Memory for Prediction: A Transformer-based Integrative Account of Sentence Processing

by

Soo Hyun Ryu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Psychology)
in the University of Michigan
2025

Doctoral Committee:

Professor Richard Lewis, Chair
Associate Professor Jonathan Brennan
Assistant Professor Richard Futrell
Professor Thad Polk

Soo Hyun Ryu
soohyunr@umich.edu
ORCID iD: 0000-0001-6636-9765

ACKNOWLEDGEMENTS

First, I want to express my deepest gratitude to my advisor, Richard Lewis. It has been an incredible privilege to study under his unwavering support and brilliant insight over the past five years. The discussions we had and the questions he encouraged me to explore taught me how to frame my thoughts and seek meaningful answers. I believe what I learned from him will be one of my most valuable assets as I move forward. I am also truly thankful for my incredible committee members — Thad Polk, Jonathan Brennan, and Richard Futrell. Their thoughtful feedback not only strengthened this thesis but also opened my mind to new ways of thinking. I feel very fortunate to have learned so much from their expertise.

I would also like to thank my former advisors. Rui Chaves, whom I worked with during my Master's program in Computational Linguistics at the University at Buffalo, played a crucial role in laying the foundation. I am also grateful to Ansuk Jeong, whom I met during my undergraduate program at Yonsei University. She helped me take my first steps into research, consistently offering invaluable academic guidance and continued moral support.

I am grateful to the friends and colleagues I met at Michigan, both inside and outside the Department of Psychology. I especially thank my lab mates and cohort, whose warmth made my time in Michigan far more enjoyable. I am also deeply thankful for my friends in South Korea. Their unwavering encouragement carried me through the toughest moments.

Lastly, I want to thank my family. My mom has believed in me every step of the way, and I wouldn't have made it this far without her love. I'm also deeply grateful to my two sisters, who have always been there as my biggest supporters. I owe so much to my husband, Deokoh, who has been by my side from day one, through every moment of this journey. I doubt I could have enjoyed it as much without his support and endless supply of laughter.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

ABSTRACT

This thesis proposes that Transformer-based language models serve as integrative sentence processing models that combine expectation-based (e.g., surprisal theory) and memory-based (e.g., cue-based retrieval theory) accounts of sentence processing. To show that attention can estimate memory retrieval interference, I introduce a novel Transformer-based metric—attention entropy. By analyzing naturalistic sentence reading time data, I demonstrate that this new metric explains variance in sentence processing difficulty not accounted for by surprisal alone, presumably reflecting memory integration difficulty. To further validate attention entropy as an estimate of memory integration difficulty, I show that its effects on predicting sentence reading times vary depending on the speed–accuracy trade-offs imposed on participants, with the strongest effects observed in accuracy-emphasized conditions. I also show that attention entropy accounts for memory interference effects in psycholinguistic phenomena such as subject–verb agreement, embedded sentence structures, and relative clause processing. This thesis concludes with a discussion of limitations and future work.

CHAPTER 1

Introduction

This thesis aims to demonstrate that recent state-of-the-art neural network language models using the Transformer architecture (Vaswani et al., 2017) provide a new foundation for a cognitive model of human sentence processing, seamlessly integrating probabilistic expectation-based (Hale, 2001, 2016; Levy, 2008, 2013) and working memory interference-based accounts (Lewis et al., 2006; Lewis and Vasishth, 2005; Vasishth and Engelmann, 2021). Because the language models compute a probability distribution over the lexicon conditioned on left context, they yield word-by-word *surprisals* (Hale, 2001; Levy, 2008) which may be used to predict reading times from an expectation-based theoretic perspective. Additionally, I show that the Transformer's internal attention mechanism—which underlies its prediction process—enables the computation of a simple, word-by-word *attention entropy* metric. This metric has independent empirical power in accounting for processing difficulty and reading times, reflecting the working memory-based aspects of sentence processing

The insight motivating this novel attention entropy metric is that the Transformer is a kind of multi-level *cue-based retrieval* architecture, and the entropy measure is an index of *similarity-based interference*—the more similar (with respect to retrieval cues) the candidates are that compete to be retrieved, the more diffuse is the attention (Ryu and Lewis, 2021). The theoretical connection between cue-based retrieval parsing and attention architectures is deep, because they are based on the same functional motivation: provide a memory mechanism that is capable of handling long-distance dependencies in natural language, given sufficiently discriminating cues and distinct memory representations (Lewis and Vasishth,

2005; Lewis et al., 2006). But modern Transformers differ sharply from earlier cue-based parsers in that the cues and representations are *learned*. Furthermore, the revolutionary AI advances made possible by Transformer-based language models provide important evidence for the *functional sufficiency* of cue-based architectures as the basis of human language processing.

Analyses in this thesis suggest that Transformers are promising candidates for developing richer, adaptive and mechanistic models of human language processing and acquisition, due to their relative simplicity, the ease with which the surprisal and entropy metrics may be computed, and the increasing availability of both pre-trained models and software for training new models. This contribution thus provides a novel approach to integrating expectation and memory, complementing previous integrative models that combine expectation and memory. Specifically, this new approach complements lossy memory surprisal models by addressing a different aspect of memory: while lossy surprisal emphasizes the degradation of encoded memory over time, the present approach focuses on the memory retrieval process during the interpretation of new input.

In Chapter 2, I provide a brief overview of how neural language models have been applied to explain (psycho)linguistic phenomena. I also outline the two main theoretical accounts of sentence processing difficulty: expectation-based and memory-based accounts. I then motivate the need for an integrative perspective and review prior efforts in this direction.

In Chapter 3, I elaborate on the connection between cue-based retrieval parsers and the Transformer's attention mechanism. Building on this link, I introduce a novel metric—attention entropy—which estimates processing difficulty from the perspective of similarity-based interference during memory retrieval. I also describe the process of selecting attention heads for computing attention entropy, based on the functions the heads serve in processing grammar dependencies. I conclude the chapter by presenting the Transformer as an integrative model of sentence processing that seamlessly combines surprisal and cue-based retrieval theories.

In Chapter 4, I demonstrate that attention entropy provides a distinct explanation for sentence processing times, independently of surprisal. Specifically, I fit Bayesian mixed-effects regression models to predict word-level reading times using surprisal, attention entropy, and control predictors (e.g., word length, word position). The results reveal dissociable effects of surprisal and attention entropy and suggest that attention entropy captures memory integration difficulty that emerges at a later stage of sentence processing.

In Chapter 5, I conduct a self-paced reading experiment to collect sentence reading times under different speed–accuracy emphasis conditions. This study aims to examine how the effects of attention entropy vary as a function of the speed–accuracy tradeoff. Results from Bayesian mixed-effects regression modeling indicate that the effects of attention entropy are strongest when accuracy is prioritized over speed, reinforcing its role as a measure of memory retrieval difficulty during sentence processing.

In Chapter 6, I show that a range of psycholinguistic phenomena related to memory interference can be accounted for by attention entropy in ways that surprisal alone cannot. These include interference in subject-verb agreement, non-agreement interference effects, the processing of self-embedded sentences, relative clause processing, and garden-path effects.

In Chapter 7, I summarize the key findings, discuss directions for future work, and provide concluding remarks.

Taken together, the contributions of this thesis are as follows:

1. To show that the Transformer's attention mechanism functions as a cue-based retrieval model, enabling multi-level, parallel retrievals of linguistic representations, based on which next-word prediction is generated.

2. To illustrate Transformer-based language models can combine expectation-based and memory-based accounts for sentence processing difficulty prediction, especially exploiting the internal processing mechanism of the Transformer model for the first time.

3. To demonstrate that attention entropies and surprisals serve distinct roles in explaining

sentence processing difficulties — surprisals explain the early word identification stage and the probabilistic belief update stage; attention entropies explain the later memory integration stage.

4. To show how the combination of surprisal and attention entropy provides an integrated account of psycholinguistic phenomena, including subject-verb agreement interference, non-agreement effects, self-embedded sentence processing, relative clause processing and garden-path processing.

CHAPTER 2

Theoretical Foundations

In this chapter, I begin by discussing how large-scale neural network language models have contributed to ongoing debates in psycholinguistics. I then discuss two major accounts for explaining human sentence processing: the expectation-based and memory-based accounts. After outlining the specific aspects each account focuses on, and the motivation for integrating them, I review previous integrative models that have attempted to combine the two. I conclude with a discussion of the promise of interpreting Transformers as novel integrative sentence processing models, and suggest how this could complement prior integrative approaches.

2.1 Neural Network Language Models and Psycholinguistics

The success of neural network-based Language Models (LMs), such as Long Short-Term Memory (LSTM) models (Hochreiter and Schmidhuber, 1997) and Transformer models (Vaswani et al., 2017), has been used to answer interesting questions in the fields of linguistics and psycholinguistics, shedding light on issues related to language processing and acquisition (refer to Futrell and Mahowald (2025) for a more detailed discussion). The development of these models provides an opportunity not only to simulate complex linguistic behaviors (Futrell et al., 2018; Qian and Levy, 2019) but also to gain insights into human cog-

nition and natural language (Millière, 2024a,b; Piantadosi, 2023). Piantadosi (2023) boldly claims that large language models can replace traditional theories of human language, though this idea has caused debate (Kodner et al., 2023).

One line of discussion concerns language acquisition and innate linguistic knowledge, where experiments with LMs can provide evidence for or against the existence of innate linguistic constraints. Wilcox et al. (2024a), for example, provide evidence against the Poverty of the Stimulus theory (Chomsky, 1980) by showing that *wh*-movement can be adequately learned by LLMs. Contreras Kallens et al. (2023) also claim that grammatical knowledge can be learned without innate linguistic knowledge. However, Lan et al. (2024) provide a contrasting viewpoint by showing that LMs fail to learn *wh*-movement and filler-gap contingencies when trained on a corpus size comparable to the linguistic input that children receive. Frank (2023) further argue that the amount of training data LMs require is vastly greater than what children are exposed to, supporting the view that innate knowledge and multi-modal information are crucial for language learning.

LMs have also become valuable tools in the study of sentence processing patterns. One important line of research treats them as psycholinguistic subjects—systems that can simulate human-like responses to specific sentence properties (Futrell et al., 2019; Qian and Levy, 2019). This perspective is motivated by the fact that recent neural network-based language models, trained on vast amounts of linguistic data, can generate naturalistic text that closely mirrors human linguistic behavior. As a result, they provide a powerful experimental tool for investigating a wide range of (psycho)linguistic phenomena (Kallini et al., 2024; Hu et al., 2020).

