_surp} + \mathrm{clt} \\ & \quad \| \mathrm{participant}) + (1 \mid \mathrm{story})\end{aligned} \tag{3}$$

加えて、両者の言語モデルの NAE の説明力が包含関係にあるかどうかについても検証した。具体的には、両者の言語モデルの NAE を説明変数およびランダム効果として式 3 に加えた回帰モデルの ΔLogLik と、どちらか片方の言語モデルの NAE のみを説明変数およびランダム効果として式 3 に加えた回帰モデルの、ΔLogLik の差が有意かどうかを尤度比検定した。

## 5 結果・考察

### 5.1 TG の NAE は読み時間に対して説明力を持つか？

**表 1** TG・Transformer の NAE の読み時間の予測への貢献度 (ΔLogLik)。参考のため、標準偏差あたりの効果量を、同一の回帰モデルにおける統語理論に基づく記憶減衰のメトリクス (clt) の効果量と共に示した。*** は $p < 0.001$ を表す。読み時間の平均値は 335ms である。

| 説明変数 | ΔLogLik (↑) | 効果量 |
|---|---|---|
| tg_nae | **173** | 2.62 ms*** |
| clt | N/A | 0.709 ms*** |
| transformer_nae | 72.4 | 1.63 ms*** |
| clt | N/A | 0.754 ms*** |

TG・Transformer の NAE の読み時間の予測への貢献度を表 1 に示した。まず、どちらの言語モデルの NAE も、ベースライン回帰モデルに含まれるサプライザルなどの説明変数とは独立に、読み時間に対して有意に説明力を持っていた。これは、[5, 6] で主張されているように、アテンションによる先行要素の重みづけというアルゴリズムは、純粋に工学的な目的で開発されたにも関わらず、人間の記憶想起

4) TG のデフォルト実装には単語同期ビームサーチが搭載されていないため、RNNG [17, 18] を用いて各単語時点で同期された 10 の統語構造を導出した。アクションビームサイズは 100、ファストトラックは 1 に設定した。RNNG の学習には TG の学習と同様のデータ (§4) を用いた。

5) 心理言語学では、統語理論に基づく記憶減衰のメトリクスが複数提案されているが、Natural Stories コーパスでの説明力が確認されている Category Locality Theory (CLT; [12]) を採用した。

のモデルとしても一定の妥当性を持つことを示す。さらに、本研究の主眼とは異なるが特筆すべきは、NAE の説明力と、（先行研究では説明変数に含まれていなかった）統語理論に基づく記憶減衰のメトリクス (clt) の説明力が、確かに独立であることが確認できたという点である。この結果は、NAE が記憶想起の中でも記憶減衰ではなく干渉度合いを定量化している (§ 2.1) ことを裏付けているといえる。これまで、心理言語学では、自然な刺激文一般に利用できる記憶想起の干渉度合いのメトリクスは提案されておらず、NAE がその最有力であることをより精緻に確かめたという意味で意義のある結果である。そして、重要なことに、TG の NAE は Transformer の NAE に比べて読み時間に対する説明力が高かった。この結果は、アテンションをアルゴリズムとする記憶想起のモデルにおいては、表現単位として統語構造を採用する方がより支配的な要素を捉えられることを示している。

### 5.2 TG・Transformer の NAE の説明力は包含関係にあるか？

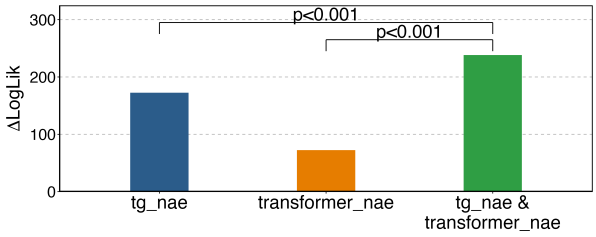


**図 2** NAE の説明力が包含関係にあるかについての検定結果。TG の NAE と Transformer の NAE がそれぞれ独立に説明する読み時間の分散が存在する。

TG・Transformer の NAE の説明力が包含関係があるかについての検定結果を図 2 に示した。両者の言語モデルの NAE を含む回帰モデルは、どちらか片方の言語モデルの NAE のみを含む回帰モデルよりも説明力が高く、またその差は共に有意であった。これは、TG の NAE が Transformer の NAE に説明できない読み時間の分散を確かに説明できているだけでなく、その逆、全体的な説明力は低い Transformer の NAE が、TG の NAE には説明できない独立した分散を説明できていることを意味している。心理言語学では、作業記憶からの要素の取り出しのモデルとして、動詞と項など統語関係に基づくもの [23] だけでなく、bag-of-words 的な意味的類似に基づくもの [24] も提案されているように、TG のアテンションと Transformer のアテンションは、それぞれ人間の記憶想起の異なる側面を捉えているモデルとして両立しうる可能性がある。

