# Investigating Psychometric Predictive Power of Syntactic Attention

**Ryo Yoshida**
The University of Tokyo
yoshiryo0617@g.ecc.u-tokyo.ac.jp

**Yushi Sugimoto**
The University of Osaka
sugimoto.yushi.hmt@osaka-u.ac.jp

**Yohei Oseki**
The University of Tokyo
oseki@g.ecc.u-tokyo.ac.jp

## Abstract

In computational psycholinguistics, Merkx and Frank (2021) demonstrated that surprisal values from Transformers exhibit a closer fit to measures of human reading effort than those from Recurrent Neural Networks (RNNs), suggesting that Transformers' attention mechanisms may capture cue-based retrieval-like operations in human sentence processing. Meanwhile, explicit integration of syntactic structures has been shown to improve language models' ability to model human sentence processing—for example, Hale et al. (2018) demonstrated that Recurrent Neural Network Grammars (RNNGs), which integrate RNNs with explicit syntactic structures, account for human brain activities that vanilla RNNs cannot capture. In this paper, we investigate the psychometric predictive power of Composition Attention Grammars (CAGs), which integrate Transformers with explicit syntactic structures, to test whether they provide a better fit to human reading times than both vanilla Transformers and RNNGs. We hypothesized that CAGs' syntactic attention mechanisms capture cue-based retrieval-like operations over syntactic memory representations—operations that may be involved in human sentence processing. The results of our strictly controlled experiments demonstrate that CAGs outperformed vanilla Transformers and RNNGs, suggesting that syntactic attention mechanisms of CAGs may serve as a mechanistic implementation of cue-based retrieval from syntactic memory.

## 1 Introduction

In computational psycholinguistics, language models (LMs) developed in Natural Language Processing (NLP) have been evaluated for their ability to model human sentence processing. Recurrent Neural Networks (RNNs; Elman, 1990), which process sequential representations recurrently, have traditionally been considered a practical implementation that demonstrates strong correspondence with human sentence processing, with their surprisal values successfully correlating with human reading times (Goodkind and Bicknell, 2018) and brain activities (Frank et al., 2015). Recently, Transformers (Vaswani et al., 2017), which have achieved state-of-the-art results on various downstream tasks, have also been tested for their power to predict human reading effort. Merkx and Frank (2021) demonstrated that Transformers outperformed RNNs in predicting human reading times and brain activities, suggesting that Transformers' attention mechanisms may provide a computational parallel to cue-based retrieval (Van Dyke and Lewis, 2003), a theory of human memory retrieval proposed in psycholinguistics.

While RNNs and Transformers primarily process sequential representations, the previous literature on computational psycholinguistics has empirically shown that explicit integration of syntactic structures can significantly improve LMs' ability to model human sentence processing. For instance, Hale et al. (2018) showed that Recurrent Neural Network Grammars (RNNGs; Dyer et al., 2016), which integrate RNNs with explicit syntactic structures, capture variance in human brain activities that cannot be accounted for by vanilla RNNs.[1]

Given that (i) Transformers may capture cue-based retrieval-like operations in human sentence processing and (ii) LMs integrated with explicit syntactic structures may capture variance in human syntactic processing, we investigate whether the integration of these two approaches might provide a better fit to measures of human reading effort

[1]More recently, Wolfman et al. (2024) showed that surprisal values from Transformer Grammars (TGs; Sartran et al., 2022), which integrate Transformers with explicit syntactic structures, also explain human brain activities that vanilla Transformers cannot predict. While their work and ours are similar in that both investigate the advantage of explicit integration of syntactic structures on Transformers, we additionally investigate the advantage of syntactic attention over syntactic recurrence, a research question not addressed in Wolfman et al. (2024).



*Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 407–418
July 31 - August 1, 2025 

than LMs employing either approach in isolation. Specifically, we investigate the psychometric predictive power of Composition Attention Grammars (CAGs; Yoshida and Oseki, 2022), which integrate Transformers with explicit syntactic structures, to test whether they provide a better fit to human reading times than both vanilla Transformers and RNNGs. We hypothesize that CAGs' syntactic attention mechanisms capture cue-based retrieval-like operations over syntactic memory representations—operations that may be involved in human sentence processing. The results of our controlled experiments demonstrate that CAGs outperformed vanilla Transformers and RNNGs, suggesting that syntactic attention mechanisms of CAGs may serve as a mechanistic implementation of cue-based retrieval from syntactic memory.[2]