These attempts have been largely successful, as many studies show that LMs' responses — measured in terms of negative log probability, or surprisal (Hale, 2001; Levy, 2008)—to natural language sentences are human-like (Wilcox et al., 2018). However, this approach should be taken with caution, as LMs often fail to explain the magnitude of language processing difficulty (Huang et al., 2024; Van Schijndel and Linzen, 2021).

Another important line of studies considers LMs as cognitive models that go beyond simple language modeling to offer insights into broader cognitive abilities such as reasoning. Trott et al. (2023), for example, provide evidence that GPT models can perform false-belief tasks, although their performance remains below that of humans. In contrast, Mahowald et al. (2024) claim that while LLMs may be useful for investigating formal linguistic competence, they show clear limitations when it comes to modeling functional linguistic competence.

Among the many aspects where neural network-based LMs can contribute to psycholinguistics, this thesis specifically focuses on interpreting Transformer-based models (Vaswani et al., 2017) as providing algorithmic explanations for sentence processing, based on the functional similarity between their core mechanism —*attention*— and psycholinguistic computational architectures that explain sentence processing patterns. Furthermore, this thesis shows the Transformer-based models can integrate two key accounts of sentence processing: memory-based and expectation-based accounts. In the remainder of this chapter, I provide explanations of these two psycholinguistic accounts of sentence processing, review previous attempts to integrate them, and propose Transformer-based language models (Vaswani et al., 2017) as a novel integrative theory for human sentence processing.

2.2 Memory-based Accounts of Human Sentence Processing

Memory-based accounts for explaining human sentence processing have focused on the fact that human sentence processing is subserved by a bounded linguistic working memory (Caplan and Waters, 1990; Carpenter, 1989; Chomsky and Miller, 1968; Gibson, 1991; Gordon et al., 2001; Lewis, 1993; Marcus, 1979; McElree et al., 2003; Miller and Isard, 1964). According to these theories, the difficulty of processing a sentence is found to link to the demands placed on this memory system. Specifically, the longer the memory system has to work to maintain processed linguistic information until dependencies can be resolved, the more

difficult the sentence is to process. This difficulty also arises from the need to retrieve the relevant dependent information from competing linguistic representations in order to form dependency relationships.

Dependency Locality Theory (DLT) (Gibson, 1998; Gibson et al., 2000) posits that the difficulty of sentence processing is influenced by two memory-related cognitive costs: integration cost and memory cost. The integration cost refers to the effort involved in integrating new linguistic information with the existing memory representation to form dependency relation, and memory cost involves the cognitive load of storing parts of the input that may be needed for later stages of parsing. More broadly, the theory posits that the distance between dependents—measured either in terms of the number of words or the number of discourse referents—affects the ease or difficulty of forming dependencies.

In cue-based retrieval theory (Lewis and Vasishth, 2005; Lewis et al., 2006), dependency formation is bounded by similarity-based interference (Chomsky, 1965; Lewis, 1996, 1998)—retrieving previous partial representations becomes more difficult when there is interference from similar candidate representations. In this view, linguistic working memory operates under principles governing all other kinds of human memory: the limiting factor is the capacity to discriminate target memoranda from similar distractors (Lewis, 1996; Shiffrin, 2003).

2.3 Expectation-based Accounts of Human Sentence Processing

Expectation-based accounts of human sentence processing propose that comprehenders make continuous, moment-to-moment predictions about upcoming words, and that perception and comprehension are strongly influenced by these probabilistic expectations. Accordingly, patterns in human sentence processing can be partly explained by the predictability of words given the preceding context.

Predictability has been defined and measured in several ways, each addressing distinct but complementary aspects. Ehrlich and Rayner (1981) operationalized predictability through contextual constraints, demonstrating that words are more likely to be skipped during reading when the context highly constrains their predictability. Another measure of predictability is cloze probability (Taylor, 1953), which involves asking participants to fill in the blank in a sentence with the most likely word. However, those methods are hard to use to measure word predictability at large scale because constructing experimental sentences and collecting data requires a lot of resources. Optimal Bayesian perception (Lewis et al., 2013; Norris, 2009) explains predictability as the strength or clarity of prior knowledge that shapes readers' expectations. Within this framework, the more obvious the linguistic context, the higher the predictability—and the less additional information needs to be accumulated.

With the availability of linguistic theories and computational language models that can estimate the probability distribution of words given context, more formal and generalizable estimations of predictability have been introduced on the basis of Shannon (1948)'s information theory. Surprisal (Hale, 2001; Levy, 2008) measures the predictability of a word as the negative log probability of it given the preceding context as defined in Equation 2.1 (see Hale (2016) for a tutorial on information theoretic metrics in psycholinguistics).

$$\text{Surprisal}(w) = -\log_2 P(w|context) \tag{2.1}$$

Hale (2001) and Levy (2008) have shown that surprisal holds promise for accounting for garden path effects (though precise quantitative tests reveal that surprisal may underestimate sentence processing difficulties; Van Schijndel and Linzen (2021)). Subsequent work has demonstrated that word reading times vary linearly with surprisal (equivalently, they are a logarithmic function of probability), and that this relationship holds over a range of several orders of magnitude—a kind of quantitative relationship that is rare in psychology.

Another information theoretic approach to explaining sentence processing with pre-

dictability is Entropy Reduction Hypothesis (ERH) (Hale, 2006, 2016). Entropy in ERH is a quantity measuring uncertainty about sentence interpretation[1]. ERH posits that information processing work is done when the sentence processor reduces this uncertainty and that this uncertainty reduction work takes time that should be measurable in word-by-word empirical measures.

Prediction-based estimations have extraordinary empirical scope in accounting for sentence processing phenomena. Hale (2001) and Levy (2008) have shown that surprisal holds promise for accounting for a wide range of garden path effects (though precise quantitative tests reveal that surprisal may underestimate the magnitude of some garden path effects, Van Schijndel and Linzen (2021)). Subsequent work Smith and Levy (2013) has demonstrated that word reading times vary linearly with surprisal (equivalently, they are a logarithmic function of probability), and that this relationship holds over a range of several orders of magnitude—a kind of quantitative relationship that is rare in psychology.

Despite the success of the expectation-based accounts (more specifically the surprisal theory) in explaining sentence processing patterns, two considerations motivate including bounded memory in an integrated model, one empirical and one theoretical. The empirical motivation is that surprisal does not provide a complete account of comprehension difficulty associated with complex syntactic structures. A specific instance is that surprisal does not account for the difficulty at the embedded verb in an English object relative clauses, compared to an English subject relative (Levy and Gibson, 2013). Object relative are less frequent but the higher surprisal comes at the onset of the relative clause, not the embedded verb (though there is some evidence for slowing at the onset NP as well; Staub (2010)). This single construction is interesting in part because it is just one of a class of difficult embedded structures that have motivated memory-based accounts (Futrell et al., 2020; Gibson, 1998; Lewis and Vasishth, 2005). Recent analyses of data obtained from participants

[1]Please note that the entropy of attention entropy introduced in the following section is a quantitative measure of diffuseness of the attentional pattern; there is no commitment in the theory that it characterizes uncertainty about what to attend to or that it characterizes uncertainty about sentence interpretation.

reading natural stories or other content provide further evidence for a contribution of structural complexity independent of surprisal. Word-by-word metrics motivated by Dependency Locality Theory and a left-corner retrieval parser have been shown to account for reading times independent of surprisal (Shain et al., 2016), and structural complexity metrics have been shown to account for variance in fMRI (Brennan et al., 2020) and EEG signals that is independent of surprisal (Hale et al., 2022).

2.4 Previous Attempts to Integrate Expectation-based and Memory-based Accounts

Expectation-based and memory-based estimates of processing difficulties address different aspects in sentence comprehension. In order to successfully combine expectation-based and memory-based theories of sentence processing, two working-memory-related costs need to be considered, which are defined as *memory cost* and *integration cost* as described in Gibson (1998). First, memory cost needs to be considered as human working memory capacity is limited and thus storing the representation of linguistic information is burdensome as new information is read as sentence processing unfolds. However, the original version of surprisal theory (Hale, 2001) does not take this into account and the representation of the preceding context is assumed to remain intact without incorporating the property of memory degradation.

Second, integration cost needs to be considered as well as incremental sentence reading requires dependency relation formation among words stored in memory in order to represent sentential meaning from the sequence of the words. This means that the presence of competing distractors can cause retrieval interference, affecting sentence processing difficulty. However, surprisal theory does not itself provide an account that specifies the computational architecture of sentence processing (Hale et al., 2022; Lewis, 2000)—the memories, representations, processes, and control structure of comprehension. In other words, regardless of

whether surprisal provides a complete account of reading times, there is still the theoretical task of uncovering the computational mechanisms of building sentential representation by forming the dependency relation among words.

There have been recent works pursuing the goal of integrating expectation-based and memory-based accounts addressing the *memory cost* in sentence processing. For example, the lossy context surprisal theory (Futrell et al., 2020) adds noise to contextual memory when computing surprisal, and shows that the surprisals computed with this assumption of noisy context better predict human sentence reading time. Also, Kuribayashi et al. (2022) showed that neural network language models better estimate human sentence reading patterns with limited context access, reflecting the limitation in memory storage. The resource rational surprisal model of Hahn et al. (2022) combines lossy context surprisal with the assumption that comprehenders make rational (given distributional facts about language) inferences given noisy memory representations. In each of these recent models, reading times are influenced by noisy memory through the surprisal bottleneck, that is, surprisal remains the key quantitative predictor but it is a function of a noisy memory of past input.

There also have been attempts to address the *integration cost* to combine expectation-based and memory-based theories. Demberg and Keller (2008) provides empirical evidence that shows surprisals and dependency locality effects explain distinct variation in sentence reading times by analyzing eye-tracking data by computing the integration cost in Dependency Locality Theory. However, as DLT only concerns the discourse referents in computing the cost, the integration cost is only assigned to nouns and verbs, constraining the generalizability in the accounts. Futrell et al. (2020) makes a link between Lossy-Surprisal theory and the Dependency Locality Theory (Gibson et al., 2000) that discusses memory retrieval process in sentence reading in that the theory shows processing difficulties increase as two dependents are far apart, and the difficulty gets even worse if those words are syntactic dependents. However, such a link is not yet strong as it still does not provide accounts for the memory retrieval interference effects. In particular, it does not provide explanation how

interference can be caused by the presence of other discourse referents that do not participate in dependency, which plays a large part in computing integration costs.

Psycholinguistically Motivated Tree-Adjoining Grammar (PMTAG) (Demberg and Keller, 2009) is another example that attempted to combine surprisal and dependency-locality theories by extending Tree-Adjoining Grammar (TAG). It showed the promise in providing integrative accounts for processing certain types of sentences. Despite its successful integration, PMTAG has a limitation in providing comprehensive accounts for sentence processing. In particular, as TAG involves defining the elementary trees and the operations (substitution and adjunction) that guide how these trees can be combined, it can become extremely unwieldy and computationally expensive.

