If Attention Serves as a Cognitive Model of Human Memory Retrieval, What is the Plausible Memory Representation?

Ryo Yoshida[♠] **Shinnosuke Isono**[♠♡] **Kohei Kajikawa**[♠♡] **Taiga Someya**[♠]
Yushi Sugimoto[◇] **Yohei Oseki**[♠]
[♠]The University of Tokyo [♡]NINJAL [◇]The University of Osaka
{yoshiryo0617,taiga98-0809,oseki}@g.ecc.u-tokyo.ac.jp
{s-isono,kohei-kajikawa}@ninjal.ac.jp sugimoto.yushi.hmt@osaka-u.ac.jp

Abstract

Recent work in computational psycholinguistics has revealed intriguing parallels between attention mechanisms and human memory retrieval, focusing primarily on vanilla Transformers that operate on token-level representations. However, computational psycholinguistic research has also established that syntactic structures provide compelling explanations for human sentence processing that token-level factors cannot fully account for. In this paper, we investigate whether the attention mechanism of Transformer Grammar (TG), which uniquely operates on syntactic structures as representational units, can serve as a cognitive model of human memory retrieval, using Normalized Attention Entropy (NAE) as a linking hypothesis between models and humans. Our experiments demonstrate that TG's attention achieves superior predictive power for self-paced reading times compared to vanilla Transformer's, with further analyses revealing independent contributions from both models. These findings suggest that human sentence processing involves dual memory representations—one based on syntactic structures and another on token sequences—with attention serving as the general memory retrieval algorithm, while highlighting the importance of incorporating syntactic structures as representational units.

1 Introduction

Whether language models (LMs) developed in natural language processing (NLP) are plausible as cognitive models of human sentence processing is a central question in computational psycholinguistics. Over the past two decades, this question has been primarily addressed from the perspective of *expectation-based theories*—one of the two major classes of human sentence processing theory—examining whether LMs' next-token prediction can serve as a model of human predictive processing (Hale, 2001; Levy, 2008; Wilcox et al., 2020; Merkx and Frank, 2021; *inter alia*).

The recent success of Transformers (Vaswani et al., 2017) in NLP has unexpectedly opened a new avenue of investigation from the perspective of *memory-based theories*, the other major class of sentence processing theory. Researchers have proposed that the attention mechanism, despite its engineering origins, can implement a human memory retrieval theory known as cue-based retrieval (Van Dyke and Lewis, 2003). Recent studies have revealed intriguing parallels between the weighted reference patterns exhibited by the attention mechanism and the elements that humans may retrieve during online sentence comprehension (Ryu and Lewis, 2021; Oh and Schuler, 2022; Timkey and Linzen, 2023).

Computational psycholinguistics has also established that human sentence processing cannot be fully explained by token-level factors; rather, *syntactic structures* have provided compelling explanations for it. For instance, next-token prediction from LMs that explicitly incorporate syntactic structure building demonstrates superior performance in accounting for human brain activity compared to vanilla RNNs and Transformers (Hale et al., 2018; Wolfman et al., 2024); the number of syntactic nodes hypothesized to be constructed per word correlates significantly with both reading times and neural activity patterns (Kajikawa et al., 2024; Brennan et al., 2012).

Given these findings, if attention can serve as a general algorithm for memory retrieval in human sentence processing, human memory retrieval should be captured by the attention mechanism operating on syntactic structures as well as that operating on token sequences. In this paper, we investigate whether the attention mechanism of Transformer Grammar (TG; Sartran et al., 2022), which uniquely operates on syntactic structures as representational units, can serve as a cognitive model of human memory retrieval, using Normalized Attention Entropy (NAE; Oh and Schuler, 2022) as the

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9795–9812

linking hypothesis between models and humans. Our experiments demonstrate that TG's attention achieves superior predictive power for self-paced reading times compared to vanilla Transformer's, with further analyses revealing independent contributions from both models. These findings suggest that human sentence processing involves dual memory representations—one based on syntactic structures and another on token sequences—with attention serving as the general memory retrieval algorithm, while highlighting the importance of incorporating syntactic structures as representational units.[1]

2 Background

2.1 Cue-based retrieval

Many psycholinguistic studies assume that human sentence processing involves memory retrieval, where based on the various cues provided by the current input word (e.g., verbs), elements (e.g., their arguments) are retrieved from working memory. In Example 1, taken from Van Dyke (2002), when the verb *was complaining* is input, its subject *the resident* must be retrieved from working memory.

(1) a. The worker was surprised that the **resident**$_{\text{[subj,anim]}}$ [who was living near the dangerous warehouse] *was complaining* about the investigation.

b. The worker was surprised that the **resident**$_{\text{[subj,anim]}}$ [who said that the warehouse$_{\text{[subj]}}$ was dangerous] *was complaining* about the investigation.

According to the cue-based retrieval theory (Van Dyke and Lewis, 2003), such retrieval becomes more difficult when similar elements exist in the sentence because the cues are overloaded; for example, only in Example 1b, *warehouse* may interfere with *resident* since they both have the feature [subj] as a retrieval cue. Van Dyke (2002) showed that humans read *was complaining* more slowly in Example 1b than in Example 1a, providing empirical support for the cue-based retrieval theory. Such interference effects have been observed across various syntactic and semantic features (Van Dyke and Lewis, 2003; Van Dyke and McElree, 2011; Nicenboim et al., 2018).

[1] Code for reproducing our results is available at `https://github.com/osekilab/TG-NAE`.

2.2 Normalized Attention Entropy (NAE)

In recent computational psycholinguistics, attempts have been made to interpret the attention mechanism—a weighted reference of preceding tokens based on Query and Key vectors—as a computational implementation of cue-based retrieval. Notably, Ryu and Lewis (2021) proposed Attention Entropy (AE) as a linking hypothesis, where the diffuseness of attention weights is assumed to quantify the degree of retrieval interference. While AE was initially proposed for modeling interference effects in specific constructions, Oh and Schuler (2022) extended it to naturally occurring text by introducing two normalizations: (i) division by the maximum entropy achievable given the number of preceding tokens, and (ii) sum-to-1 renormalization of attention weights over preceding tokens (**Normalized AE, NAE**).[2]

$$\mathrm{NAE}_{l,h,i} = -\frac{1}{\log_2 |T|} \sum_{j \in T} \tilde{a}_{l,h,i,j} \log_2 \tilde{a}_{l,h,i,j} \tag{1}$$

where T is the set of preceding token positions,[3] $\tilde{a}_{l,h,i,j} = \frac{a_{l,h,i,j}}{\sum_{k \in T} a_{l,h,i,k}}$ is the renormalized attention weight, and $a_{l,h,i,j}$ represents the attention weight from the i-th token (Query) to the j-th preceding token (Key) in the h-th head of layer l.[4] In this paper, we employ this NAE as a linking hypothesis between attention mechanisms and human memory retrieval.[5]

2.3 Transformer Grammar (TG)

Transformer Grammar (**TG**; Sartran et al., 2022) is a type of syntactic LM, a generative model that jointly generates token sequences $\boldsymbol{x}$ and their corresponding syntactic structures $\boldsymbol{y}$. TG formulates the generation of $(\boldsymbol{x}, \boldsymbol{y})$ as modeling a sequence of actions, $\boldsymbol{a}$ (e.g., `(S (NP The blue bird NP) (VP sings VP) S)`), constructing both token sequences

[2] Oh and Schuler (2022) showed that regression models for predicting reading times fail to converge with vanilla AE.