## 2 Background

### 2.1 Psychometric predictive power

In psycholinguistics, it is well established that humans predict the next word during sentence processing (i.e., expectation-based theories), and the less predictable the next word is, the more effort is required to process it. The computational psycholinguistics literature (Hale, 2001; Levy, 2008) quantifies this predictability as *surprisal*, the negative log probability of a word given the context:

$$\text{surprisal} = -\log p(\text{word}|\text{context}). \quad (1)$$

Previous work has employed this information-theoretic complexity metric to link LMs' probability estimates with human reading effort (Smith and Levy, 2013; Goodkind and Bicknell, 2018). Building upon this paradigm, the computational psycholinguistics community has investigated LMs with high psychometric predictive power—i.e., LMs that can compute surprisal values with trends similar to measures of human reading effort—by comparing surprisal from various models with reading times or brain activities from humans (Frank and Bod, 2011; Fossum and Levy, 2012; Frank et al., 2015; Hale et al., 2018; Brennan and Hale, 2019; Wilcox et al., 2020; Brennan et al., 2020; Merkx and Frank, 2021; Kuribayashi et al., 2022; Wolfman et al., 2024, *inter alia*).

[2]Code for reproducing our results is available at `https://github.com/osekilab/CAG-EyeTrack`.

### 2.2 Sequential recurrence vs. sequential attention

RNNs (Elman, 1990) process sequential information (i.e., word embeddings) in a recurrent manner; they maintain a single vector representing a "context" and, at each time step, update this context vector with the embedding of the current input word (implementing sequential recurrence; Figure 1a). In contrast, recently introduced Transformers (Vaswani et al., 2017) employ an attention mechanism; they maintain all previous word embeddings and, at each time step, generate a context vector by selectively attending to them (implementing sequential attention; Figure 1b). Taking advantage of direct access to previous information, Transformers have been shown to outperform RNNs in various NLP tasks (cf. Wang et al., 2018, 2020).

Recently, the computational psycholinguistics community has also investigated whether Transformers have an advantage over RNNs in psychometric predictive power. Merkx and Frank (2021) compared Transformers and RNNs on their predictive power for human reading times and brain activities. The results showed that Transformers generally outperformed RNNs, suggesting that sequential attention, implemented by Transformers, captures aspects of human reading effort that sequential recurrence, implemented by RNNs, cannot account for.

Based on these findings, Merkx and Frank (2021) argued that the explained effort might be attributed to cue-based retrieval-like operations during human sentence processing (Van Dyke and Lewis, 2003). The cue-based retrieval theory posits that human sentence processing involves memory retrieval, where elements are retrieved from working memory based on cues provided by the current input word. Merkx and Frank's (2021) argument was that Transformers' attention mechanism—selective attention to previous word embeddings based on Queries from current input and Keys from previous words—might serve as a mechanistic implementation of this cue-based memory retrieval. Consequently, surprisal values from the attention mechanism would show similar trends to human reading effort, serving as the *causal bottleneck* (Levy, 2008).[3]

[3]Complementary research has examined the relationship between attention-based metrics (such as attention entropy) and human reading effort to assess the validity of attention mechanisms as a mechanistic implementation of cue-based retrieval (Ryu and Lewis, 2021; Oh and Schuler, 2022).

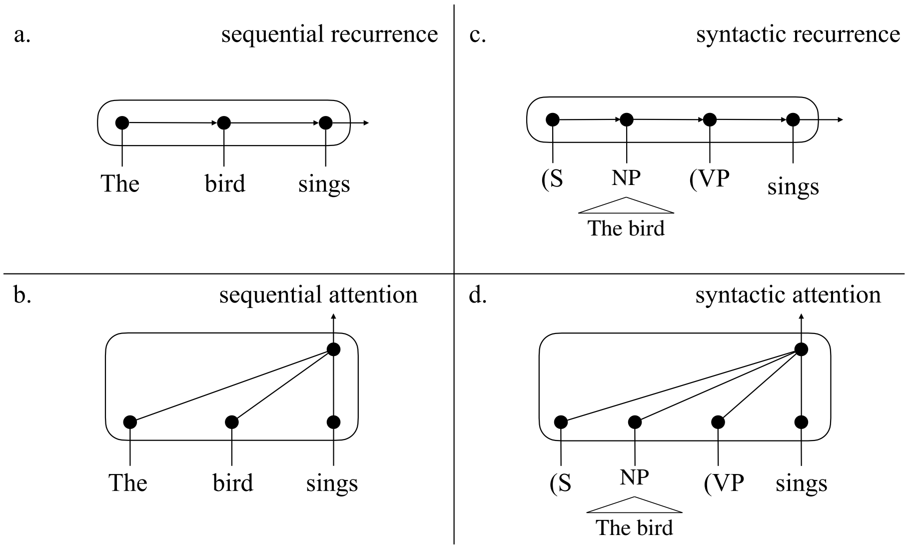


Figure 1: Four types of architectures. Previous work has investigated three types of architectural comparisons: (i) recurrence vs. attention in sequential architectures (a vs. b), (ii) sequential vs. syntactic in recurrent architectures (a vs. c), and (iii) sequential vs. syntactic in attention architectures (b vs. d). In this paper, we complete this comparison framework by directly comparing recurrence vs. attention in syntactic architectures (c vs. d).