As a novel approach to combining the two accounts of sentence processing, this thesis aims to interpret Transformer language models (Vaswani et al., 2017) as predictive cue-based retrieval models. This approach is motivated by two key aspects of Transformers: their ability to predict words based on preceding sentential context, which enables the model to build expectations, and their attention mechanism, which functionally resembles the cue-based retrieval theory (Lewis and Vasishth, 2005; Lewis et al., 2006)—a framework that explains memory retrieval interference effects in sentence processing. In what follows, I provide a detailed explanation of how Transformer models integrate expectation-based and memory-based theories. In particular, I introduce a novel metric, *attention entropy*, which captures sentence processing difficulty from the perspective of memory integration, leveraging the functional similarity between Transformers' attention mechanism and cue-based retrieval parsers.

CHAPTER 3

Transformers as Predictive Cue-based Retrieval Models

This section will establish the conceptual link between Transformer's attention mechanism and cue-based retrieval theory. I first review the key aspects of cue-based retrieval theory and current challenges in developing it as a psycholinguistic theory and describe how Transformers help to address those challenges. Then I will formally define the attention entropy metrics that will be used in subsequent analyses, especially using attention heads with grammatical function. I then conclude this section with how the similarity between attention mechanism and cue-based retrieval parsing enables Transformers to function as predictive cue-based retrieval model.

3.1 Cue-based Retrieval Parsing and its Methodological Challenges

Figure 3.1, adapted from Lewis et al. (2006), illustrates the core assumptions of cue-based retrieval parsing (Lewis et al., 2006; Lewis and Vasishth, 2005; Van Dyke and Lewis, 2003; Lissón et al., 2021, 2023): (a1) word-by-word incremental formation of linguistic relations is mediated by the retrieval from memory of relevant prior linguistic representations; (a2) the retrieval process happens via a parallel match of cues against all candidate memory items; (a3) the cues are derived from the current word being processed via the application of

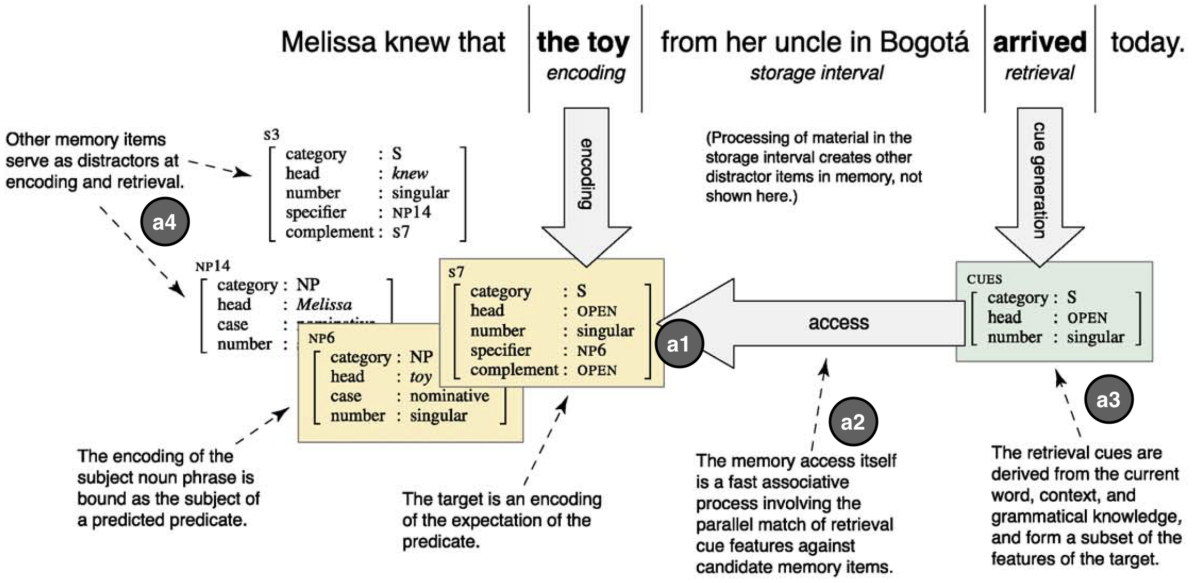


Figure 3.1: Cue-based retrieval parsing, adapted from Lewis et al. (2006). The labels a1–a4 refer to assumptions enumerated in the text.

proceduralized knowledge, which includes grammatical knowledge; and (a4) similarity-based interference arises to the degree that cues partially match distractor items in memory. This interference may lead to longer retrieval times of the correct target. It can also lead to increased probability of retrieving an incorrect distractor, which in the case of grammatical illusions can lead to a mixed distribution of correct and incorrect retrievals, and retrieval times with a mean that is faster than non-interfering baselines.

This broad sketch leaves unspecified many key details that are required to make empirical predictions. These include commitments to basic parsing strategies, whether single or multiple interpretations are maintained during processing, the nature of the intermediate memory representations, and the possible features that may be used as cues. For example, the ACT-R model of Lewis and Vasishth (2005) used a left-corner parser and syntactic representations and cues modeled after X-bar phrase structure trees (Baker, 1995; Chomsky, 1970).

The requirement that cue-based parsing models must commit to linguistic representations and specific cues is both a methodological virtue and challenge. One benefit is that it

allows for empirical tests of different assumptions about cues (e.g., Yadav et al. 2022). The challenge is that the application to arbitrary linguistic stimuli requires a broad-coverage cue-based parser, and there have been few attempts to create one. A related challenge is that it has not been clear how to model a limited beam multi-path parser that would presumably require multiple memory retrievals at each word. Pre-trained Transformers provide one way to address all of these challenges: they are broad coverage language models that may be applied to any linguistic stimuli, the representations and cues (queries) are learned, and the model provides a concrete mechanism that supports multiple parallel retrievals and, at least implicitly, multi-path parsing.

3.2 Memory Retrieval in the Transformer: Scaled Dot-product Attention

The function of attention in the Transformer is to create dynamic pathways from representations of earlier parts of the sentence to the computations that create representations for the current word or token. Explicit memory retrievals open up these pathways through simple computations localized in each *attention head*. There are multiple attention heads operating at multiple layers of abstract representation. I describe here the computations local to a single attention head and layer[1].

In each attention head, the input embeddings of words in a sentence are converted into *keys*, *queries*, and *values*[2], which are real-numbered vectors. *Queries* serve as the cue in cue-based retrieval parsing, and *values* are the representations associated with prior words. Consider reading the sixth word (*were*) of a sentence, as in Figure 3.2: there are five value vectors for the previous words (denoted $v_1, v_2, ...v_5$ in Figure 3.2), excluding the value vector

[1]Recent work by Timkey and Linzen (2023) shows that interference effects arise in a recurrent neural net with only one attention head.

[2]This is unfortunate terminology when combining Transformers with value-based decision making. Value vectors in Transformers do not represent *value* in the decision theoretic or reinforcement learning sense.

for the sixth word itself.

At this point, if the Transformer were a direct implementation of the simplest kind of cue-based retrieval model, it would be natural to assume that q_6 is matched against $v_1 ... v_5$, and the closest matching value vector is retrieved. If this were the case, attention would be a *content-based retrieval* because the cues would be matched directly against the content of candidate retrieval items. But there are two important twists here that make the Transformer different.

The first twist is that the query vectors are not matched directly against the value vectors. Instead, they are matched against *key* vectors of the same size that are associated with prior words (denoted $k_1 ... k_5$ in Figure 3.2). The match produces a positive real scalar quantity that represents the degree of match, and it is computed by simply taking the dot-product of the query and key vectors and dividing by a scaling constant that is a function of the vector size; Vaswani et al. (2017) call this 'scaled dot-product attention'.

The second twist is the following. Rather than retrieving the value vector associated with the maximum score, a softmax computation transforms the set of match scores into a vector of quantities between 0 and 1 that sum to one, and these quantities are used to compute a *weighted sum of the value vectors*. This weighted sum of value vectors is the output of the attention process. These retrieved weighted-sum of values are then passed through further computations including a multi-layer perceptron to compute new vector representations as output. The weighted sum retrieval and the use of separate key vectors is a generalization of cue-based retrieval; it subsumes the simple case where the key vectors are just equal to the value vectors and where a maximum is taken to retrieve a single value. Formally, if all the query vectors are combined as columns of a matrix Q, and keys and values combined into matrices K and V, the output of attention is a matrix of output vectors defined by:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}V\right) \tag{3.1}$$

, where d_k is the dimensionality of the key and query vectors and QK^T accomplishes the dot products of all the query and key vectors. The focus here is solely on 'causal' attention, where attention at a given word position is directed only to preceding words and not to future ones; this approach has become the standard paradigm in generative AI.

Most Transformer language models are built with multiple layers. In each layer, multiple attention heads perform multiple scaled dot-product attention retrievals and the outputs of each head are concatenated. These concatenated outputs are then projected to vectors that form the input to a subsequent layer of multi-headed attention, until finally a probability distribution over next-words is computed which is the final output of the Transformer at each word. For example, GPT2-small that will be used in the thesis, there are 12 layers each with 12 independent attention heads. There are thus two distinct kinds of parallelism: each attention head can make a retrieval that combines multiple items in parallel, and there are multiple attention heads operating in parallel at each layer.

What is remarkable about the Transformer is that it is possible to train the architecture end-to-end so that the word embeddings, keys, values, query representations, and softmax parameters are all learned from the task of predict-the-next-token. (And with sufficient data, yields multi-modal models that form the basis of systems like ChatGPT).

3.3 Attention as Cue-based Retrieval Subject to Interference

The mapping of Transformer attention to cue-based retrieval is clear: all word-by-word computations in the Transformer are mediated by retrieval from prior representations (assumption (a1), above); attention happens via a parallel match of queries against all candidates for retrieval (a2); and queries are computed from representations of each word (a3). What remains to be established is that attention patterns show evidence of interference from distractor items that are not in grammatical relations with the current word (a4).