[3] For vanilla Transformers, $T = \{1, 2, \ldots, i-1\}$.

[4] Oh and Schuler (2022) explored NAE calculation using various attention weight formulations, but in this paper, we adopt the norm-based attention weight formulation (Kobayashi et al., 2020), which achieved the highest predictive power on the self-paced reading time corpus.

[5] While Oh and Schuler (2022) also proposed other metrics based on distances between attention weights at consecutive time steps, we exclusively adopt NAE because (i) in TG, the number of preceding elements varies with time, making distance definition non-trivial, and (ii) Oh and Schuler (2022) demonstrated that NAE's predictive power subsumes that of distance-based metrics in the self-paced reading time corpus.

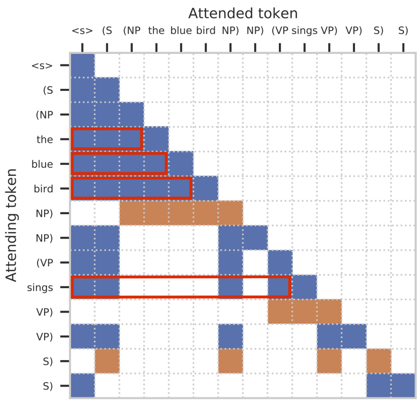

Figure 1: TG's attention mask with COMPOSE/STACK attention mechanisms, adapted from Sartran et al. (2022). COMPOSE generates a vector representation of the closed phrase, while subsequent STACK operations reference this vector as the phrase representation for next-action prediction. Red boxes indicate the attention weights used to calculate NAE for each word.

and their syntactic structures in a top-down, left-to-right manner. The action sequence $\boldsymbol{a}$ comprises three types of operations:

- `(X`: Generate a non-terminal symbol `(X`, where `X` represents a phrasal tag such as `NP`;
- `w`: Generate a terminal symbol `w`, where `w` represents a token such as `bird`;
- `X)`: Generate `X)` to close the most recent opened non-terminal symbol, where `X` matches the phrasal tag of the targeted non-terminal symbol.

The probability of action sequence $\boldsymbol{a} = (a_1, a_2, \dots, a_n)$ is decomposed using the chain rule. Formally, TG is defined as:

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{a}) = \prod_{t=1}^{n} p(a_t | a_{<t}). \qquad (2)$$

TG's key innovation lies in its handling of closed phrases: immediately after generating `X)`, it computes a vector representation of the closed phrase, which subsequent next-action predictions use as the representation for that phrase. Technically, this operation is realized via two components: `X)` action duplication and a specialized attention mask. The duplication process transforms $\boldsymbol{a}$ into $\boldsymbol{a}'$ by duplicating all `X)` actions (e.g., `(S (NP The blue bird NP) NP) (VP sings VP) VP) S) S)`), while preserving the modeling space $p(\boldsymbol{a})$ by preventing predictions for duplicated positions. The attention mask implements two distinct attention mechanisms: COMPOSE and STACK (Figure 1). COMPOSE operates exclusively at the first occurrence of each `X)` to generate the phrasal representation by attending only to vectors between the corresponding `(X` and `X)` (without making predictions). STACK operates at all other positions to compute representations for next-action prediction, with attention restricted to positions on the *stack* (comprising unclosed non-terminals, not-composed terminals, and closed phrases).

Previous research has demonstrated that TG's probability estimates align more closely with human offline grammaticality judgments (Sartran et al., 2022) and online brain activity (Wolfman et al., 2024) than vanilla Transformers. This paper investigates whether the attention mechanism of TG, which uniquely operates on syntactic structures as representational units, can serve as a cognitive model of human memory retrieval.

3 Methods

3.1 NAE calculation with TG

The calculation of NAE with TG requires assumptions regarding two key perspectives:

1. What syntactic structures should be assumed for a given token sequence?
2. How should the cognitive load of attention from non-lexical symbols (i.e., `(X` and `X)`) be attributed to lexical tokens?

In response to these considerations, we make the following assumptions:

1-A. We assume only the globally correct syntactic structure (i.e., "perfect oracle"; Brennan, 2016).

2-A. We consider only attention from lexical tokens, excluding attention from non-lexical symbols.

The adoption of 1-A. is motivated by two factors. First, the self-paced reading time corpus we utilized here provides gold-standard syntactic structures for each sentence, and previous studies have developed predictors based on these annotations (Shain et al., 2020; Isono, 2024). Using the same structural assumptions enables fair comparison with these established predictors, considering the possibility of parsing errors. Second, TG's current implementation lacks beam search procedure (Stern et al., 2017; Crabbé et al., 2019), an inference technique commonly used in cognitive modeling to handle local ambiguities through parallel parsing (Hale et al., 2018; Sugimoto et al., 2024).[6]

Regarding 2-A., given the multiple possible approaches to attributing processing load from non-lexical symbols to lexical tokens, we adopt the most straightforward and theoretically neutral approach. Figure 1 denotes the attention weights used to calculate NAE for each word, with red boxes.

3.2 Settings

Language models We used 16-layer, 8-head TG and Transformer (252M parameters).[7] All hyperparameters followed the default settings described in Sartran et al. (2022) (see Appendix A). Following Oh and Schuler (2022), we computed NAE separately for each attention head at the top layer and then summed the values across heads.

Training data We used BLLIP-LG, a dataset containing 42M tokens (1.8M sentences) from the Brown Laboratory for Linguistic Information Processing (BLLIP) 1987–89 WSJ Corpus Release 1 (Charniak et al., 2000).[8] The corpus was re-parsed using a state-of-the-art constituency parser (Kitaev and Klein, 2018) and split into train-val-test sets by Hu et al. (2020). BLLIP-LG has been widely used for training syntactic LMs, including TG. Following Sartran et al. (2022), we trained a 32K SentencePiece tokenizer (Kudo and Richardson, 2018) on the training set and segmented each sentence into subword units.

Both TG and Transformer were trained at the sentence level: TG maximized the joint probability $p(\boldsymbol{x}, \boldsymbol{y})$ on action sequences, while Transformer maximized the probability $p(\boldsymbol{x})$ on terminal subword sequences. For training hyperparameters, we largely followed the default settings in Sartran et al. (2022) but adjusted the batch size to fit within the memory constraints of our hardware (NVIDIA A100, 40GB). Accordingly, we tuned other hyperparameters (e.g., learning rate) to maintain training stability.[9] We trained three models with different random initialization seeds and selected the checkpoint with the lowest validation loss for each run.

Reading time data We used the Natural Stories corpus (Futrell et al., 2018),[10] "a series of English narrative texts designed to contain many low-frequency and psycholinguistically interesting syntactic constructions while still sounding fluent and coherent." We selected this corpus for these "interesting" syntactic constructions, which provide an ideal testbed for investigating memory retrieval effects that might be less pronounced in simpler, more naturalistic texts. The corpus has also been used in several studies investigating memory-related processing mechanisms (Shain et al., 2016; Dotlačil, 2021; Isono, 2024).