More recently, Michaelov et al. (2021) replicated Merkx and Frank's (2021) results and presented additional analysis suggesting that Transformers can better capture human semantic facilitation effects than RNNs.

### 2.3 Sequential vs. syntactic

Although RNNs and Transformers have shown non-negligible results in psychometric predictive power, these architectures are fundamentally "sequential" models that process sequential information—without explicitly modeling the hierarchical syntactic structures of natural languages. The distinction between vanilla LMs and *syntactic LMs* such as RNNGs lies in this structural aspect—syntactic LMs not only generate a word sequence but also explicitly construct its underlying syntactic structure. Specifically, syntactic LMs jointly generate sentences and their syntactic structures through next-action prediction for the following three actions:

- `(X`: Generate a non-terminal symbol `(X`, where `X` represents a phrasal tag (e.g., `NP`). The vector representing the phrasal tag is placed on top of the *stack*, which maintains a list of vectors corresponding to the current context in syntactic LMs.

- `w`: Generate a terminal symbol `w`, where `w` represents a word (e.g., `bird`). The vector representing the word is placed on top of the stack.

- `)`: Close the most recent open non-terminal symbol. The vectors that constitute the closed phrase (i.e., the closed phrasal tag and its constituent vectors) are typically combined into a single vector representation using a *composition function* and placed on top of the stack. However, some syntactic LMs omit this composition step and simply place a vector representing the phrase closure on top of the stack (henceforth, we denote this type of syntactic LM with the subscript $_{-\text{comp}}$).

Computational psycholinguistics studies have shown that syntactic LMs outperform their vanilla LM counterparts in psychometric predictive power, suggesting that syntactic LMs can capture non-trivial variance in human syntactic processing. For instance, RNNGs, which recurrently summarize the stack state using RNNs (Dyer et al., 2015) (implementing syntactic recurrence; Figure 1c), can predict patterns in human brain activity (Hale et al., 2018) and human reading time (Yoshida et al., 2021) that vanilla RNNs cannot. Hale et al. (2018) also showed the advantage of the composition function, demonstrating that RNNGs$_{-\text{comp}}$ cannot explain the brain activity that RNNGs can.

More recently, Wolfman et al. (2024) demonstrated that Transformer Grammars (TGs; Sartran et al., 2022), which summarize the stack state by selectively attending to previous vectors using Transformers (implementing syntactic attention; Figure 1d), also explain human brain activities more successfully than vanilla Transformers.[4]

## 3 Syntactic recurrence vs. syntactic attention

As reviewed in Section 2, previous work has investigated three types of architectural comparisons: (i) recurrence vs. attention in sequential architectures (Merkx and Frank, 2021; Michaelov et al., 2021) (Figure 1a vs. 1b), (ii) sequential vs. syntactic in recurrent architectures (Hale et al., 2018; Yoshida et al., 2021) (Figure 1a vs. 1c), and (iii) sequential vs. syntactic in attention architectures (Wolfman et al., 2024) (Figure 1b vs. 1d). In this paper, we complete this comparison framework by directly comparing recurrence vs. attention in syntactic architectures (Figure 1c vs. 1d).

We hypothesize that syntactic attention—where previous vectors "in the stack" are selectively attended to based on Queries from current input and Keys from previous vectors—might show superior psychometric predictive power over syntactic recurrence by capturing cue-based retrieval-like operations over "syntactic memory representations"—operations that may be involved in human sentence processing. This hypothesis extends Merkx and Frank's (2021) argument that sequential attention (implemented by vanilla Transformers) outperforms sequential recurrence (implemented by RNNs), capturing cue-based retrieval-like operations over sequential memory representations.

LMs that implement syntactic attention include Transformer Grammars (TGs; Sartran et al., 2022) and Composition Attention Grammars (CAGs; Yoshida and Oseki, 2022). Both TGs and CAGs are syntactic LMs based on Transformers and employ composition functions. For our investigation, we employ CAGs for three reasons. First, CAGs' implementation includes word-synchronous beam search (Stern et al., 2017), an inference technique commonly used in computational psycholinguistics to model human local ambiguity resolution through parallel parsing (Hale et al., 2018; Sugimoto et al., 2024) (see Section 4.3 for details), whereas TGs lack this capability. Second, CAGs' probability estimation aligns more closely with human offline grammaticality judgments than TGs (Yoshida and Oseki, 2022). Third, CAGs employ bidirectional LSTMs for the composition function, which is the same implementation used in RNNGs, while TGs implement the composition function via attention masks. This design choice enables a more controlled comparison between syntactic recurrence and syntactic attention, as the architectures differ only in their stack summarization process.