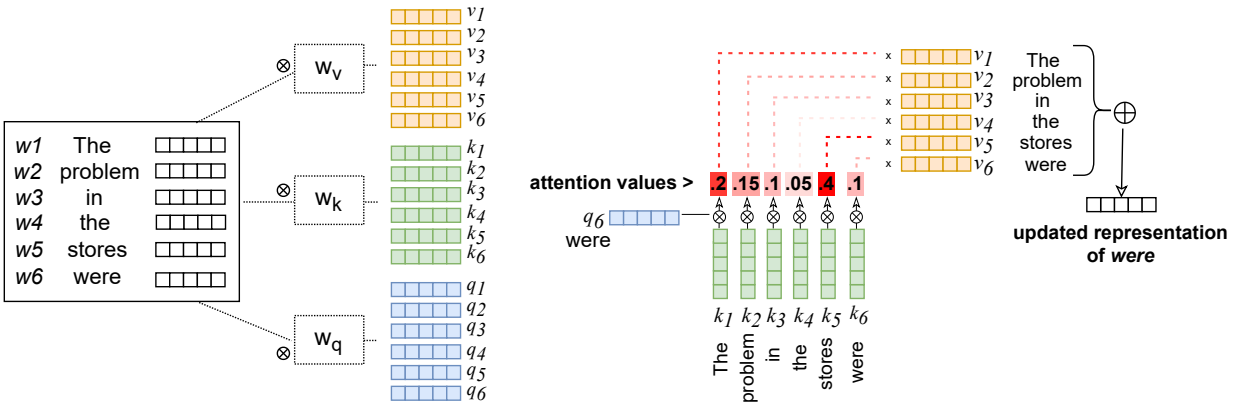
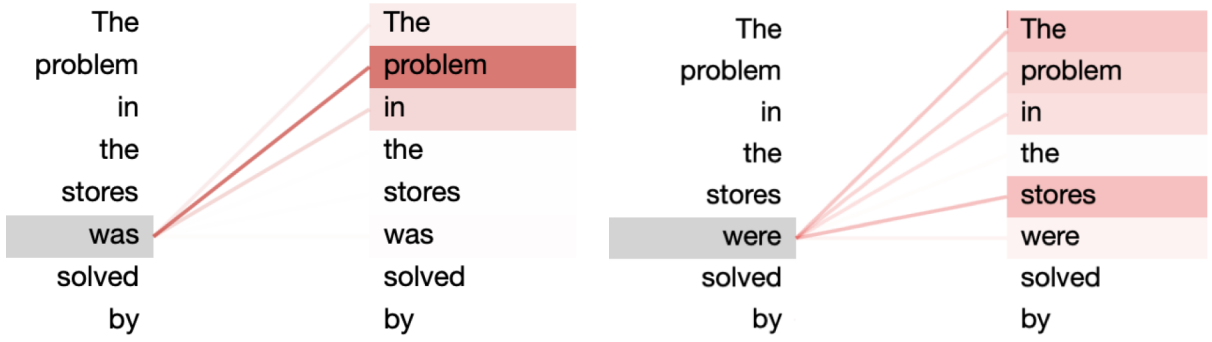


Figure 3.2: How attention weights between a word (*were*) and representations associated with preceding words are computed in a Transformer's self-attention head. *Left:* Input embeddings of each word are multiplied by learned weight matrices and converted into value, key, and query vectors. *Right:* The query vector of *were* is matched against key vectors of preceding words. The value vectors of preceding words are summed using the attention weights to contribute to a contextual representation of *were* .

The remainder of the paper provides extensive evidence for this, but I provide one simple illustrative example of interference explained with attention here.

Consider the example sentences in (1) involving subject-verb agreement in English. The attention patterns of interest are those observed at the verb (*was* in sentence (1a) and *were* in sentence (1b)). In both cases the preceding words are identical, and the head noun of the subject phrase is *problem*, although there is an agreement violation in (1b). Prior work has shown that some attention heads in Transformer-based language models are specialized for specific grammar dependencies such as subject-verb (Vig and Belinkov, 2019; Voita et al., 2019). It is therefore expected that if an attention head is found to be specialized for processing subject-verb relations, this attention head would show relatively focused attention on the head noun *problem* in (1a). But under the assumption that the plural number feature is part of the retrieval query, there should be additional attention weight given to the distractor *stores* in (1b).



(a) Non-interfering (b) Interfering

Figure 3.3: Attention distributions at the critical verb (*was* and *were*) in the self-attention head that is specialized for building subject-verb relations. The distribution is more diffuse in the interfering condition because of attraction to the distractor (*stores*) that matches the number feature of *were*.

(1) Agreement interfering examples.

a. NON-INTERFERING

The **problem** in the stores **was** solved by

b. INTERFERING

The **problem** in the stores **were** solved by

This is exactly the pattern observed in Figure 3.3, which shows the attention weights computed by the attention head in GPT2-small that is known to be responsible for building subject-verb relations, using a visualization tool that Vig (2019) introduced[3]. Section 3.5.1 provides the detailed procedure to identify attention heads with the ability to process a certain type of grammar dependency.) The attention in (1a) is relatively focused and the attention in (1b) is more diffuse, because of the additional weight given to the distractor NP. I will provide a more in-depth illustration of similarity-based interference in agreement phenomena in Section 6.1.

[3]The specific head is the third head in the fourth layer; henceforth referred to as $\text{head}_{4,3}$ this head was identified using the method introduced in Voita et al. (2019)

3.4 Attention Entropy as a Word-by-word Processing Metric

The key idea in the thesis is that diffuse attention patterns in pre-trained Transformers are a signature of similarity-based retrieval interference in memory integration. To quantify the interference effects at each word, I use *attention entropy*[4] metric.

Recall that at each word w_i, each attention head computes a softmax vector of attention weights allocated to each previous word position. I denote the attention weight from a source word at position i to a target word at position $j < i$ by attention head h as

$$\text{Attn}_h(w_i, w_j) \tag{3.2}$$

which must be between 0 and 1.

To quantify the diffuseness of the attention from a source word i at a given attention head, I use Shannon (1948)'s information entropy, which has the properties that we need: it is well-defined over weights that sum to 1, it is at maximum when the attention weights are equal, and minimum when all the weight is on one element. Formally, the attention entropy at word i for attention head H is:

$$\text{AttnEntr}_h(w_i) = -\sum_{j=1}^{i-1} \text{Attn}_h(w_i, w_j) \times \log_2 \text{Attn}_h(w_i, w_j) \tag{3.3}$$

, where i refers to the location of the current source word, and the j's are locations of prior words.

The attention entropy can be measured either using a single attention head (for example one that is known to be specialized for finding a specific grammatical dependency; *local*

[4]Note that the entropy used for attention entropy is different from entropy in Hale (2006)'s Entropy Reduction Hypothesis. *Attention entropy* is a metric designed to estimate 'retrieval interference' while *entropy reduction* estimates the changes of uncertainty in *prediction*. There could be possible interesting connections among these quantities which needs to be studied further.

attention entropy henceforth), or a *aggregate attention entropy* that aggregates the attention entropy over multiple attention heads. After selecting a set of attention heads to be used for aggregate attention entropy computation, a simple mean in Equation 3.4 is used as a measure of aggregate attention, defined as:

$$\text{AggrAttnEntr}(w_i) = \frac{\sum_{(l,h)\in H} \text{AttnEntr}_{lh}(w_i)}{|H|} \tag{3.4}$$

, where l and h indicate the indices of layers and heads in Transformer model, H is a set of selected attention heads, and AttnEntr_{lh} refers to attention entropy measured at the hth head in the lth layer.

In the next section, I will introduce a method for selecting a set of attention heads to be used for aggregate attention entropies, relying on a previous method (Voita et al., 2019) to find attention heads that capture grammar dependencies.

3.5 Attention Entropies from Grammar Dependency Attention Heads

In the previous section, I introduced two types of attention entropy: local attention entropy and aggregate attention entropy. Both require a selection process. For local attention entropy, a single attention head must be chosen based on its ability to capture grammatical dependencies relevant to the phenomenon under investigation. For example, interference effects in subject-verb agreement can be effectively illustrated using an attention head that consistently attends to subject-verb relations, as briefly described in Section 3.3.

In order to compute aggregate attention entropy, a subset of attention heads must be selected to avoid noise from irrelevant attention heads—those that do not contribute meaningfully to memory retrieval and integration. A straightforward approach to computing aggregate attention entropy might be to compute the mean across all 144 attention heads;

however, this is problematic because many heads exhibit linguistically uninformative attention patterns, such as consistently attending to the first or immediately preceding token. Additionally, some attention heads may not play a meaningful role in the Transformer's final output (Voita et al., 2019; Vig and Belinkov, 2019), introducing noise if included in the aggregate attention entropy computation.

In what follows, I describe the selection process of grammatical-dependency attention heads using the method proposed by Voita et al. (2019). This enables the selection of an attention head for local attention entropy computation and also supports the computation of aggregate attention entropy over selected grammar-dependency heads. I define this as Grammar Dependency Attention Entropy (GDAE), and it will be used throughout the following chapters.

3.5.1 Grammar-dependency Attention Heads Selection Process

Voita et al. (2019) presents a method to identify Transformer's attention heads that capture specific grammar dependencies. This method works by comparing the accuracy of attention heads in focusing on syntactic dependents (i.e., paying the highest attention to syntactic dependents) with the proportion of the most frequent relative positions observed between those dependents. By using the relative position frequencies as a baseline, the method ensures that the attention heads with grammar knowledge assign the greatest attention to a particular syntactic dependent not simply due to positional information. The detailed steps are illustrated below.

1. Selection Materials for Analysis Using the Natural Language Toolkit (NLTK; Bird (2006)), I collected 148,376 sentences from the Brown Corpus and the Gutenberg Corpus. Grammar dependencies among words were annotated using the CoreNLP dependency parser (Manning et al., 2014).

2. Assigning grammar dependencies among words

Using CoreNLP (Manning et al., 2014), each word in the sentences was assigned the position of its governor and the type of grammar dependency it forms with the governor. The relative position of the dependents is then computed by taking the difference between the positions of the target word and its governor.

3. Calculating the proportion of the most frequent relative positions of dependents in grammar dependencies

For each grammar dependency, I recorded the relative positions of dependents in relation to their governors across all instances. For example, in a nominal subject-verb (`nsubj`) relation, the most common relative position is -1, meaning that subjects typically appear immediately before its governor (*verb*), which accounts for 42% of all instances.

4. Examine attention heads' accuracies in relating grammar dependents

To determine which attention heads capture certain dependency relations better than simple prediction based on relative position, I computed the percentage of instances for each attention head where the largest attention is paid to the corresponding grammar dependents from their governors, or to the governors from the dependents if the governor precedes the dependents. In the case of the nominal subject and verb (`nsubj`) relation, the highest percentage was found in $\text{head}_{4,3}$ (the third head in the fourth layer of GPT2), where the largest attention is paid to the subject from the verb in 59% of `nsubj` instances.

5. Determination of grammar dependency attention heads

Just as Voita et al. (2019) did, I consider attention heads can reflect a certain grammar dependency if their accuracy in allocating the largest attention to syntactically corresponding dependents is at least 10% higher than the proportion of instances that can be explained by the most frequent relative positions. For example, for subject-verb relations, it must allocate

maximum attention to subjects in at least 45.1% (41+4.1 %) of cases. For some dependency types, multiple attention heads meet this threshold, and I select the one with the highest score.