The Natural Stories corpus consists of 10 stories (485 sentences, 10,245 words) with self-paced reading times collected from 181 anonymized native English speakers. Following Futrell et al.'s preprocessing, data points were removed if (i) a participant scored less than 5/6 on comprehension questions for a story or (ii) individual reading times were less than 100 ms or greater than 3,000 ms. Following Oh and Schuler (2022), we also excluded sentence-initial and sentence-final data points. We further removed sentence-second data points, as they lack the log trigram frequency of the previous token required for our baseline regression model. After preprocessing, 725,875 data points from 180 participants remained for statistical analysis, out of the original 848,875 data points.

Statistical analysis We evaluated each LM's NAE contribution to reading time prediction by measuring improvements in regression model fit when adding NAE as predictors. For each model (TG/Transformer), we included both the current word's NAE (`tg_nae`/`tf_nae`) and the previous word's NAE (`tg_nae_so`/`tf_nae_so`) to account for spillover effects (Mitchell, 1984).[11][12] Model improvement was quantified as the increase in log-likelihood (ΔLogLik). This evaluation was conducted for each random seed, and we report the mean ΔLogLik with standard deviation.

[6] As a proof of concept, we also conducted experiments using multiple syntactic structures generated by word-synchronous beam search with Recurrent Neural Network Grammar (Dyer et al., 2016; Kuncoro et al., 2017; Noji and Oseki, 2021), obtaining similar results (Appendix D).

[7] https://github.com/google-deepmind/transformer_grammars

[8] https://catalog.ldc.upenn.edu/LDC2000T43

[9] For the detailed hyperparameters, see Appendix A.

[10] https://github.com/languageMIT/naturalstories. We used the corrected version that addresses the data misalignment issue identified in May 2025.

[11] _so indicates spillover.

[12] Following Oh and Schuler (2022), we summed the subword NAE values for each word.

Model	ΔLogLik (↑)	Predictor	Effect size [ms]	p-value range	Significant seeds
TG	**76.6** (±8.1)	`tg_nae`	1.42 (± 0.2)	<**0.001**	3/3
		`tg_nae_so`	2.26 (± 0.1)	<**0.001**	3/3
Transformer	42.8 (±9.5)	`tf_nae`	1.32 (± 0.2)	<**0.001**	3/3
		`tf_nae_so`	1.46 (± 0.2)	<**0.001**	3/3

Table 1: TG's and Transformer's NAE contribution to reading time prediction (ΔLogLik). The effect size per standard deviation is shown for each model-derived predictor, along with the p-value range across random seeds and the number of seeds showing significant contributions. Standard deviations across seeds for ΔLogLik and effect sizes are shown in parentheses. The mean reading time in the analysis is 334 ms.

Following previous studies such as Dotlačil (2021), Shain et al. (2016), and Isono (2024), the baseline regression model controlled for non-structural, basic aspects of text known to affect reading times:

- `zone` and `position` (integer): word position in the story and sentence;
- `wordlen` (integer): number of characters in the word;
- `unigram`, `bigram`, and `trigram` (continuous): log-transformed n-gram frequencies.

We additionally included the following predictors:

- `tg_surp` and `tf_surp` (continuous): surprisal from TG and Transformer;
- `stack_count` (integer): number of elements in the *stack* (comprising unclosed non-terminals, not-composed terminals, and closed phrases).

Following Oh and Schuler (2022), we included surprisal to test NAE's significance in the presence of surprisal predictors from the same LMs.[13] Stack count was included to isolate the cost of holding elements (Joshi, 1990; Abney and Johnson, 1991; Resnik, 1992) from their interference effects, which TG's NAE was designed to capture. For the correlations between the predictors, see Appendix B.

All predictors were z-transformed, and we also included the previous word's values as predictors to model spillover, except for the positional information. The baseline regression model was a linear mixed-effects model (Baayen et al., 2008) with these fixed effects and by-subject and by-story random intercepts:

$$\begin{aligned}\log(\texttt{RT}) \sim\ & \texttt{zone} + \texttt{position} + \texttt{wordlen} + \\ & \texttt{unigram} + \texttt{bigram} + \texttt{trigram} + \\ & \texttt{tg\_surp} + \texttt{tf\_surp} + \\ & \texttt{stack\_count} + \texttt{wordlen\_so} + \\ & \texttt{unigram\_so} + \texttt{bigram\_so} + \\ & \texttt{trigram\_so} + \texttt{tg\_surp\_so} + \\ & \texttt{tf\_surp\_so} + \texttt{stack\_count\_so} + \\ & (1 \mid \texttt{participant}) + (1 \mid \texttt{story}) \quad (3)\end{aligned}$$

To assess each LM's independent contribution to reading time prediction, we also conducted likelihood ratio tests (Wurm and Fisicaro, 2014) by extending Equation 3 in two ways: adding both LMs' NAE versus adding only one LM's NAE. Note that a larger ΔLogLik from one LM does not necessarily indicate that it contributes above and beyond the other LM, nor does a smaller ΔLogLik indicate no unique contribution. Following Aurnhammer and Frank (2019), we used NAE and surprisal values averaged across random seeds for these nested model comparisons.

4 Results

4.1 Does TG's NAE have predictive power for reading times?

Table 1 presents the contributions of TG's and Transformer's NAE to reading time prediction. First, Transformer's NAE exhibited significant predictive power for reading times, independent of baseline predictors such as surprisal. The effect size was in the expected positive direction (higher NAE values corresponding to longer reading times), showing both immediate and spillover effects. This corroborates the arguments of Ryu and Lewis (2021) and Oh and Schuler (2022) that the attention

[13]For an experiment on the predictive power of surprisal itself, see Appendix E.

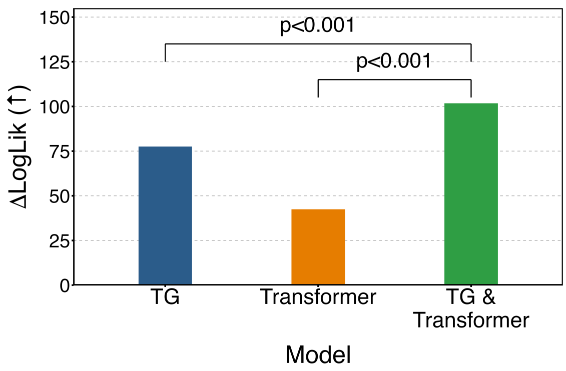

Figure 2: Likelihood ratio test results examining the independence of NAE's predictive power

mechanism—the weighted reference of preceding tokens—functions as a cognitive model of human memory retrieval, despite its engineering-oriented origins.

Second, TG's NAE exhibited robust predictive power, demonstrating significant positive effects in both immediate and spillover conditions. This finding not only provides additional evidence for incremental construction of syntactic structures in human sentence processing (e.g., Fossum and Levy, 2012), but also suggests that TG's attention mechanism effectively models memory retrieval from these constructed syntactic representations.

Finally, TG's NAE made a substantially stronger contribution to reading time prediction (ΔLogLik=76.6) compared to Transformer's NAE (ΔLogLik=42.8). This finding suggests that retrieval from syntactic memory representations plays a more dominant role in human sentence processing than retrieval from lexical memory representations. This underscores the importance of incorporating syntactic structures as a unit of memory representation, which we implemented through the integration of TG and NAE here.