### 5.3 TG・Transformer の NAE はそれぞれいつ効果的か？

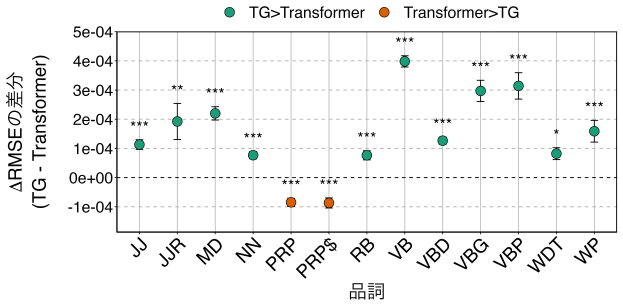


**図 3** 品詞ごとの読み時間予測の改善度 (ΔRMSE) の差分 (TG - Transformer)。改善度および改善度の差分が有意である品詞のみを抜粋した。*は $p < 0.05$、**は $p < 0.01$、***は $p < 0.001$ を表す（Bonferroni 補正後）。

TG・Transformer の NAE がそれぞれ読み時間のどのような側面の予測に効果的であるかを検証するため、Natural Stories コーパスに付与されている品詞タグ（付録 C）ごとに、ベースライン回帰モデルからの読み時間予測の改善度 (Δ Root Mean Squared Error, ΔRMSE) の差分 (TG - Transformer) を示した（図 3；改善度および改善度の差分が有意である品詞のみを抜粋）。[6] 特に興味深い結果として、TG の NAE は動詞群 (VB, VBD, VBG, VBP) で、Transformer の NAE は代名詞群 (PRP, $PRP) で、一貫して改善度が高いという点が挙げられる。この結果は、動詞を入力とした際の想起（例：主語の想起）と、代名詞を入力とした際の想起（例：先行詞の想起）は、それぞれ統語構造/トークン列を表現単位とするアテンションという異なる認知モデルによって、より高い妥当性で捉えられることを示している。

## 6 おわりに

本研究では、統語構造を表現単位として扱う Transformer Grammar (TG) のアテンションが、人間の記憶想起のモデルとしての妥当であることを示した。工学で発展してきたアテンションと言語学で想定されてきた統語構造との統合により妥当な認知モデルが得られるというこの結果が、両分野の協働を促進することを期待する。

6) 全ての品詞のうち以下の条件を全て満たすものを抜粋した：(i) 頻度が 1000 以上、(ii) TG・Transformer いずれかの ΔRMSE が有意、(iii) TG と Transformer の ΔRMSE の差分が有意。有意差検定には共に Wilcoxon の符号順位検定を用い、有意水準 ($p < 0.05$) にはそれぞれ Bonferroni 補正を施した。

## 謝辞

本研究は JSPS 科研費 JP24H00087, JP24KJ0800, JST さきがけ JPMJPR21C2, JST SPRING JPMJSP2108 の助成を受けたものです。


## 参考文献


[1] Ethan G. Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger P. Levy. On the Predictive Power of Neural Language Models for Human Real-TimeComprehension Behavior. In **Proceedings of the Annual Meeting of the Cognitive Science Society**, Vol. 42, 2020.

[2] John Hale. A Probabilistic Earley Parser as a Psycholinguistic Model. In **Second Meeting of the North American Chapter of the Association for Computational Linguistics**, 2001.

[3] Roger Levy. Expectation-based syntactic comprehension. **Cognition**, Vol. 106, No. 3, pp. 1126–1177, March 2008.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.

[5] Soo Hyun Ryu and Richard Lewis. Accounting for Agreement Phenomena in Sentence Comprehension with Transformer Language Models: Effects of Similarity-based Interference on Surprisal and Attention. In **Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics**, pp. 61–71, Online, June 2021. Association for Computational Linguistics.

[6] Byung-Doh Oh and William Schuler. Entropy- and Distance-Based Predictors From GPT-2 Attention Patterns Predict Reading Times Over and Above GPT-2 Surprisal. In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 9324–9334, Abu Dhabi, United Arab Emirates, February 2022. Association for Computational Linguistics.

[7] Julie A Van Dyke and Richard L Lewis. Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. **Journal of Memory and Language**, Vol. 49, No. 3, pp. 285–316, October 2003.

[8] David Marr. **Vision: A Computational Investigation into the Human Representation and Processing of Visual Information**. W. H. Freeman and Company, San Francisco, 1982.

[9] John T. Hale. **Automaton Theories of Human Sentence Comprehension**. CSLI Studies in Computational Linguistics. CSLI Publications, Stanford, CA, 2014.