3.5.2 Identified Grammar Dependency Attention Heads

Among the 43 dependency types supported by the CoreNLP package, 14 types failed to be associated with the patterns of any attention heads. There are twenty attention heads that are selected based on criteria listed above: (0,6), (1,0), (1,1), (2,0), (2,3), (2,8), (2,9), (3,2), (3,5), (3,6), (3,8), (3,9), (4,0), (4,2), (4,3), (4,9), (4,11), (6,7), (7,8), and (10,5) where the first and second numbers in each pair indicate the layer and head numbers, respectively. The twenty selected attention heads are graphically represented in Figure 3.4, and the full results are provided in Appendix A.1.

A single grammar dependency attention head can be selected for computing local attention entropies based on the specific research question. For example, if the goal is to analyze interference effects in subject-verb agreement, $\text{head}_{4,3}$ can be chosen, as it achieves the highest score among attention heads identified as detecting `nsubj` dependencies, such as $\text{head}_{6,0}$, $\text{head}_{3,6}$, and $\text{head}_{2,9}$.

To compute aggregate attention entropy from grammar-dependency attention heads (or Grammar Dependency Attention Entropy; GDAE), which reflects general interference effects rather than those arising from a single grammatical dependency, I take the average of attention entropies from the twenty selected attention heads mentioned above.

3.5.3 Limitation of Grammar Dependency Attention Entropies

One limitation of the grammar-dependency-based method for computing aggregate attention entropies is that the method focuses exclusively on heads that encode grammatical dependencies. Although this syntactic focus aligns with many established psycholinguistic effects, it is notable that memory-based retrieval interference can also influence semantic processing

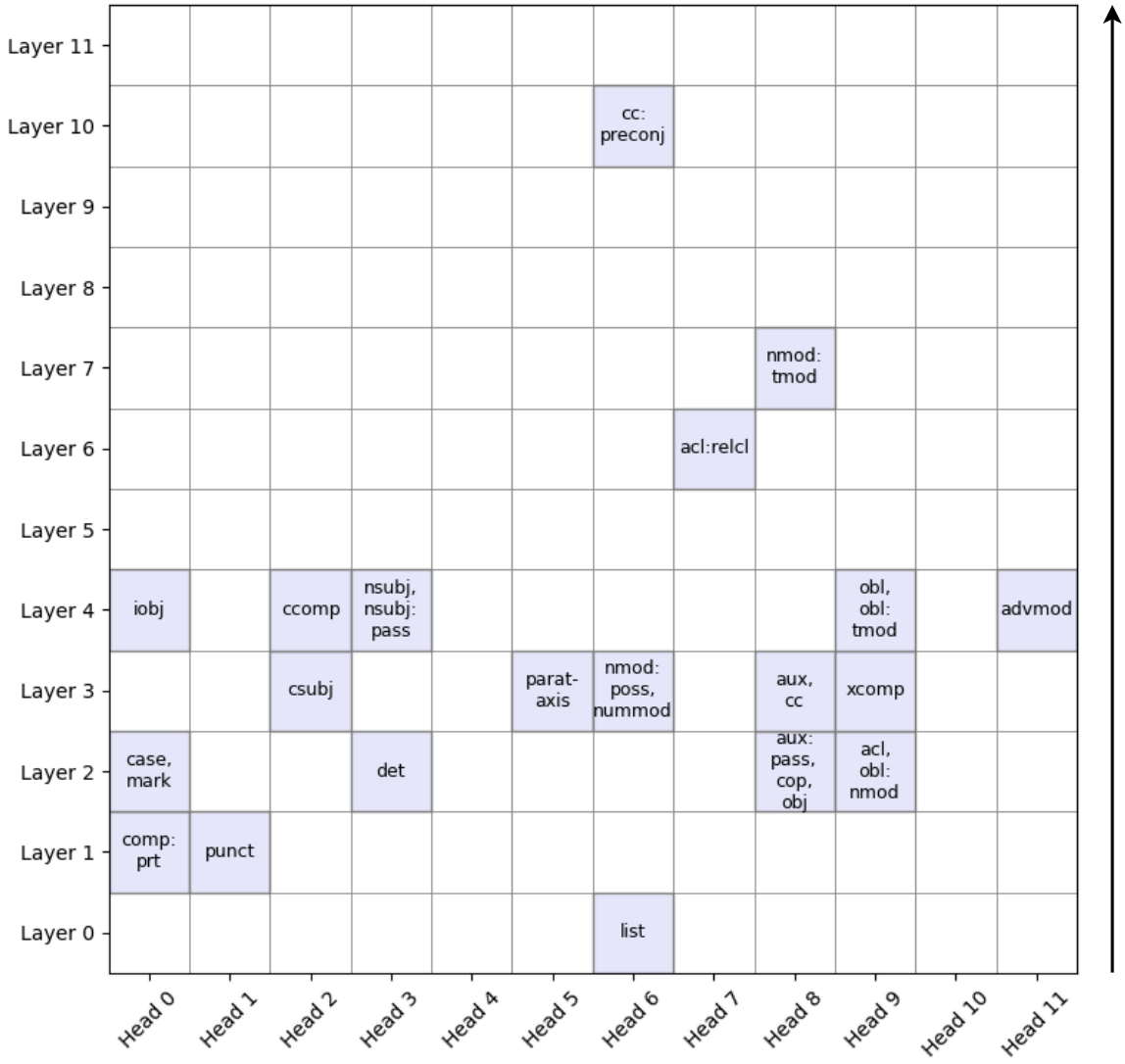


Figure 3.4: Twenty selected attention heads based on their capacity to capture grammatical dependencies. The Y-axis represents the layer positions, while the X-axis shows the position of heads within each layer. The arrow indicates the processing direction.

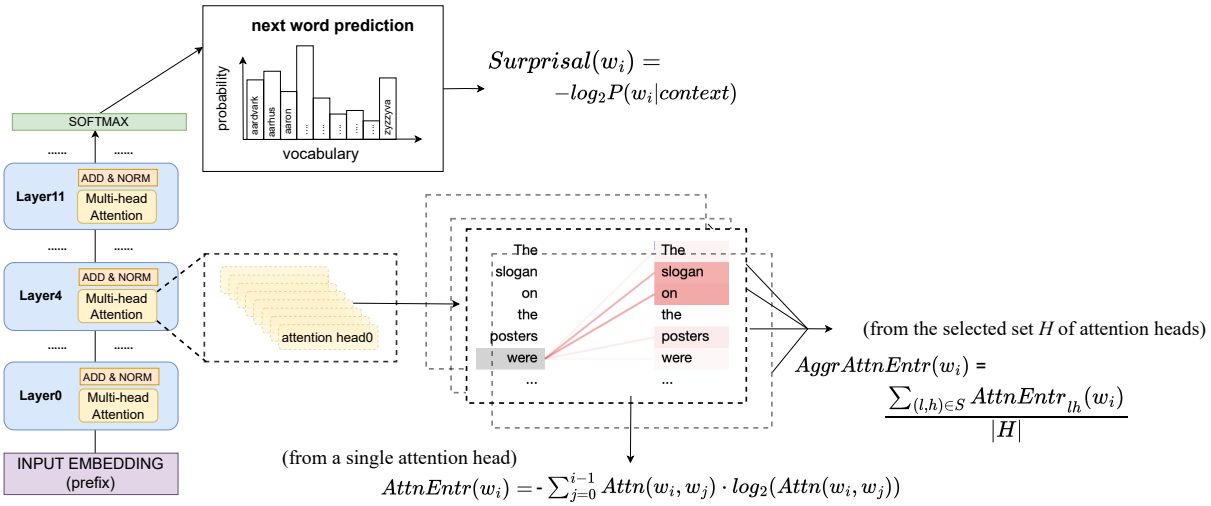


Figure 3.5: The internal working mechanisms of GPT2-small, showing how the Transformer architecture integrates expectation-based and memory-based aspects of sentence processing. Surprisals are computed from the last hidden state of the model, and attention entropies are computed based on attention distributions in attention heads. Attention entropy is proposed as an index of similarity-based retrieval interference (Ryu and Lewis, 2021)

and interpretation (Cunnings and Sturt, 2018). As such, attention heads that contribute to semantic coherence, discourse integration, or other non-syntactic processes may also be identified and included for aggregate attention entropy computation.

I will return to these broader considerations and propose future directions for attention head selection in Chapter 7.

3.6 Transformers as Integrative Sentence Processing Models

Figure 3.5 provides an overview of how Transformer-based language models (specifically GPT2-small) operate with predictive cue-based retrieval processing. As Transformers are a type of neural network language model that estimates a probability distribution over words given a prefix, surprisals can be easily computed from their final layer representations. The final representations are built through attention mechanism that incorporates a process

that behaves like cue-based retrieval model. Using attention entropies, Transformer models can also estimate sentence processing difficulty from the perspective of memory retrieval interference.

The integrative properties that combine expectation-based and memory-based perspectives are shared in many other Transformer-based Language Models with causal attention heads. I chose to use the GPT2-small model for two reasons. First, small models make analysis more widely accessible and future interpretability analyses more tractable. Second, Oh et al. (2022) found that the GPT2-small model is the most predictive of human reading times among a set of larger models. There have been more recent and complex models after Oh et al. (2022), but it is not likely that the recent models will provide the basis for better predictions of human reading times despite their higher performance in downstream tasks. In short, larger size and longer training do not make Transformer-based models more predictive of human reading times (Oh and Schuler, 2023b,a; Wilcox et al., 2024b).

In the upcoming chapters, I will present evidence showing that attention entropies offer an additional explanation of naturalistic sentence reading, as well as psycholinguistically interesting phenomena (such as garden-path sentences, agreement interference, and embedded sentences), which surprisals alone cannot account for.

CHAPTER 4

Modeling Sentence Reading Times with Surprisal and Attention Entropy

This chapter aims to examine whether and how word-by-word aggregate attention entropy that I measure with Grammar Dependency Attention Entropy (GDAE) provides additional predictive power beyond surprisal in explaining sentence processing difficulty.

I start with two preliminary analyses: first, I examine the optimal attention window size to be used to compute GDAE for reading time prediction; second, I examine the predictive power of GDAE in comparison with the aggregate attention entropies that are with the entire set of attention heads in GPT2-small and with non grammar dependency attention heads.

I then fit three Bayesian mixed-effect regression models that predict three different types of word-level reading times: (1) self-paced reading times, (2) first-fixation durations from eye-tracking data, and (3) go-past times from eye-tracking data. The models included attention entropy, surprisal, and other psycholinguistic factors (e.g., word frequency, word length, etc.) as predictors. Results from the models show that attention entropy provides additional predictive power beyond surprisal in accounting for sentence processing difficulty in a way that is consistent with the interpretation that attention entropy reflects memory retrieval interference during sentence processing.