4.2 Do TG's and Transformer's NAE have independent contributions?

Figure 2 presents the results of likelihood ratio tests examining the independence of TG's and Transformer's NAE contributions. The regression model incorporating NAE from both LMs ('TG & Transformer') demonstrated significantly higher predictive power than the models containing NAE from either LM alone ('TG' or 'Transformer'). This reveals that TG's NAE certainly captures variance in reading times that Transformer's NAE cannot explain, while Transformer's NAE, despite its lower overall predictive power, accounts for unique variance not captured by TG's NAE. This finding aligns with psycholinguistic literature, where cognitive models of memory retrieval encompass both syntax-based approaches (e.g., verb-argument relationships; Lewis and Vasishth, 2005) and semantic-based approaches (e.g., bag-of-words-like similarity; Brouwer et al., 2012), suggesting that the attention mechanisms of TG and Transformer serve as complementary cognitive models, each capturing distinct aspects of human memory retrieval.

4.3 What aspects of memory retrieval do TG's and Transformer's NAE capture?

To investigate the aspects of human memory retrieval captured by TG's and Transformer's NAE, we analyzed differences in prediction improvement across part-of-speech (POS) tags annotated in the Natural Stories corpus.[14] Our analysis followed three steps: (i) selecting POS tags with more than 1,000 occurrences, (ii) for each POS tag, testing the significance of improvement from the baseline regression model (measured in Δ Root Mean Squared Error, ΔRMSE) when adding NAE of the current and previous word as fixed effects,[15] and (iii) examining the significance of *differences* in ΔRMSE between TG and Transformer for POS tags where either model showed significant improvement. We assessed significance using Wilcoxon signed-rank tests with Bonferroni correction ($p < 0.05$).

Figure 3 presents the differences in prediction improvement across POS tags. Consistent with the larger ΔLogLik value, TG's NAE demonstrated advantages over Transformer's NAE across a broader range of POS tags. Notably, TG's NAE exhibited superior improvement across verbs (`VB`, `VBG`, `VBN`, and `VBP`), while Transformer's NAE excelled for nouns (`NN` and `NNP`). This pattern aligns with our earlier argument regarding the complementary nature of these models (Section 4.2), indicating that different types of retrieval operations—verb-triggered retrieval, which often relies on syntactic features (e.g., argument structure), and noun-triggered retrieval, which often relies on semantic features (e.g., referential associations)—are better captured by distinct attention mechanisms: attention with syntactic and token memory representa-

[14] For a complete list of POS tags in the Natural Stories corpus, see Appendix C.

[15] We used the same regression models as in Section 4.2, where surprisal and NAE values were averaged across seeds.

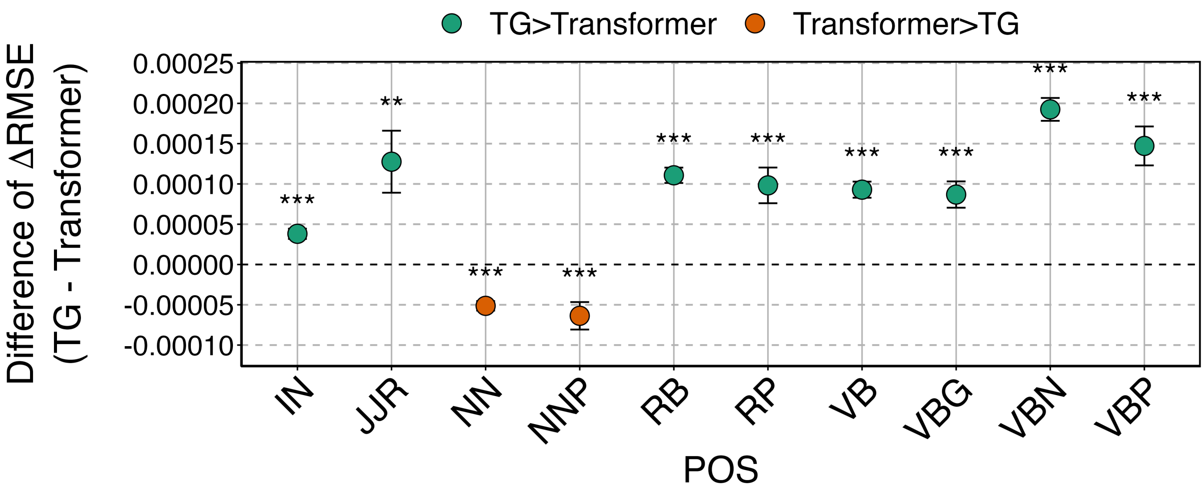

Figure 3: Differences in reading time prediction improvement (ΔRMSE) between TG and Transformer across POS tags (TG - Transformer). The y-axis shows the mean differences per word, with the error bars representing standard errors. Only POS tags showing significant improvement in either model and significant differences between models are displayed. Statistical significance after Bonferroni correction: ** $p < 0.01$, *** $p < 0.001$.

Model	ΔLogLik	Predictor	p-value
TG	**46.1** (± 9.1)	*_nae *_nae_so	** (2/3) *** (3/3)
$\text{TG}_{-\text{comp}}$	18.1 (± 9.3)	*_nae *_nae_so	** (1/3) *** (3/3)

Table 2: TG's and $\text{TG}_{-\text{comp}}$'s contribution to reading time prediction. The rightmost column shows the p-value range across random seeds that achieved significance (*** $p < 0.001$ and ** $p < 0.01$), along with the number of seeds showing significant contributions. Due to the potential multicollinearity between the Transformer's NAE and TG/$\text{TG}_{-\text{comp}}$'s NAE, the column of the effect size is omitted.

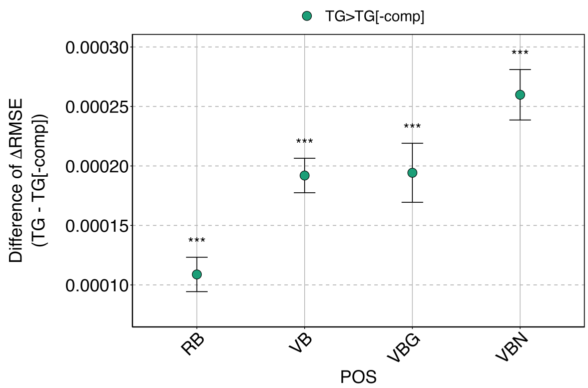

Figure 4: Differences in ΔRMSE between TG and $\text{TG}_{-\text{comp}}$ across POS tags (TG - $\text{TG}_{-\text{comp}}$). Statistical significance after Bonferroni correction: *** $p < 0.001$.

tions, respectively.