## 4 Method

We evaluate four LMs that employ either selective attention or recurrent processing on word sequences or syntactic structures, comparing their psychometric predictive power for human reading times using the Zurich Cognitive Language Processing Corpus (ZuCo; Hollenstein et al., 2018). Following Hale et al. (2018), we also include degraded versions of syntactic LMs that lack the composition function. The following subsections describe our experimental settings in detail.

### 4.1 Language models

In our experiment, we trained LMs with strictly controlled hyperparameters following Yoshida and Oseki (2022), as their model sizes were made maximally comparable.

**LSTM (sequential recurrence)** Long Short-Term Memories (LSTMs; Hochreiter and Schmidhuber, 1997) are LMs that perform recurrent processing on word sequences. We used 2-layer LSTMs with 301 hidden and input dimensions (model size: 16.59M).[5]

**RNNG (syntactic recurrence)** Recurrent Neural Network Grammars (RNNGs; Dyer et al., 2016) are LMs that perform recurrent processing on syntactic structures. RNNGs are equipped with a composition function based on bidirectional LSTMs. We used stack-only RNNGs (Kuncoro et al., 2018; Noji and Oseki, 2021) with 2-layer stack LSTMs with 276 hidden and input dimensions (model size: 16.61M).[6]

**RNNG$_{-\text{comp}}$ (degraded syntactic recurrence)** RNNGs$_{-\text{comp}}$ (Choe and Charniak, 2016; Hale et al., 2018) are a degraded version of RNNGs without the composition function. We used

[4] Yoshida et al. (2025) also demonstrated that attention entropy derived from TGs can predict human reading times more successfully than vanilla Transformers.

[5] We implemented LSTMs using the PyTorch package (https://github.com/pytorch/pytorch).

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

RNNGs$_{-\text{comp}}$ with 2-layer LSTMs with 301 hidden and input dimensions (model size: 16.58M).

**Transformer (sequential attention)** Transformers (Radford et al., 2018) are LMs that perform selective attention on word sequences. We used 3-layer 4-head Transformers with 272 hidden and input dimensions (model size: 16.62M).[7]

**CAG (syntactic attention)** Composition Attention Grammars (CAGs; Yoshida and Oseki, 2022) are LMs that perform selective attention on syntactic structures. CAGs are equipped with a composition function based on bidirectional LSTMs. We used 3-layer 4-head CAGs with 256 hidden and input dimensions (model size: 16.57M).[8]

**CAG$_{-\text{comp}}$ (degraded syntactic attention)** CAGs$_{-\text{comp}}$ (Qian et al., 2021) are a degraded version of CAGs without the composition function. We used 3-layer 4-head CAGs$_{-\text{comp}}$ with 272 hidden and input dimensions (model size: 16.63M).[9]

### 4.2 Training data

All LMs were trained using BLLIP-LG, which comprises 1.8M sentences and 42M tokens sampled from the Brown Laboratory for Linguistic Information Processing 1987-89 Corpus Release 1 (BLLIP; Charniak et al., 2000). The train-dev-test split followed Hu et al. (2020). Following Qian et al. (2021), sentences were tokenized into subwords using a Byte Pair Encoding tokenizer (Sennrich et al., 2016) from the Huggingface Transformers package (Wolf et al., 2020).

All LMs were trained at the sentence level: LSTMs and Transformers were trained on terminal subwords, whereas RNNGs, RNNG$_{-\text{comp}}$, CAGs, and CAG$_{-\text{comp}}$ were trained on both terminal subwords and syntactic structures, which were parsed by Hu et al. (2020) using a state-of-the-art constituency parser (Kitaev and Klein, 2018). All LMs shared the same training hyperparameters: a learning rate of $10^{-3}$, a dropout rate of 0.1, the Adam optimizer (Kingma and Ba, 2015), and a minibatch size of 256. Training was conducted for 15 epochs. We selected the checkpoint with the lowest loss on the development set for evaluation and conducted experiments three times with different random seeds.

### 4.3 Eye tracking data

We used reading times from the Zurich Cognitive Language Processing Corpus (ZuCo; Hollenstein et al., 2018) to evaluate whether LMs can successfully predict human reading effort. ZuCo is a collection of single sentences from the Stanford Sentiment Treebank and the Wikipedia relation extraction corpus, accompanied by simultaneous eye-tracking and electroencephalography (EEG) recordings from 12 native English speakers. Although ZuCo comprises data from both normal reading and task-specific reading tasks, we used only 700 sentences from the natural reading task, following previous work (e.g., Hollenstein et al., 2021). During the natural reading task, sentences were displayed one by one, and participants read them at their own pace. During preprocessing by Hollenstein et al. (2018), fixations that were (i) shorter than 100 ms or (ii) recorded when EEG amplitude exceeded $\pm 90\,\mu\text{V}$ were removed due to irrelevance to reading activity or data quality concerns.