4.1 Preliminary Study 1: Finding the Optimal Attention Window Size

Gao and Yu (prep) demonstrated that surprisal becomes more predictive of human reading times when GPT2-small uses a large context window (approximately 1,000 tokens). This result is unsurprising, as a larger context enables readers to draw on broader linguistic information to form more accurate expectations about upcoming words.

To use Grammar Dependency Attention Entropy (GDAE) as an estimator of human reading times, a similar question must be addressed: *How many words ahead should be considered when computing GDAE?* If the influence of GDAE on reading times reflects memory integration, it is likely that a much smaller window than 1,000 tokens will be optimal. This is because memory integration relies on rich lexical and syntactic details that are typically retained only over short contexts. Consequently, integration is most likely to occur within a single sentence or across a few adjacent sentences, rather than over long-distance contexts spanning hundreds of tokens.

To explore this question, I conducted two empirical analyses that aim to examine how the predictive power of attention entropy varies as a function of attention window size. First, I compared the coefficients of GDAE from simple frequentist regression models using the Equation 4.1, which was fit with attention entropies whose attention window sizes vary.

$$\begin{aligned} ReadingTimes \sim lm(Surprisal + GDAE\_N + Frequency + WordLength + \\ SpilloverSurprisal + SpilloverGDAE\_N + SpilloverFrequency + \\ SpilloverWordLength + Position) \end{aligned} \tag{4.1}$$

, where N refers to the size of the attention window used to compute GDAEs.

The token representations were derived using the maximum context window size, and attention values for the most recent N tokens were included for attention entropy with window size N. These models included surprisals and other standard control variables such as word length, word position, and word frequency. To account for spillover effects, predictors derived from the immediately preceding word were also included. I used the Natural Stories corpus (Futrell et al., 2021) that includes word-level reading times collected using a self-paced reading paradigm. For the details of the Natural Stories corpus and the preprocessing steps, see 4.3 and 4.4 respectively.

In the second analysis, I evaluated how much additional variance in reading times could be explained by attention entropy computed with different window sizes, compared to attention entropy computed only over the sentence-level context (i.e., considering attention only to words within the same sentence). I fit a baseline model where attention entropies are computed at the sentence level (i.e., the context representation is computed only with the prefix of a sentence and attentions are paid only to the previous words in the same sentence), and compared log-likelihood differences between the baseline models and models fit with attention entropies computed from fixed number of preceding tokens (i.e., fixed attention window size) regardless of sentence border. The models were fit also using the Equation 4.1.

The results from the first and second analyses are in Figure 4.1 and Figure 4.2, respectively. Figure 4.1 shows GDAE has the highest coefficient when it is computed with an attention window size of 30. Also, Figure 4.2 shows that the log-likelihood is the highest when the GDAE is computed with a window size of 30. These two results together suggest that attention entropy benefits from a relatively limited window size (30) compared to the large optimal context window for surprisal.

A window size of 30 means that the entropy metric focuses on attentional patterns primarily within the current sentence and the preceding sentence — aligning with the selection of attention heads based on intra-sentential grammatical dependencies.

The shorter optimal window size for attention entropy, in contrast to the much longer

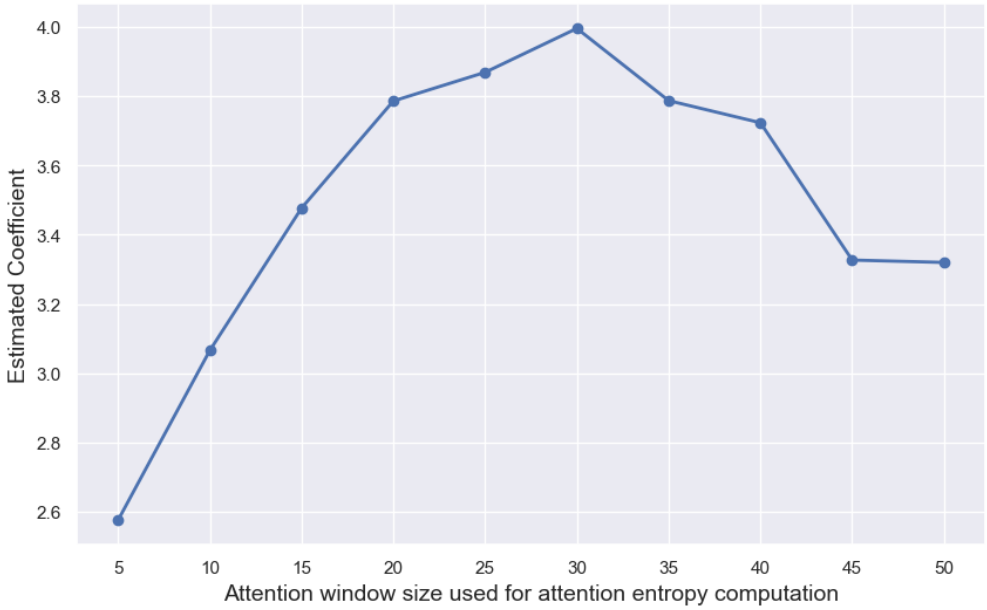


Figure 4.1: Change in coefficients of attention entropies by the different context window size used for attention entropy computation.

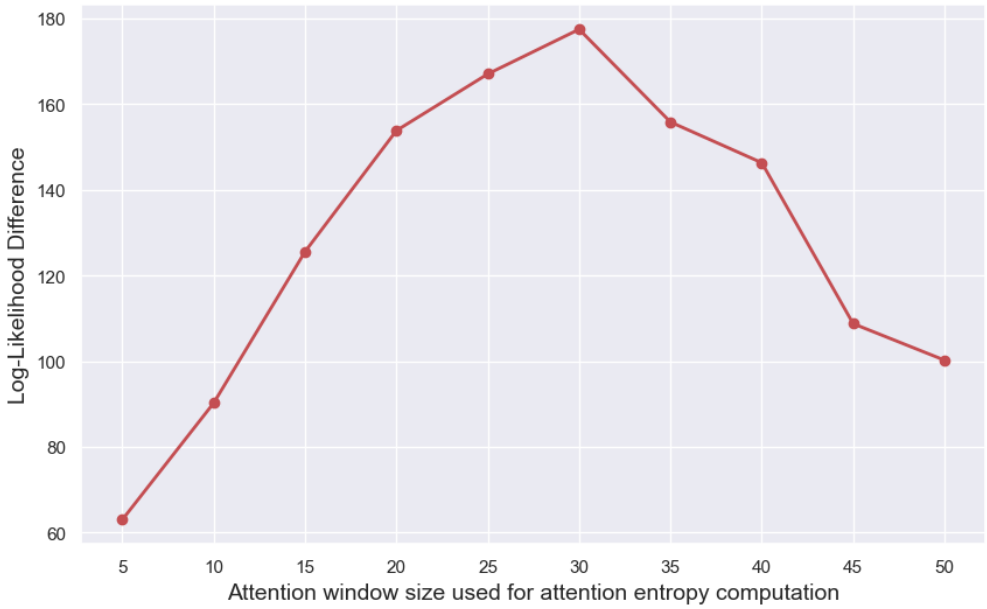


Figure 4.2: Change in the log-likelihood difference between the base model and the model with attention entropies.

window size required for surprisal, supports the interpretation of their respective roles in sentence processing. Attention entropy reflects memory interference, which relies on fine-grained linguistic features that are likely to be held in short-term memory. Surprisal, on the other hand, captures expectation-based processing, which draws on broader contextual information and can extend across longer spans of discourse, as it does not depend on detailed linguistic representations[1].

Based on the findings from this preliminary study, I use an attention window size of 30 to compute Grammar Dependency Attention Entropies (GDAE) used for sentence reading times modeling in this chapter.

4.2 Preliminary Study 2: Predictive Power of Grammar Dependency Attention Entropy

Prior to modeling naturalistic sentence reading times with GDAE, I assess whether GDAE increases predictive power in predicting human reading times compared to the aggregate attention entropies that are computed using all the attention heads in GPT2-small model. To this end, I computed three types of aggregate attention entropies: (a) attention entropy using the 20 selected grammar-dependency attention heads, which is GDAE, (b) attention entropy computed using the 122 non-selected attention heads, and (c) attention entropy using all 144 attention heads as a baseline. I then fit three simple frequentist regression models incorporating these different attention entropy measures, surprisal, and other predictors (e.g., word frequency, word length, word position in the sentence), using the model structure in Equation 4.2. The same Natural Stories Corpus (Futrell et al., 2021) is used as in the previous

[1]While I identified 30 as the optimal attention window in a simplified experiment, I believe that the true optimal window size may depend on various factors. Several open questions remain to be answered. For instance, the ideal window size may vary depending on the type of grammar dependency a word forms or the word position in the sentence. Additionally, different reading time measures—such as self paced reading times, first fixation duration or go-past time—may benefit from different attention window sizes when modeling processing difficulty.

section. Attention window size of 30 was used for aggregate attention entropies and context window size of 1000 was used for surprisal. I then compared the log-likelihood differences between (a) and (c), and between (b) and (c), to evaluate the relative contributions of the selected grammar dependency attention heads.

$$\begin{aligned} ReadingTimes \sim lm(Surprisal + AggrEntropy + Frequency + WordLength + \\ SpilloverSurprisal + SpilloverAggrEntropy + SpilloverFrequency + \\ SpilloverWordLength + Position) \end{aligned} \tag{4.2}$$

The results showed that log-likelihood difference between (a) and (c) was 117.29, whereas the difference between (b) and (c) was –11.73. The results indicate that GDAE has the predictive power for explaining human sentence reading times, compared to the aggregate attention entropies from all 144 attention heads in GPT2-small or from non grammar dependency attention heads.

Based on the results from the two preliminary analyses, GDAEs computed with the attention window size of 30 are used to model naturalistic sentence reading times in the following analysis.

4.3 Three Naturalistic Sentence Reading Times to Model

In order to examine the effects of GDAE in predicting naturalistic sentence reading times in detail, I used three word-level reading time materials. First, I included self-paced reading times from the Natural Stories corpus (Futrell et al., 2021), which consist of reading times for 485 sentences read by 200 participants collected using a self-paced reading paradigm.