5 Follow-up analysis

5.1 Do TG's advantages stem from the COMPOSE attention?

As described in Section 2.3, TG's key feature is the COMPOSE attention, which explicitly generates single vector representations for closed phrases. Here, we investigate whether TG's predictive power derives from merely considering syntactic structures or from explicitly treating closed phrases as single representations (see Hale et al., 2018; Brennan et al., 2020). To address this question, we developed $\text{TG}_{-\text{comp}}$, a TG variant that processes each action in the action sequence $\boldsymbol{a}$ as an individual token without the COMPOSE attention (i.e., Choe and Charniak's *Parsing as Language Modeling* approach). We trained $\text{TG}_{-\text{comp}}$ with identical hyperparameters as TG. The baseline regression model (Equation 3) was augmented with (i) $\text{TG}_{-\text{comp}}$'s surprisal and (ii) Transformer's NAE to (i) ensure a fair comparison between TG and $\text{TG}_{-\text{comp}}$ and (ii) distinguish between the effects of direct terminal token access and syntactic structure consideration in $\text{TG}_{-\text{comp}}$.[16]

Table 2 presents the ΔLogLik values obtained when incorporating either TG's or $\text{TG}_{-\text{comp}}$'s NAE as fixed effects into the baseline regression model. Note that due to the potential multicollinearity

[16]For direct comparison between TG and $\text{TG}_{-\text{comp}}$ under the baseline regression model without Transformer's NAE, see Appendix F.

between Transformer's NAE and TG/TG$_{-\text{comp}}$'s NAE, we focus on the ΔLogLik values and significance of the contribution rather than individual effect sizes.

Our analysis reveals two key findings. First, TG$_{-\text{comp}}$'s NAE demonstrates significant predictive power for reading times, even in the presence of Transformer's NAE, implying that consideration of syntactic structures alone captures certain memory retrievals that the token-level attention mechanism cannot capture. Second, TG's NAE outperforms TG$_{-\text{comp}}$'s, suggesting that the attention mechanism that treats closed phrases as single representations more effectively captures variance in syntax-based memory retrieval. The likelihood ratio tests further revealed that TG's NAE captured reading time patterns unexplainable by TG$_{-\text{comp}}$ ('TG & TG$_{-\text{comp}}$'>'TG$_{-\text{comp}}$', $p < 0.001$), while TG$_{-\text{comp}}$ did not explain unique variance beyond what TG already accounts for ('TG & TG$_{-\text{comp}}$'>'TG', $p = 0.478$).

We analyzed the ΔRMSE differences across POS tags to investigate which aspects of human memory retrieval are better captured by the COMPOSE attention. Figure 4 showed that TG's NAE was consistently superior to TG$_{-\text{comp}}$'s NAE across verbs (VB, VBG, and VBN), highlighting the critical role of the COMPOSE attention in capturing verb-triggered retrieval, a type of retrieval that was identified as distinctively better captured by TG compared to vanilla Transformers (Section 4.3).

5.2 Does TG's NAE capture interference effects?

Psycholinguistic research has considered two primary types of memory retrieval costs: *interference* effects, which NAE aims to capture, and *decay* effects—the cognitive load associated with accessing elements at greater linear distances (e.g., Gibson, 1998, 2000). Here, we examine whether TG's NAE genuinely captures interference effects by testing its independence from variables that model memory decay effects. For modeling decay effects, we employed Category Locality Theory (CLT; Isono, 2024),[17] which treats phrases in syntactic structure[18] as representational units of memory and quantifies decay effects using the distance (measured in content words) between an input and the phrases to be composed with it.

Model	Predictor	Effect size [ms]
TG & CLT	`tg_nae`	1.18^{***}
	`tg_nae_so`	2.38^{***}
	`clt`	0.06
	`clt_so`	1.30^{***}

Table 3: Effect sizes per standard deviation are shown for TG's NAE and CLT predictors. Significance levels: *** $p < 0.001$.

To assess independence, we tested whether TG's NAE and CLT maintain their contributions when simultaneously included in the baseline regression model (Equation 3), and examined their independence through likelihood ratio tests.[19] The results (Table 3) show that TG's NAE exhibited significant effects in both immediate and spillover conditions, and CLT demonstrated a significant spillover effect. The likelihood ratio tests confirmed that these effects were independent ('TG & CLT'>'CLT', $p < 0.001$; 'TG & CLT'>'TG', $p < 0.001$).

These results provide empirical evidence that NAE quantifies interference rather than decay in memory retrieval—extending beyond previous studies on NAE (Ryu and Lewis, 2021; Oh and Schuler, 2022). This finding is significant because, as far as we are aware, while psycholinguistics has developed various implementations of memory decay effects, it has lacked broad-coverage implementations of interference effects applicable to naturally occurring texts. Our results suggest that NAE represents a promising approach for quantifying interference effects in a broad-coverage manner.

6 Level of description

In cognitive modeling studies based on surprisal theory, explanations typically follow the form "if these LMs were models of human prediction, the difficulty of next-word disambiguation that humans solve would be approximated as follows." Such explanations typically operate at the most abstract of Marr's three levels of description—the computational level. Recently, Futrell et al. (2020) proposed

[17]Although Dependency Locality Theory (DLT; Gibson, 1998, 2000) is widely recognized as one of the most prominent models for capturing decay effects, we opted for CLT here, following Isono's finding that DLT-based predictors fail to achieve statistical significance in explaining reading times in the Natural Stories corpus.

[18]CLT assumes syntactic structure based on Combinatory Categorial Grammar (Steedman, 2000).

[19]As in other likelihood ratio tests, we used surprisal and NAE values averaged across random seeds.

lossy-context surprisal to integrate memory representation perspectives into surprisal theory. However, as they explicitly stated, this theory remains at the computational level, relaxing assumptions about memory representations in human predictive processing. In contrast, cognitive models of human memory, such as cue-based retrieval, generally provide explanations about *mechanisms* that deal with specific mental representations. These explanations typically move down one level to the algorithmic level of description.

While not explicitly stated by the authors, we argue that the work of Ryu and Lewis (2021) and Oh and Schuler (2022)—conceptualizing attention mechanisms as implementations of cue-based retrieval—represents movement toward the algorithmic level, approaching cue-based retrieval itself. Our research advances this direction by investigating a fundamental question at this level: the nature of memory representations (see Hale, 2014). Future work could fully operate at the algorithmic level by incorporating more realistic elements such as parallel parsing, left-corner parsing strategies, and memory decay mechanisms.

7 Conclusion

In this paper, we have demonstrated that attention can serve as the general algorithm for modeling human memory retrieval from two representational systems. Furthermore, we have shown that among the LMs examined in this paper (TG, $\text{TG}_{-\text{comp}}$, and Transformer), TG—whose attention mechanism uniquely operates on syntactic structures as representational units—best captures dominant factors in human sentence processing. Our results suggest that the integration of attention mechanisms (developed in NLP) with syntactic structures (theorized in linguistics) constitutes a broad-coverage candidate implementation for human memory retrieval. We hope these findings will foster greater collaboration between these two fields.

Limitations

Our NAE calculation comprised three steps: (i) computing NAE for each attention head in the top layer, (ii) adding the values across heads, and (iii) summing subword-level values into word level. While this procedure strictly adhered to Oh and Schuler (2022), alternative approaches to handling layers, attention heads (Ryu and Lewis, 2021), and subword tokens (Oh and Schuler, 2024; Giulianelli et al., 2024) warrant investigation.

While our study provides an in-depth investigation using the Natural Stories corpus—an English self-paced reading time corpus containing many interesting syntactic constructions—the breadth of our analysis has certain limitations. The generalizability of our findings to different languages (e.g., Japanese self-paced reading time corpus from Asahara, 2022) and other cognitive load (e.g., gaze duration from Kennedy et al., 2003 or EEG and fMRI from Bhattasali et al., 2020) remains to be investigated.