[10] John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan Brennan. Finding syntax in human encephalography with beam search. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2727–2736, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[11] Michael Wolfman, Donald Dunagan, Jonathan Brennan, and John Hale. Hierarchical syntactic structure in human-like language models. In **Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics**, pp. 72–80, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[12] Shinnosuke Isono. Category Locality Theory: A unified account of locality effects in sentence comprehension. **Cognition**, Vol. 247, p. 105766, June 2024.

[13] Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is Not Only a Weight: Analyzing Transformers with Vector Norms. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 7057–7075, Online, January 2020. Association for Computational Linguistics.

[14] Laurent Sartran, Samuel Barrett, Adhiguna Kuncoro, Miloš Stanojević, Phil Blunsom, and Chris Dyer. Transformer Grammars: Augmenting Transformer Language Models with Syntactic Inductive Biases at Scale. **Transactions of the Association for Computational Linguistics**, Vol. 10, pp. 1423–1439, 2022.

[15] Daniel Jurafsky. A Probabilistic Model of Lexical and Syntactic Access and Disambiguation. **Cognitive Science**, Vol. 20, No. 2, pp. 137–194, 1996.

[16] Mitchell Stern, Daniel Fried, and Dan Klein. Effective Inference for Generative Neural Parsing. In **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing**, pp. 1695–1700, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[17] Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. Recurrent Neural Network Grammars. In **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 199–209, San Diego, California, June 2016. Association for Computational Linguistics.

[18] Hiroshi Noji and Yohei Oseki. Effective Batching for Recurrent Neural Network Grammars. In **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021**, pp. 4340–4352, Online, August 2021. Association for Computational Linguistics.

[19] Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. A Systematic Assessment of Syntactic Generalization in Neural Language Models. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 1725–1744, Online, July 2020. Association for Computational Linguistics.

[20] Nikita Kitaev and Dan Klein. Constituency Parsing with a Self-Attentive Encoder. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2676–2686, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[21] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[22] Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven Piantadosi, and Evelina Fedorenko. The Natural Stories Corpus. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[23] Richard L. Lewis and Shravan Vasishth. An activation-based model of sentence processing as skilled memory retrieval. **Cognitive Science**, Vol. 29, No. 3, pp. 375–419, May 2005.

[24] Harm Brouwer, Hartmut Fitz, and John Hoeks. Getting real about Semantic Illusions: Rethinking the functional role of the P600 in language comprehension. **Brain Research**, Vol. 1446, pp. 127–143, March 2012.

[25] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 2978–2988, Florence, Italy, July 2019. Association for Computational Linguistics.

**表 2**　TG・Transformer のハイパーパラメタ

| | |
|---|---|
| ベースモデル | Transformer-XL [25] |
| 語彙数 | 32,768 |
| 埋め込み次元数 | 1,024 |
| フィードフォワード次元数 | 4,096 |
| 層数 | 16 |
| ヘッド数 | 8 |
| セグメント長 | 256 |
| メモリ長 | 256 |
| 最適化器 | Adam ($\beta_1 = 0.9, \beta_2 = 0.999$) |
| バッチサイズ | 16 |
| 学習ステップ数 | 400,000 |
| 学習率スケジューラ | 線形ウォームアップ& コサイン減衰 |
| ウォームアップステップ数 | 32,000 |
| 開始学習率 | $2.5 \times 10^{-8}$ |
| 最大学習率 | $3.75 \times 10^{-5}$ |
| 最終学習率 | $7.5 \times 10^{-8}$ |
| ドロップアウト率 | 0.1 |

**表 3**　回帰モデルの説明変数

| | |
|---|---|
| zone | 物語中の単語位置 |
| position | 文中の単語位置 |
| wordlen | 単語中の文字数 |
| unigram | 単語ユニグラム頻度 |
| bigram | 単語バイグラム頻度 |
| trigram | 単語トライグラム頻度 |
| transformer_surp | Transformer のサプライザル |
| tg_surp | TG のサプライザル |
| clt | 統語理論に基づく 記憶減衰のメトリクス |
| transformer_nae | Transformer の NAE |
| tg_nae | TG の NAE |