In this paper, first-pass reading time (the sum of all fixation times on a word before the eye moves away from it) was used as the prediction target.[10] Following the convention of psycholinguistic studies, we excluded words with missing values (e.g., non-fixations) or at sentence-initial and sentence-final positions from our statistical analysis. We further removed words that were out of vocabulary (OOV) in the large corpus (Wikitext-2; Merity et al., 2017) or words following OOV words, as frequency values are required for our baseline regression model. Consequently, 80,853 data points were included in the statistical analysis out of 161,597 total data points. The high proportion of deleted data points during preprocessing was mainly due to the large number of missing values (52,240 data points).

In previous computational psycholinguistic research, there was often a mismatch between LMs' processing level and human data collection procedures—for instance, LMs trained at the sentence level were evaluated against human data col-

[7] We implemented Transformers using the Huggingface Transformers package (https://github.com/huggingface/transformers).

[8] https://github.com/osekilab/CAG

[9] https://github.com/IBM/transformers-struct-guidance

[10] We first conduct validation using reading time as the most accessible and interpretable human data source, given that the specific event-related potential (ERP) components of EEG that would best reflect cue-based retrieval-like operations over syntactic memory representations remain to be determined.

lected during document-level reading (cf. Wilcox et al., 2020). In this paper, we address this gap by conducting more strictly controlled experiments using ZuCo, a corpus where eye-tracking data was recorded during sentence-level reading.[11]

Since only word sequences were input during surprisal calculation, we employed word-synchronous beam search (Stern et al., 2017) to infer syntactic structures for CAGs and RNNGs. Word-synchronous beam search retains a collection of the most likely syntactic structures given a partial word sequence and marginalizes their probabilities to approximate next-word probabilities. Hale et al. (2018) argued that the combination of syntactic LMs and word-synchronous beam search successfully captured human local ambiguity resolution during online sentence processing.[12]

### 4.4 Statistical analysis

We analyzed how well surprisal from each LM predicts human reading time, measuring improvements in regression model fit when adding surprisal values as predictors. For each LM, we included both the surprisal of the current word and the previous word to account for spillover effects (Mitchell, 1984).[13] As a measure of psychometric predictive power, we evaluated the per-token increase in log-likelihood (ΔLogLik) on the entire dataset. This evaluation was conducted for each random seed, and we report the mean psychometric predictive power with standard deviation.

Following previous studies such as Merkx and Frank (2021), the baseline regression model controlled for several predictors relevant to reading activity:

- `order` (integer): sentence display order during the reading task;
- `position` (integer): word position in the sentence;
- `length` and `prev_length` (integer): number of characters in the current and previous word;
- `freq` and `prev_freq` (continuous): log-transformed frequencies of the current and previous word.

Previous words' values were included for modeling the spillover effect. All numeric factors were $z$-transformed.

The baseline regression model was a linear mixed-effects model (Baayen et al., 2008) with these fixed effects and a by-subject random intercept:[14]

$$\begin{aligned}\log(\mathtt{RT}) \sim\ & \mathtt{order} + \mathtt{position} + \\ & \mathtt{length} + \mathtt{prev\_length} + \\ & \mathtt{freq} + \mathtt{prev\_freq} + \\ & (\mathtt{1} \,|\, \mathtt{subj}). \end{aligned} \qquad (2)$$

Before evaluating psychometric predictive power, we conducted baseline regression model-based data omission, removing data points beyond three standard deviations. This removed 559 data points, leaving 80,294 data points for the final statistical analysis.

### 4.5 Nested model comparison

We conducted nested model comparisons (Wurm and Fisicaro, 2014) to evaluate whether the differences in ΔLogLik are statistically significant. Specifically, we extended Equation 2 by adding surprisal values from two LMs versus adding surprisal values from only one LM, and tested the statistical significance of the deviance using the $\chi^2$ test ($p \leq 0.05$). Following Aurnhammer and Frank (2019), we used surprisal values averaged across different random seeds for these nested model comparisons.

## 5 Results

### 5.1 Overall

The Psychometric Predictive Power (PPP, per-token ΔLogLik) of each LM is summarized in Figure 2. The psychometric predictive power averaged across different random seeds (the vertical axis) is plotted against the LMs investigated in this paper (the horizontal axis). Error bars denote standard deviations across random seeds. We confirmed that the psychometric predictive power was statistically

[11] An alternative approach would be to train LMs at the document level and evaluate them on document-level reading data. However, we adopt the sentence-level setting because syntactic LMs are conventionally trained on sentences, and RNNGs and CAGs lack implementations applicable to document-level training.

[12] We set the action beam size to 100, word beam size to 10, and fast-track to 1. Word beam size corresponds to the number of syntactic structures to be marginalized.

[13] Following the convention of previous studies (e.g., Wilcox et al., 2020; Kuribayashi et al., 2021), the word-level surprisal was calculated as the cumulative surprisal of its constituent subwords.

[14] We implemented the regression model using the lme4 package (Bates et al., 2015) in R (R Core Team, 2024).