As discussed in Section 3.1, we employed "perfect oracles" as syntactic structures behind token sequences. While this assumption has been widely adopted in previous studies on human syntactic processing (Brennan, 2016; Shain et al., 2016; Stanojević et al., 2021; Isono, 2024), this idealization leaves the resolution of local ambiguities, which humans encounter during actual sentence processing, outside the scope of our study (for a conceptual case study, see Appendix D).

Although we adopted the default TG implementation of a top-down parsing strategy, psycholinguistic literature has suggested that a left-corner parsing strategy might be more plausible for human sentence processing from a perspective of memory capacity (cf. `stack_count`) (Abney and Johnson, 1991; Resnik, 1992). However, when contrasting memory representations based on syntactic structures versus token sequences, the attention mechanism of the top-down TG can serve as a reasonable approximation of retrieval from structure-based memory—this aligns with previous work contrasting predictions based on syntactic structures versus token sequences, which used top-down TG or RNNG to represent structure-based prediction (Wolfman et al., 2024; Hale et al., 2018; Brennan et al., 2020).

Finally, while this paper focused on investigating the attention mechanism through the lens of memory-based theory, exploring TG and Transformer as integrated implementations for expectation-based theory (via surprisal) and memory-based theory (via NAE) represents a promising future direction (Michaelov et al., 2021; Ryu and Lewis, 2022). Specifically, future work could investigate the attention mechanism as the underlying driver of surprisal's predictive power (Appendix E), analyzing the relationship between surprisal and NAE.

Ethical considerations

We employed AI-based tools (Claude, ChatGPT, GitHub Copilot, and Grammarly) for writing and coding assistance. These tools were used in compliance with the ACL Policy on the Use of AI Writing Assistance.

Acknowledgments

We sincerely thank Michael Wolfman and John Hale for providing detailed information on Wolfman et al. (2024). We also appreciate the insightful reviews provided by anonymous ARR 2025 February reviewers. This work was supported by JSPS KAKENHI Grant Number 24H00087, Grant-in-Aid for JSPS Fellows JP24KJ0800, JST PRESTO Grant Number JPMJPR21C2, and JST SPRING Grant Number JPMJSP2108.

References

Steven P. Abney and Mark Johnson. 1991. Memory requirements and local ambiguities of parsing strategies. *Journal of Psycholinguistic Research*, 20(3):233–250.

Masayuki Asahara. 2022. Reading Time and Vocabulary Rating in the Japanese Language: Large-Scale Japanese Reading Time Data Collection Using Crowdsourcing. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5178–5187, Marseille, France. European Language Resources Association.

Christoph Aurnhammer and Stefan L. Frank. 2019. Comparing Gated and Simple Recurrent Neural Network Architectures as Modelsof Human Sentence Processing. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 41(0).

R. H. Baayen, D. J. Davidson, and D. M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Shohini Bhattasali, Jonathan Brennan, Wen-Ming Luh, Berta Franzluebbers, and John Hale. 2020. The Alice Datasets: fMRI & EEG Observations of Natural Language Comprehension. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 120–125, Marseille, France. European Language Resources Association.

Jonathan Brennan. 2016. Naturalistic Sentence Comprehension in the Brain. *Language and Linguistics Compass*, 10(7):299–313.

Jonathan Brennan, Yuval Nir, Uri Hasson, Rafael Malach, David J. Heeger, and Liina Pylkkänen. 2012. Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and Language*, 120(2):163–173.

Jonathan R. Brennan, Chris Dyer, Adhiguna Kuncoro, and John T. Hale. 2020. Localizing syntactic predictions using recurrent neural network grammars. *Neuropsychologia*, 146:107479.

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

Eugene Charniak, Don Blaheta, Niyu Ge, Keith Hall, John Hale, and Mark Johnson. 2000. BLLIP 1987-89 WSJ Corpus Release 1.

Do Kook Choe and Eugene Charniak. 2016. Parsing as Language Modeling. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2331–2336, Austin, Texas. Association for Computational Linguistics.

Benoit Crabbé, Murielle Fabre, and Christophe Pallier. 2019. Variable beam search for generative neural parsing and its relevance for the analysis of neuro-imaging signal. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1150–1160, Hong Kong, China. Association for Computational Linguistics.

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

Jakub Dotlačil. 2021. Parsing as a Cue-Based Retrieval Model. *Cognitive Science*, 45(8):e13020.

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

Victoria Fossum and Roger Levy. 2012. Sequential vs. Hierarchical Syntactic Models of Human Incremental Sentence Processing. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2012)*, pages 61–69, Montréal, Canada. Association for Computational Linguistics.

Richard Futrell, Edward Gibson, and Roger P. Levy. 2020. Lossy-Context Surprisal: An Information-Theoretic Model of Memory Effects in Sentence Processing. *Cognitive Science*, 44(3):e12814.

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

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, pages 94–126. The MIT Press, Cambridge, MA, US.

Mario Giulianelli, Luca Malagutti, Juan Luis Gastaldi, Brian DuSell, Tim Vieira, and Ryan Cotterell. 2024. On the Proper Treatment of Tokenization in Psycholinguistics. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18556–18572, Miami, Florida, USA. Association for Computational Linguistics.

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

John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan Brennan. 2018. 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)*, pages 2727–2736, Melbourne, Australia. Association for Computational Linguistics.

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

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

Shinnosuke Isono. 2024. Category Locality Theory: A unified account of locality effects in sentence comprehension. *Cognition*, 247:105766.

Aravind K. Joshi. 1990. Processing crossed and nested dependencies: An automation perspective on the psycholinguistic results. *Language and Cognitive Processes*, 5(1):1–27.

Kohei Kajikawa, Ryo Yoshida, and Yohei Oseki. 2024. Dissociating Syntactic Operations via Composition Count. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46(0).

Alan Kennedy, Robin Hill, and Joël Pynte. 2003. The dundee corpus. In *Proceedings of the 12th European Conference on Eye Movement*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. 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)*, pages 7057–7075, Online. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. 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*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Adhiguna Kuncoro, Miguel Ballesteros, Lingpeng Kong, Chris Dyer, Graham Neubig, and Noah A. Smith. 2017. What Do Recurrent Neural Network Grammars Learn About Syntax? In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1249–1258, Valencia, Spain. Association for Computational Linguistics.

Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13):1–26.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Richard L. Lewis and Shravan Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3):375–419.

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

Danny Merkx and Stefan L. Frank. 2021. Human Sentence Processing: Recurrence or Attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 12–22, Online. Association for Computational Linguistics.

James A. Michaelov, Megan D. Bardolph, Seana Coulson, and Benjamin Bergen. 2021. Different kinds of cognitive plausibility: Why are transformers better than RNNs at predicting N400 amplitude? *Proceedings of the Annual Meeting of the Cognitive Science Society*, 43(43).

D. C. Mitchell. 1984. An Evaluation of Subject-Paced Reading Tasks and Other Methods for Investigating Immediate Processes in Reading 1. In *New Methods in Reading Comprehension Research*. Routledge.

Bruno Nicenboim, Shravan Vasishth, Felix Engelmann, and Katja Suckow. 2018. Exploratory and Confirmatory Analyses in Sentence Processing: A Case Study of Number Interference in German. *Cognitive Science*, 42(S4):1075–1100.