Sentences in this corpus were designed to include syntactic constructions that generate psycholinguistically interesting phenomena which show the role of memory in sentence processing more clearly than most commonly-used sentence structures. I also included two types of reading times from eye-tracking data: the first fixation duration, which is the time spent on the first fixation of the current word and the go-past time, which is the total of all fixations before moving to the right of the current word, including any regressions to earlier words. I chose the two different reading times in eye-tracking as they reflect different aspects in sentence reading. Specifically, the first fixation duration indexes only the early stage of word processing while the go-past time also incorporates the later stage of word processing. I used the Ghent Eye tracking corpus (GECO, Cop et al. (2017)) in which 5,031 sentences from novels are read by 14 monolingual English speakers.

4.4 Bayesian Modeling of Sentence Reading Times

For every word in the corpus, both surprisals and GDAEs were computed using GPT2-small. To take spillover effects into account, predictors from the immediately preceding word are also included. Surprisal for each word was calculated using the largest possible context window size, which is 1,024, and GDAE was calculated with the context window size of 30.

Data points are excluded if a word is read in less than 100 ms or in more than 3,000 ms from analyses in both reading time data. When GPT2's byte-pair encoding (BPE) tokenizer recognizes a word as a combination of multiple tokens, I took the maximum values of surprisal and attention entropy of the subtokens. I chose maximum rather than sum or mean to avoid disproportionately increasing (with sum) or decreasing (with mean) the entropy metric for individual low-frequency words which happened to be split into subtokens.

A Bayesian linear mixed model was fit with surprisal and GDAE for the current and previous words and many other control variables including position, word length and word frequency computed from Google N-gram corpus (Michel et al., 2011) (see Tables 4.1 and

4.2). I included participant IDs as a random slope and word types as a random intercept. All predictors were transformed into z-scores except for the word position. The model was fit using the R package brms Bürkner (2017). Uninformative priors were specified as follows: The intercept prior assumes a normal distribution centered at 1000 with a standard deviation of 1000; for the regression coefficients (b), I assume a normal distribution with a mean of 0 and a standard deviation of 500; The standard deviation (sd) prior also follows a normal distribution with a mean of 0 and a standard deviation of 500; the correlation matrix prior employs an LKJ distribution with a shape parameter of 1. The model specification follows the approach outlined in previous work on reading time modeling by Boyce and Levy (2023).

4.5 Results

Table 4.1: Posterior estimates for the fixed effects of predictors on self-paced word reading times in the Natural Stories Corpus (Futrell et al., 2021).

Parameters	Estimate	95% CrI
intercept	**321.64**	**[310.54, 333.68]**
word position	-0.01	[-0.10, 0.07]
word length	**4.44**	**[2.47, 6.43]**
surprisal	**7.74**	**[6.70, 8.84]**
GDAE	**2.27**	**[1.47, 3.05]**
word frequency	-5.32	[-7.82, 2.57]
surprisal × GDAE	**0.90**	**[0.46, 1.35]**
(previous word) word length	-0.09	[-1.13, 0.87]
(previous word) surprisal	**5.03**	**[4.30, 5.80]**
(previous word) GDAE	0.36	[-0.27, 0.96]
(previous word) word frequency	**-2.33**	**[-3.10, -1.56]**

(previous word) surprisal × GDAE	**0.83**	**[0.44, 1.20]**

Table 4.2: Posterior estimates for the fixed effects of predictors on word reading times in the GECO Corpus (Cop et al., 2017). Predictors with significant effects are in bold.

First Fixation Duration

Parameters	Estimate	95% CrI
intercept	**213.70**	**[194.19, 234.24]**
word position	**-0.11**	**[-0.25, -0.03]**
word length	-0.23	[-1.66, 1.24]
surprisal	**3.29**	**[2.46, 4.12]**
GDAE	-0.98	[-2.40, 0.39]
word frequency	**-4.06**	**[-5.46, -2.68]**
surprisal × GDAE	**0.39**	**[0.01, 0.77]**
(previous word) word length	**-2.77**	**[-3.97, -1.61]**
(previous word) surprisal	**1.75**	**[0.93, 2.57]**
(previous word) GDAE	**2.35**	**[0.86, 3.91]**
(previous word) word frequency	-0.06	[-0.83, 0.71]
(previous word) surprisal × (previous word) GDAE	0.28	[-0.06, 0.63]

Go-past Time

Parameters	Estimate	95% CrI
intercept	**331.23**	**[293.99, 367.93]**
word position	0.38	[-0.14, 0.90]
word length	**16.38**	**[8.62, 24.58]**
surprisal	**9.84**	**[7.45, 12.23]**

GDAE	**3.27**	**[0.06, 6.43]**
word frequency	-2.53	[-6.44, 1.14]
surprisal × GDAE	1.34	[-0.31, 3.02]
(previous word) word length	**-10.16**	**[-13.69, -6.72]**
(previous word) surprisal	**10.95**	**[7.87, 14.01]**
(previous word) GDAE	-3.12	[-6.40, 0.16]
(previous word) word frequency	**-5.11**	**[-8.32, -1.76]**
(previous word) surprisal × GDAE	-1.08	[-2.65, 0.52]

Results from the models are provided in Tables 4.1 and 4.2. The posterior distribution of estimates for predictors of interest — surprisal and attention entropy — are shown in Figure 4.3. Not surprisingly, quite strong surprisal and spillover surprisal effects are found. In addition to that, the results show the attention entropy has an influence on reading times in both self-paced reading and eye-tracking, independent of surprisal and the other effects.

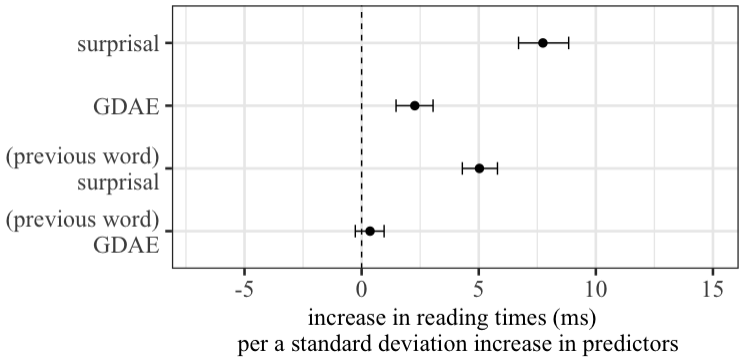
More specifically, positive entropy effects (slowdowns caused by increase in attention entropy) are found on the current target words for self-paced reading times and go-past times in eye-tracking data. In first-fixation durations in eye-tracking data, positive attention entropy effects of the previous word (i.e., spillover attention entropy effects) are found.

There are also positive interaction effects between surprisal and attention entropy, indicating an over-additive effect. This suggests that the attention entropy effect is increased for more unexpected words. It is premature to draw conclusions about their nature, in the absence of theoretical models that make predictions about the interactions.

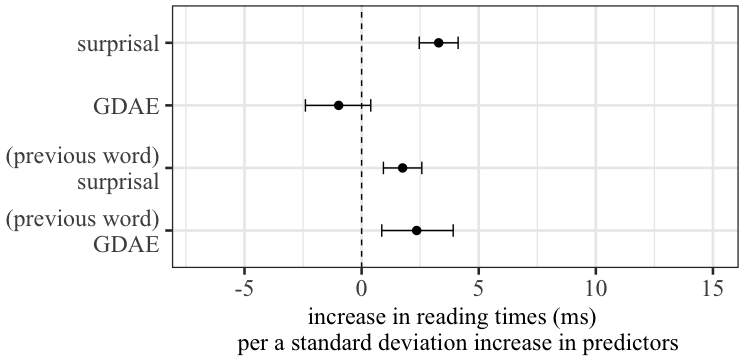
4.6 Discussion

The three robust effects of interest here are (i) the effect of surprisal of the current word, (ii) the effect of surprisal of the previous word, and (iii) the positive effect of GDAE of the current word in self-paced reading times and go-past time in eye-tracking data, or of the

(a) Coefficient estimates from Natural Stories corpus.



(b) Coefficient estimates from GECO corpus (first fixation duration).



(c) Coefficient estimates from GECO corpus (go-past time).

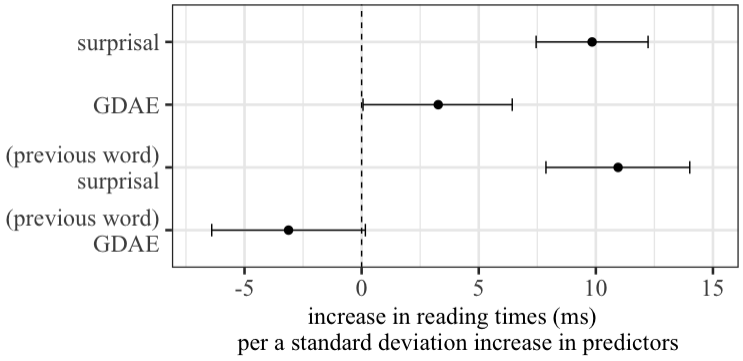


Figure 4.3: Results from Bayesian regression models on predicting reading times. The distribution of posterior estimates of the coefficients of standardized predictors from self-paced reading time and eye-tracking data. The lines indicate 95% credible intervals, and the dots indicate means of estimates.

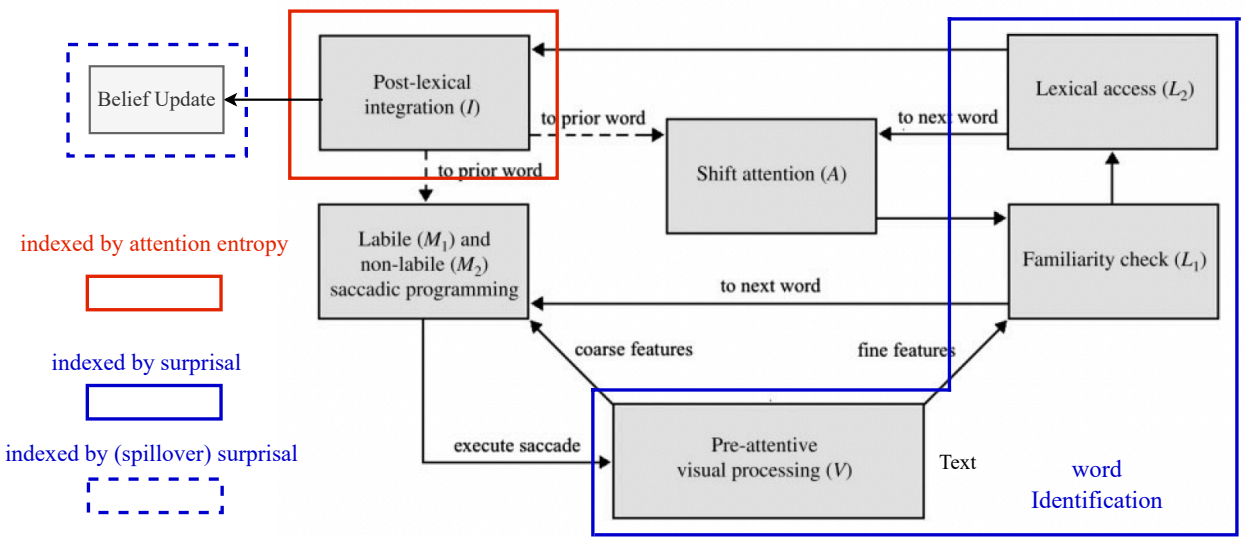


Figure 4.4: Representation of influences from surprisal and attention entropy being integrated on the E-Z Reader model of eye-movement control. The original diagram is from Mancheva et al. (2015)

previous word with first fixation duration in eye-tracking data. In what follows, I will briefly sketch a theoretical interpretation of these effects.