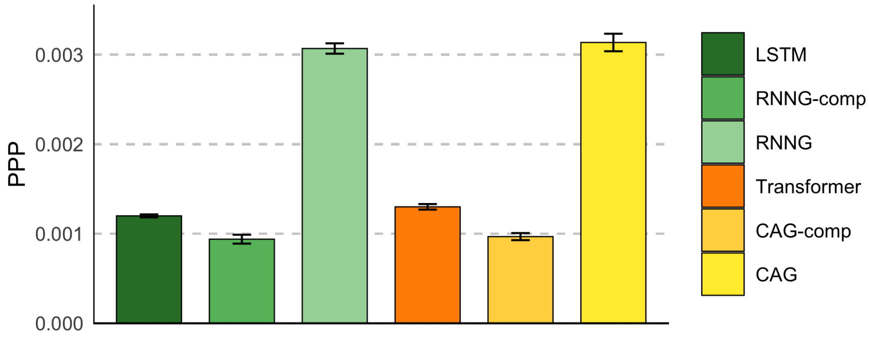


Figure 2: Psychometric Predictive Power (PPP, per-token ΔLogLik) of each LM. The psychometric predictive power averaged across different random seeds (vertical axis) is plotted against the LMs investigated in this paper (horizontal axis). Error bars denote standard deviations across random seeds.

significant for all LMs under nested model comparisons against the baseline regression model, and the direction was appropriate for reading time—that is, higher surprisal values corresponded to longer reading times. The results demonstrated that CAGs achieved the highest psychometric predictive power: CAG > RNNG > Transformer > LSTM > CAG$_{-\text{comp}}$ > RNNG$_{-\text{comp}}$, showing that the architecture performing syntactic attention captures the most variance in human reading time.

**Reproduction of sequential recurrence vs. sequential attention** In our experiment, Transformers outperformed LSTMs in psychometric predictive power. To confirm that this difference is statistically significant, the result of the nested model comparison is shown in the top block of Table 1. The nested model comparison revealed that Transformers significantly outperformed LSTMs, corroborating Merkx and Frank's (2021) finding that Transformers, which implement sequential attention, capture variance in human reading effort that RNNs, which implement sequential recurrence, cannot.[15]

**Reproduction of sequential vs. syntactic** In our experiment, RNNGs and CAGs outperformed LSTMs and Transformers, respectively. To confirm that these differences are statistically significant, the results of nested model comparisons are shown in the middle block of Table 1. The nested model comparisons revealed that RNNGs and CAGs significantly outperformed LSTMs and Transformers, respectively, supporting the findings of Hale et al. (2018) and Wolfman et al. (2024) that syntactic LMs can account for human reading effort that vanilla LMs cannot predict.

In addition, RNNGs and CAGs also significantly outperformed RNNGs$_{-\text{comp}}$ and CAGs$_{-\text{comp}}$, respectively, corroborating Hale et al.'s (2018) argument that the composition function is crucial for syntactic LMs to capture human syntactic processing. As a side note, RNNGs$_{-\text{comp}}$ and CAGs$_{-\text{comp}}$ underperformed LSTMs and Transformers, respectively. This implies that stack representations without the composition function not only harm the ability to account for syntactic processing but also cause a loss in simulating general human predictive processing. Hale et al. (2018) also showed a null result when comparing the psychometric predictive power of RNNGs$_{-\text{comp}}$ to that of LSTMs.

**Syntactic recurrence vs. syntactic attention** In our experiment, CAGs outperformed RNNGs in the absolute value of psychometric predictive power. To confirm that the difference between CAGs and RNNGs is statistically significant, the result of the nested model comparison is shown in the bottom block of Table 1. The nested model comparison revealed that CAGs significantly outperformed RNNGs, suggesting that CAGs, which implement syntactic attention, can successfully capture variance in human reading time that RNNGs, which implement syntactic recurrence, cannot account for.

[15] Incidentally, Merkx and Frank (2021) found the advantage of Transformers on self-paced reading times and EEG but obtained mixed results on first-pass reading time. Our more definitive findings may be attributed to our strictly controlled experimental settings, where Transformer advantages could become more consistently observable.

| | $\chi^2$ | df | $p$ |
|---|---|---|---|
| Sequential recurrence vs. sequential attention | | | |
| LSTM < TF | 16.75 | 2 | **0.00023** |
| Sequential vs. syntactic | | | |
| LSTM < RNNG | 315.7 | 2 | **<0.0001** |
| TF < CAG | 308.5 | 2 | **<0.0001** |
| RNNG$_{-c.}$ < RNNG | 369.8 | 2 | **<0.0001** |
| CAG$_{-c.}$ < CAG | 372.0 | 2 | **<0.0001** |
| **Syntactic recurrence vs. syntactic attention** | | | |
| RNNG < CAG | 11.42 | 2 | **0.00331** |

Table 1: Results of nested model comparisons from three perspectives: (i) reproduction of sequential recurrence vs. sequential attention, (ii) reproduction of sequential vs. syntactic, and (iii) syntactic recurrence vs. syntactic attention. TF and $_{-c.}$ indicate Transformer and $_{-comp}$, respectively.