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

Byung-Doh Oh and William Schuler. 2022. 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*, pages 9324–9334, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Byung-Doh Oh and William Schuler. 2024. Leading Whitespaces of Language Models' Subword Vocabulary Pose a Confound for Calculating Word Probabilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3464–3472, Miami, Florida, USA. Association for Computational Linguistics.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

Philip Resnik. 1992. Left-Corner Parsing and Psychological Plausibility. In *COLING 1992 Volume 1: The 14th International Conference on Computational Linguistics*.

Soo Hyun Ryu and Richard Lewis. 2021. 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*, pages 61–71, Online. Association for Computational Linguistics.

Soo Hyun Ryu and Richard L. Lewis. 2022. Using Transformer language model to integrate surprisal, entropy, and working memory retrieval accounts of sentence processing. In *Proceedings of the 35th Annual Conference on Human Sentence Processing*, Santa Cruz, CA, USA.

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

Cory Shain, Idan Asher Blank, Marten van Schijndel, William Schuler, and Evelina Fedorenko. 2020. fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, 138:107307.

Cory Shain, Marten van Schijndel, Richard Futrell, Edward Gibson, and William Schuler. 2016. Memory access during incremental sentence processing causes reading time latency. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 49–58, Osaka, Japan. The COLING 2016 Organizing Committee.

Miloš Stanojević, Shohini Bhattasali, Donald Dunagan, Luca Campanelli, Mark Steedman, Jonathan Brennan, and John Hale. 2021. Modeling Incremental Language Comprehension in the Brain with Combinatory Categorial Grammar. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 23–38, Online. Association for Computational Linguistics.

Mark Steedman. 2000. *The Syntactic Process*. MIT Press, Cambridge, MA, USA.

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

Yushi Sugimoto, Ryo Yoshida, Hyeonjeong Jeong, Masatoshi Koizumi, Jonathan R. Brennan, and Yohei Oseki. 2024. Localizing Syntactic Composition with Left-Corner Recurrent Neural Network Grammars. *Neurobiology of Language*, 5(1):201–224.

William Timkey and Tal Linzen. 2023. A Language Model with Limited Memory Capacity Captures Interference in Human Sentence Processing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8705–8720, Singapore. Association for Computational Linguistics.

Julie A Van Dyke and Richard L Lewis. 2003. 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*, 49(3):285–316.

Julie A. Van Dyke and Brian McElree. 2011. Cue-dependent interference in comprehension. *Journal of Memory and Language*, 65(3):247–263.

Julie Ann Van Dyke. 2002. *Retrieval Effects in Sentence Parsing and Interpretation*. University of Pittsburgh ETD, University of Pittsburgh.

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

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

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

Lee H. Wurm and Sebastiano A. Fisicaro. 2014. What residualizing predictors in regression analyses does (and what it does not do). *Journal of Memory and Language*, 72:37–48.

A Hyperparameters

The hyperparameters are shown in Table 4. All model hyperparameters follow Sartran et al. (2022); Wolfman et al. (2024), while training hyperparameters were adjusted to accommodate the batch size suitable for our computational resources (NVIDIA A100, 40GB). The total computational cost required for all experiments was approximately 225 GPU hours.

B Correlations between the predictors

The correlations between the predictors in our statistical analysis are shown in Table 5. While the NAE from different LMs shows very high correlations with each other, TG and Transformer provide independent predictive power for the self-paced reading times; TG subsumes the predictive power of $\text{TG}_{-\text{comp}}$ (see Section 4.2 and 5.1).[20]

C Part-of-speech tags

Table 5 presents the complete list of part-of-speech (POS) and symbol tags in the Natural Stories corpus. As reading times are annotated for each whitespace-delimited region, for data points containing symbol tags (e.g., `NNP.`), we used the stripped version (e.g., `NNP`) in our analysis. Additionally, we excluded from our analysis any data points containing multiple POS tags (e.g., `NNP POS`).

D Parallel parsing experiment

As a conceptual case study for the local ambiguity resolution in syntactic structures behind token sequences, we implemented TG's NAE calculation using 10 syntactic structures obtained through word-synchronous beam search (Stern et al., 2017) with Recurrent Neural Network Grammar (RNNG; Dyer et al., 2016; Kuncoro et al., 2017; Noji and Oseki, 2021).[21][22] NAE was computed individually for each syntactic structure and then aggregated as a weighted average:

$$\text{NAE\_BS}_{l,h,w} := \frac{\sum_{t \in \text{Beam}_w} p(t) \cdot \text{NAE}^{t}_{l,h,\tau(w,t)}}{\sum_{t \in \text{Beam}_w} p(t)}, \quad (4)$$

where Beam_w represents the set of syntactic structures synchronized at the w-th word ($|\text{Beam}_w|=10$), and $\tau(w,t)$ denotes the token position corresponding to the w-th word in structure t.[23]

The analysis revealed patterns consistent with those observed when considering only the globally correct syntactic structure: both LMs' NAE demonstrated significant predictive power for reading times, with TG's NAE showing stronger contributions compared to Transformer's (Table 6).[24] The likelihood ratio tests further confirmed independent contributions from both LMs ($p < 0.001$ for both comparisons: 'TG & Transformer'$>$'Transformer' and 'TG & Transformer'$>$'TG').

E Surprisal experiment

We analyzed each LM's surprisal contribution to reading time prediction using a baseline regression model that excluded both LMs' surprisal from Equation 3 but included their NAE (Table 7). While both LMs' surprisal demonstrated significant predictive power for reading times, Transformer's surprisal exhibited a stronger contribution compared to TG's. Additionally, our likelihood ratio tests using the averaged surprisal revealed that the regression model incorporating both LMs' surprisal showed significantly higher predictive power compared to models with either LM's surprisal alone ($p < 0.001$ for both comparisons: 'TG & Transformer'$>$'Transformer' and 'TG & Transformer'$>$'TG'). These findings suggest that (i) unlike attention mechanisms, next-token prediction based solely on token sequences more effectively captures dominant factors of human predictive processing, but (ii) similar to attention mechanisms, both types of next-token prediction—those based on token sequences alone and those leveraging both syntactic structures and token sequences—may coexist as models that capture distinct aspects of human predictive processing.

F Direct comparison of TG and $\text{TG}_{-\text{comp}}$

In Section 5.1, we explored the advantage of treating closed phrases as single representations beyond syntactic structure consideration. Our analysis incorporated Transformer's NAE in the baseline regression model to distinguish between two effects

[20] `_mcomp` indicates $-$comp.

[21] `https://github.com/aistairc/rnng-pytorch`

[22] RNNG was trained on BLLIP-LG using default hyperparameters. For inference, action beam size and fast track were set to 100 and 1, respectively.

[23] `stack_count` was similarly calculated as the weighted average across syntactic structures in Beam_w.

[24] `_bs` indicates beam search.