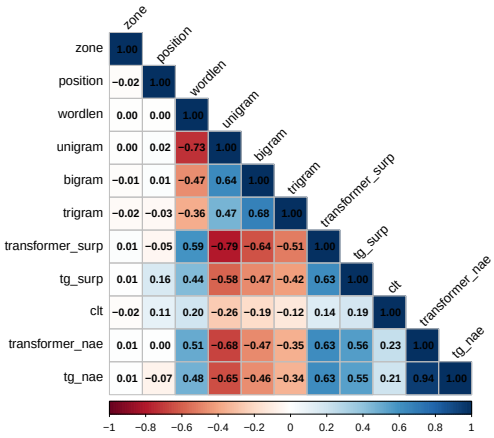


**図 4**　説明変数間の相関

# A　ハイパーパラメタ

TG・Transformer のハイパーパラメタを表 2 に示した。モデル関連のパラメタは全て [14, 11] にならい、学習関連のパラメタは実験に用いた計算資源 (NVIDIA RTX A5000, 24GB) に載るバッチサイズに合わせて調節した。本研究の全ての実験には合計約 100GPU 時間を要した。

# B　回帰モデルの説明変数

回帰モデルの説明変数の詳細を表 3 に、説明変数間の相関を図 4 に示した。

**表 4**　品詞タグおよび記号タグの一覧

| | | | |
|---|---|---|---|
| CC | 等位接続詞 | PRP$ | 所有代名詞 |
| CD | 基数詞 | RB | 副詞 |
| DT | 限定詞 | RBR | 副詞比較級 |
| EX | 存在の *there* | RBS | 副詞最上級 |
| FW | 外来語 | RP | 助詞 |
| IN | 前置詞/従属接続詞 | TO | *to* |
| JJ | 形容詞 | UH | 間投詞 |
| JJR | 形容詞比較級 | VB | 動詞原形 |
| JJS | 形容詞最上級 | VBD | 動詞過去形 |
| MD | 助動詞 | VBG | 動名詞/現在分詞 |
| NN | 名詞（単数/物質） | VBN | 過去分詞 |
| NNS | 名詞（複数） | VBP | 動詞現在形（非 3 単） |
| NNP | 固有名詞（単数） | VBZ | 動詞現在形（3 単） |
| NNPS | 固有名詞（複数） | WDT | 疑問限定詞 |
| PDT | 前置限定詞 | WP | 疑問代名詞 |
| POS | 所有格標識 | WP$ | 疑問所有代名詞 |
| PRP | 人称代名詞 | WRB | 疑問副詞 |
| -LRB- | 左丸括弧 | , | カンマ |
| -RRB- | 右丸括弧 | . | ピリオド |
| `` | 開き引用符 | : | コロン |
| '' | 閉じ引用符 | | |

**表 5**　本研究で用いたデータ・ツールのライセンス

| データ・ツール | ライセンス |
|---|---|
| BLLIP[1] | BLLIP 1987-89 WSJ Corpus Release 1 |
| Natural Stories コーパス [2] | CC BY-NC-SA 4.0 |
| Transformer Grammar[3] | Apache 2.0 |

[1] https://catalog.ldc.upenn.edu/LDC2000T43
[2] https://github.com/languageMIT/naturalstories
[3] https://github.com/google-deepmind/transformer_grammars

**表 6**　TG・Transformer のサプライザルの読み時間の予測への貢献度 (ΔLogLik)。***は $p < 0.001$ を表す。読み時間の平均値は 335ms である。

| 説明変数 | ΔLogLik (↑) | 効果量 |
|---|---|---|
| tg_surp | 200 | 1.65 ms*** |
| transformer_surp | **295** | 3.75 ms*** |

# C　品詞タグ

Natural Stories コーパスに付与されている品詞タグおよび記号タグの一覧を表 4 に示した。読み時間は原則スペース区切りで付与されているため、記号タグを含む場合（例：NNP.）は、それを除いたもの（例：NNP）を品詞タグとして用いて分析を行った。

# D　サプライザルの説明力

ベースライン回帰モデル（式 3）から両者の言語モデルのサプライザルを除き、代わりに両者の言語モデルの NAE を加えた回帰モデルをベースライン回帰モデルとして用いて、各言語モデルのサプライザルの読み時間予測への貢献度を調査した（表 6）。どちらの言語モデルのサプライザルも、読み時間に対して有意に説明力を持つが、説明力の絶対値は Transformer が TG を上回る結果となった。また、説明力の包含関係 (§4) についても調査したが、両者の言語モデルのサプライザルを含む言語モデルは、どちらか片方の言語モデルのサプライザルのみを含む言語モデルよりも有意に説明力が高かった（共に $p < 0.001$）。この結果は、(i) 記憶想起のモデルとは異なり、トークン列のみに基づく次単語予測の方が人間の予測処理のより支配的な要素を捉えられるが、(ii) 記憶想起のモデルと同じく、トークン列のみ/統語構造およびトークン列に基づく次単語予測が、共に人間の予測処理の異なる側面を捉えているモデルとして両立しうる可能性がある、ことを示している。

# E　ライセンス

本研究で用いたデータ・ツールのライセンスを表 5 に示した。