### 5.2 Longer and shorter sentences

To investigate under what conditions syntactic attention has an advantage over syntactic recurrence, we split the data points in ZuCo into two subsets based on sentences longer or shorter than the average sentence length, following Merkx and Frank (2021). Merkx and Frank (2021) conducted this analysis expecting that longer sentences could accentuate Transformers' advantage of direct access to previous information. The longer and shorter subsets include 37,578 and 43,275 data points, respectively. We removed 601 and 703 data points that were beyond three standard deviations, leaving 37,307 and 42,997 data points for the final statistical analysis, respectively.

The psychometric predictive power of CAGs and RNNGs on longer and shorter sentences is shown in Figure 3. The results show that CAGs and RNNGs achieve comparable psychometric predictive power on shorter sentences, but CAGs outperformed RNNGs on longer sentences. To confirm that these differences are statistically significant, the results of nested model comparisons are shown in Table 2. The nested model comparisons revealed that CAGs significantly outperformed RNNGs only on longer sentences, consistent with their performance on the complete dataset.

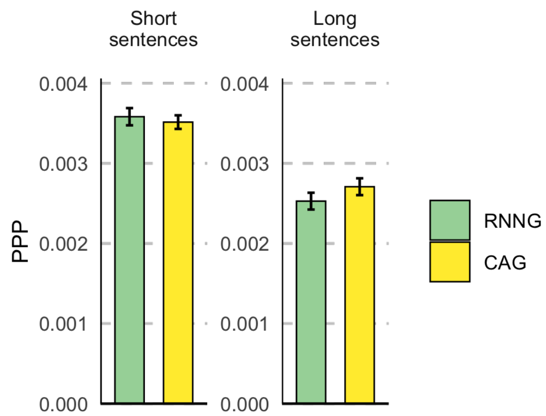


Figure 3: Psychometric predictive power (PPP, per-token ΔLogLik) of CAGs and RNNGs on longer and shorter sentences. The psychometric predictive power averaged across different random seeds (vertical axis) is plotted against the LMs (horizontal axis). Error bars denote standard deviations across random seeds.

| | $\chi^2$ | df | $p$ |
|---|---|---|---|
| Short sentences | | | |
| RNNG < CAG | 0.8359 | 2 | 0.6584 |
| Long sentences | | | |
| RNNG < CAG | 14.793 | 2 | **0.0006133** |

Table 2: Results of nested model comparisons on longer and shorter subsets of ZuCo

## 6 Discussion

In this paper, we reproduced the results of (i) sequential recurrence vs. sequential attention (cf. Merkx and Frank, 2021), (ii) sequential vs. syntactic (cf. Hale et al., 2018; Wolfman et al., 2024), and (iii) demonstrated that CAGs, which implement syntactic attention, achieve higher psychometric predictive power than both vanilla Transformers and RNNGs. Given that Merkx and Frank (2021) and Hale et al. (2018) suggest that attention mechanisms and syntactic LMs can serve as mechanistic implementations of human cue-based retrieval and syntactic processing, respectively, our results suggest that syntactic attention in CAGs may serve as a mechanistic implementation of cue-based retrieval from syntactic memory. This interpretation is also consistent with psycholinguistic studies demonstrating that memory costs derived from syntactic structures successfully predict reading times and brain activities (Isono, 2024; Shain et al., 2022)—findings that support the cognitive plausibility of syntactic memory representations and operations over them.

Furthermore, the analyses of longer versus shorter sentences suggest that cue-based retrieval-like operations over syntactic memory representations may become more prominent when processing longer sentences. Merkx and Frank (2021) demonstrated that Transformers' superior psychometric predictive power over RNNs was particularly pronounced on longer sentences, suggesting that retrieval operations may be especially important when accessing information from linearly distant words. While both CAGs and RNNGs can maintain information from linearly distant words through their composition functions, the direct access afforded by attention mechanisms nevertheless provides additional advantages as sentences get longer.

Interestingly, Wilcox et al. (2018) and Oh et al. (2021) found that syntactic LMs (i.e., RNNGs) underperformed LSTMs or Transformers in modeling human reading times and brain activities, contradicting the advantages observed by Hale et al. (2018), Wolfman et al. (2024), and our sequential vs. syntactic results. One potential explanation for these discrepancies lies in experimental control: while Wilcox et al. (2018) and Oh et al. (2021) compared LMs with varying model sizes, Hale et al. (2018), Wolfman et al. (2024) and our experiment all employed LMs with maximally comparable model sizes. Our approach further extends this methodology by aligning LMs' processing units with human data collection procedures at the sentence level. These results highlight the critical role of controlled experimental design, especially when comparing minimally different architectures.