Model architecture	Transformer-XL (Dai et al., 2019)
Vocabulary size	32,768
Model dimension	1,024
Feed-forward dimension	4,096
Number of layers	16
Number of heads	8
Segment length	256
Memory length	256
Optimizer	Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$) (Kingma and Ba, 2015)
Batch size	16
Number of training steps	400,000
Learning rate scheduler	Linear warm-up & cosine annealing
Number of warm-up steps	32,000
Initial learning rate	2.5×10^{-8}
Maximum learning rate	3.75×10^{-5}
Final learning rate	7.5×10^{-8}
Dropout rate	0.1

Table 4: Model and training hyperparameters

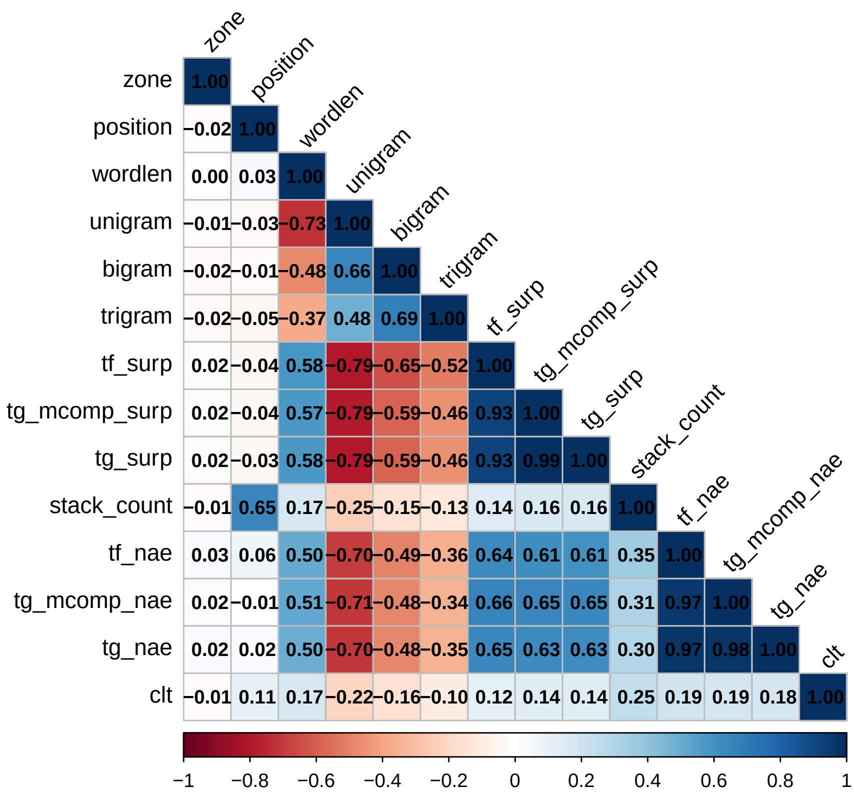

Figure 5: Correlations between the predictors in our statistical analysis

CC	Coordinating conjunction	PRP$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential *there*	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition or subordinating conjunction	TO	*to*
JJ	Adjective	UH	Interjection
JJR	Adjective, comparative	VB	Verb, base form
JJS	Adjective, superlative	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund or present participle
NN	Noun, singular or mass	VBN	Verb, past participle
NNS	Noun, plural	VBP	Verb, non-3rd person singular present
NNP	Proper noun, singuler	VBZ	Verb, 3rd person singular present
NNPS	Proper noun, plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoun
POS	Possessive ending	WP$	Possessive wh-pronoun
PRP	Personal pronoun	WRB	Wh-adverb
-LRB-	Left round bracket	,	Comma
-RRB-	Right round bracket	.	Period
“	Open double quotes	:	Colon
”	Closing double quotes		

Table 5: POS and symbol tags in the Natural Stories corpus

Model	ΔLogLik (↑)	Predictor	Effect size [ms]	p-value range	Significant seeds
TG	**56.0** (±5.5)	tg_bs_nae	1.08 (± 0.1)	<**0.001**	3/3
		tg_bs_nae_so	2.01 (± 0.1)	<**0.001**	3/3
Transformer	26.1 (±8.1)	tf_nae	0.95 (± 0.1)	<**0.001**	3/3
		tf_nae_so	1.25 (± 0.2)	<**0.001**	3/3

Table 6: TG's and Transformer's NAE contribution to reading time prediction, where TG's NAE was calculated with multiple syntactic structures generated by word-synchronous beam search with RNNG

Model	ΔLogLik (↑)	Predictor	Effect size [ms]	p-value range	Significant seeds
TG	62.6 (±6.8)	tg_surp	1.36 (± 0.0)	<**0.001**	3/3
		tg_surp_so	1.88 (± 0.1)	<**0.001**	3/3
Transformer	**159** (±12)	tf_surp	2.47 (± 0.1)	<**0.001**	3/3
		tf_surp_so	2.98 (± 0.2)	<**0.001**	3/3

Table 7: TG's and Transformer's surprisal contribution to reading time prediction

Model	ΔLogLik (↑)	Predictor	Effect size [ms]	p-value range	Significant seeds
TG	**78.2** (±6.9)	tg_nae	1.43 (± 0.2)	<**0.001**	3/3
		tg_nae_so	2.28 (± 0.3)	<**0.001**	3/3
$\text{TG}_{-\text{comp}}$	59.4 (±17)	tg_mcomp_nae	1.36 (± 0.2)	<**0.001**	3/3
		tg_mcomp_nae_so	1.90 (± 0.2)	<**0.001**	3/3

Table 8: TG's and $\text{TG}_{-\text{comp}}$'s NAE contribution to reading time prediction with Transformer's NAE excluded from the regression baseline model

in $\text{TG}_{-\text{comp}}$: syntactic structure consideration and direct terminal token access.

To evaluate which model—TG or $\text{TG}_{-\text{comp}}$—better captures more dominant factors in human sentence processing as a single model, we assessed their predictive power without Transformer's NAE in the baseline regression model (Table 8). The analysis revealed TG's superior predictive power (ΔLogLik=78.2) compared to $\text{TG}_{-\text{comp}}$ (ΔLogLik=59.4). These results highlight that TG, which explicitly treats closed phrases as single representations, outperforms $\text{TG}_{-\text{comp}}$, even when considering $\text{TG}_{-\text{comp}}$'s advantage in direct terminal token access. The likelihood ratio tests confirmed TG's independent predictive power from $\text{TG}_{-\text{comp}}$ ('TG & $\text{TG}_{-\text{comp}}$'>'$\text{TG}_{-\text{comp}}$', $p < 0.001$), while in contrast to Section 5.1, $\text{TG}_{-\text{comp}}$ also accounted for unique variance ('TG & $\text{TG}_{-\text{comp}}$'>'TG', $p < 0.01$). This bidirectional relationship likely emerges because $\text{TG}_{-\text{comp}}$'s direct terminal token access explains unique variance in the absence of Transformer's NAE.

G License

Table 9 summarizes the licenses of the data and tools employed in this paper. All data and tools were used under their respective license terms.

Dataset/Tool	License
BLLIP (Charniak et al., 2000)	BLLIP 1987–89 WSJ Corpus Release 1
Natural Stories corpus (Futrell et al., 2018)	CC BY-NC-SA 4.0
transformer_grammar (Sartran et al., 2022)	Apache 2.0
rnng-pytorch (Noji and Oseki, 2021)	MIT License
SentencePiece (Kudo and Richardson, 2018)	Apache 2.0
R (version 4.4.2) (R Core Team, 2024)	GNU GPL ≥ 2
lme4 (version 1.1.34) (Bates et al., 2015)	GNU GPL ≥ 2
lmerTest (version 3.1.3) (Kuznetsova et al., 2017)	GNU GPL ≥ 2

Table 9: Licenses of datasets and tools