## 7 Conclusion

In this paper, we investigated the psychometric predictive power of Composition Attention Grammars (CAGs) through strictly controlled experiments. Our results demonstrated that CAGs outperformed both vanilla Transformers and RNNGs, suggesting that syntactic attention may serve as a mechanistic implementation of cue-based retrieval from syntactic memory. Further analyses revealed that this result is primarily driven by improved performance on longer sentences, indicating that cue-based retrieval-like operations over syntactic memory representations became increasingly important as sentences got longer.

## Limitations

There are several limitations to this study. First, although we utilized CAGs as a model of syntactic attention, TGs could also serve as an alternative. While our choice of CAGs was motivated by (i) their word-synchronous beam search capability, (ii) better alignment to human offline grammaticality judgments, and (iii) their use of bidirectional LSTMs for the composition function (see Section 3), whether our positive results for syntactic attention generalize to TGs remains an open question.

Second, our experiments were based solely on reading time data from ZuCo. As noted earlier, we chose reading time as the most accessible and interpretable human data source, given that the specific event-related potential (ERP) components of EEG that would best reflect cue-based retrieval-like operations over syntactic memory representations remain to be determined. Future research should explore which ERP components might be most sensitive to these operations and extend the evaluation to additional measures of human sentence processing.

Third, while our sentence-level analysis provided technical advantages for controlled comparisons, extending these syntactic LMs to document-level processing would be valuable for future research, as this would enable controlled experiments on additional datasets (e.g., the Natural Stories corpus; Futrell et al., 2018).

## Acknowledgments

We would like to thank Shinnosuke Isono for his helpful suggestions on the manuscript. We also thank anonymous CoNLL 2025 reviewers for their insightful reviews. 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

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.

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

Jonathan R. Brennan and John T. Hale. 2019. Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PLOS ONE*, 14(1):e0207741.

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.

Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A. Smith. 2015. Transition-Based Dependency Parsing with Stack Long Short-Term Memory. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 334–343, Beijing, China. Association for Computational Linguistics.

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.

Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211.

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.

Stefan L. Frank and Rens Bod. 2011. Insensitivity of the Human Sentence-Processing System to Hierarchical Structure. *Psychological Science*, 22(6):829–834.

Stefan L. Frank, Leun J. Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, 140:1–11.

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).

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. 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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-term Memory. *Neural computation*, 9(8):1735–80.

Nora Hollenstein, Federico Pirovano, Ce Zhang, Lena Jäger, and Lisa Beinborn. 2021. Multilingual Language Models Predict Human Reading Behavior. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 106–123, Online. Association for Computational Linguistics.

Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific Data*, 5(1):180291.

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.

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.

Adhiguna Kuncoro, Chris Dyer, John Hale, Dani Yogatama, Stephen Clark, and Phil Blunsom. 2018. LSTMs Can Learn Syntax-Sensitive Dependencies Well, But Modeling Structure Makes Them Better. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1436, Melbourne, Australia. Association for Computational Linguistics.

Tatsuki Kuribayashi, Yohei Oseki, Ana Brassard, and Kentaro Inui. 2022. Context Limitations Make Neural Language Models More Human-Like. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10421–10436, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. 2021. Lower Perplexity is Not Always Human-Like. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5203–5217, Online. Association for Computational Linguistics.

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

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer Sentinel Mixture Models. In *International Conference on Learning Representations*.

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.

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, Christian Clark, and William Schuler. 2021. Surprisal Estimators for Human Reading Times Need Character Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3746–3757, 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.

Peng Qian, Tahira Naseem, Roger Levy, and Ramón Fernandez Astudillo. 2021. Structural Guidance for Transformer Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3735–3745, Online. Association for Computational Linguistics.

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

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. page 12.

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.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Cory Shain, Idan A. Blank, Evelina Fedorenko, Edward Gibson, and William Schuler. 2022. Robust effects of working memory demand during naturalistic language comprehension in language-selective cortex. *Journal of Neuroscience*, 42(39):7412–7430.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

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.

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.

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.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2020. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *Preprint*, arXiv:1905.00537.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN Language Models Learn about Filler–Gap Dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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.

Ryo Yoshida, Shinnosuke Isono, Kohei Kajikawa, Taiga Someya, Yushi Sugimoto, and Yohei Oseki. 2025. If Attention Serves as a Cognitive Model of Human Memory Retrieval, What is the Plausible Memory Representation? *Preprint*, arXiv:2502.11469.

Ryo Yoshida, Hiroshi Noji, and Yohei Oseki. 2021. Modeling Human Sentence Processing with Left-Corner Recurrent Neural Network Grammars. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2964–2973, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ryo Yoshida and Yohei Oseki. 2022. Composition, Attention, or Both? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5822–5834, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.