Probabilistic expectations of upcoming words speed up lexical identification and access to the degree that the word is expected (Ehrlich and Rayner, 1981; Rayner and Well, 1996; Schustack et al., 1987). In particular, assuming noisy sequential sampling of visual evidence at some fixed rate, an optimal Bayesian perceiver will require a number of samples (and thus time) that is a linear function of surprisal (Norris, 2009). This effect on word identification would be expected to show up in earlier reading time measures in most accounts of eye movement control, such as EZ-Reader (Reichle et al., 2003), Figure 4.4.

Under the interpretation that surprisal is also a measure of the cost of probabilistic belief update (Levy, 2008) (which itself might be post memory retrieval), surprisal might manifest as a spillover effect in addition to a separate early lexical identification effect. An eye movement control system that is optimized for time efficiency would also aggressively schedule saccades to next words in a manner that would make later stage, higher level linguistic dependency formation effects spillover to upcoming words, as in EZ-Reader or

SWIFT (Engbert et al., 2005).

As attention entropy is considered an indicator of interference level during memory retrieval, I hypothesize that the attention entropy effects would extend into spillover regions. The pattern of attention entropy effects found in the eye-tracking data aligns with the interpretation that attention entropy reflects the sentence processing difficulty associated with later-stage memory retrieval. Specifically, when considering go-past time, which includes the regression to the previous words presumably required for dependency formation, the entropy effects are observed at the target region, suggesting entropy's role in later memory retrieval. In contrast, when the model is fit to first fixation, the entropy effects appear as spillover indicating that later memory retrieval difficulty carries over to subsequent regions.

CHAPTER 5

Effects of Attention Entropies as a Function of Speed-accuracy Tradeoff

If attention entropy reflects the memory integration stage of sentence processing, then its effect size should vary depending on the relative emphasis on speed versus accuracy during reading. Specifically, when accuracy is emphasized over speed, attention entropy effects are expected to increase, as careful integration of linguistic input becomes more important than rapid reading. In contrast, when speed is prioritized over accuracy, readers are likely to allocate less time to memory integration, making attention entropy a weaker predictor of reading times.

To test this, I investigate how the effect of attention entropy on word-level reading times varies with changes in speed-accuracy tradeoff. Specifically, I conduct a self-paced reading experiment that manipulated the speed–accuracy trade-off. Analysis of the data shows that the effect of attention entropy on reading times increases when accuracy is emphasized, supporting that attention entropy reflects memory integration effort.

5.1 Materials

Twelve stories, each with approximately 50 sentences, were used in the experiment. Nine of these were selected from the Natural Stories Corpus (Futrell et al., 2021). The remaining three were included as practice stories: The Money-Box (Andersen, 1855), The Goblin's Club

(a Korean folk tale of unknown author and date, translated into English by Fenkl (2000)), and Louisa May Alcott: A Child's Biography (Alcott, 1915). I adapted these practice stories to match the length and structure of the texts from the Natural Stories Corpus.

5.2 Participants

A total of 52 native English speakers were recruited through the online platform Prolific. Participants were randomly assigned to one of three groups (n = 17, 17, and 18, respectively).

5.3 Procedure

Participants were required to complete the experiment across three separate sessions on different dates. Each session implemented a distinct reward scheme designed to encourage a specific reading goal: emphasizing either speed, accuracy, or a balance of both. During each session, participants read four stories, and their word-by-word reading times were recorded using a self-paced reading paradigm. After reading each story, they completed a set of 12 comprehension questions. The first story of each session served as a practice story and was excluded from data analysis.

In all conditions, participants received a base reward of $3 if they achieved at least 75% accuracy on the comprehension questions. Participants who did not meet this threshold received no compensation for that session, and their data were excluded from the analysis.

In the speed condition, participants earned an additional $0.25 for every 30 seconds saved relative to a 24-minute baseline for reading four stories. In the balance condition, they earned $0.25 for every 60 seconds saved relative to the same 24-minute baseline and received an additional $0.20 for each correct comprehension question, provided they scored at least 36 out of 48 questions across the four stories. In the accuracy condition, participants earned $0.40 for each correct comprehension question beyond the base reward.

The assignment of story order and reward conditions was pseudo-randomized across

groups to control for order effects. To encourage retention across sessions, participants also received additional return bonuses of $1 for completing the second visit and $2 for completing the third.

5.4 Bayesian Modeling of Sentence Reading Times Collected under Different Speed-Accuracy-Tradeoff

To investigate how the effect of attention entropy changes in explaining word-by-word reading times under different speed–accuracy trade-off conditions, I conducted two sets of analyses using the experimental data.

First, I examined the coefficients of Grammar Dependency Attention Entropies (GDAE) from a Bayesian mixed-effects regression model fit using Equation 5.1. The model structure closely follows that used in Chapter 4, and the same weakly informative priors were applied.

$$\begin{aligned} ReadingTimes \sim brm(&Surprisal * GDAE + Frequency + WordLength + \\ &SpilloverSurprisal * SpilloverGDAE + SpilloverFrequency + \\ &SpilloverWordLength + Position + Visit + (1|subjectID) + (1|word)) \end{aligned} \tag{5.1}$$

Second, I evaluated the extent to which attention entropy improves model fit under different speed-accuracy emphasis conditions. Specifically, I compared the log-likelihoods of two nested frequentist models: a baseline model that excluded attention entropy (Equation 5.2) and a full model that included attention entropy (Equation 5.3). By computing the log-likelihood ratio for each experimental condition, I quantified the additional explanatory power gained by incorporating attention entropy into the model.

$$\begin{aligned} ReadingTimes \sim lmer(&Surprisal + Frequency + WordLength + \\ &SpilloverSurprisal + SpilloverFrequency + SpilloverWordLength + \\ &Position + Visit + (1|subjectID) + (1|word)) \end{aligned} \tag{5.2}$$

$$\begin{aligned} ReadingTimes \sim lmer(&Surprisal * GDAE + Frequency + WordLength + \\ &SpilloverSurprisal * SpilloverGDAE + SpilloverFrequency + \\ &SpilloverWordLength + Position + Visit + (1|subjectID) + (1|word)) \end{aligned} \tag{5.3}$$

In both analyses, I included random intercepts for subjects and word types but omitted random slopes to simplify the model structure. This decision was motivated by the relatively smaller sample size in this study compared to the data used in Chapter 4, which made fitting more complex random-effects structures less stable and reliable.

5.5 Results

Figure 5.1 displays average reading times and comprehension accuracies across conditions and session (visit) numbers. As expected, reading times decrease as the task places greater emphasis on speed. Accuracy is slightly greater in the accuracy and balanced condition than in the speed condition.

The full results from the Bayesian model are presented in Table 5.1 and the coefficients of the surprisals and GDAEs are shown in Figure 5.2. The locus of surprisal effects varies by condition, indicating successful experimental manipulation. Surprisal effects appear primarily in the target region under the accuracy condition, shift to the spillover region in the speed condition, and appear in both regions in the balanced condition. This suggests that faster reading delays the processing of unexpected words, resulting in spillover effects. In all

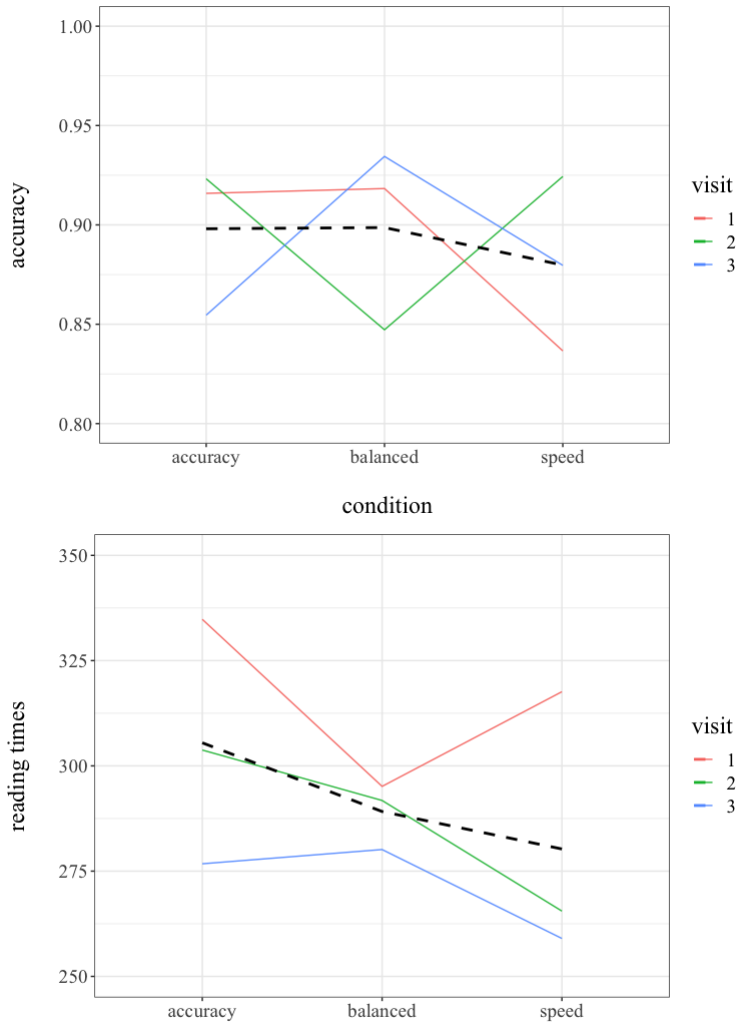


Figure 5.1: Accuracy (top) and reading times (bottom) by condition across three visits. The dashed line shows the condition-wise average across all visits.

three conditions, attention entropy effects emerge in the spillover region. As expected, the accuracy condition exhibits substantially stronger attention entropy effects compared to the speed and balanced conditions.

The results from the second analysis show that the log-likelihood ratio between the full model (including attention entropy) and the baseline model (excluding it) is largest in the accuracy condition (27.44). The ratios for the balanced (16.48) and speed (17.47) conditions, being considerably lower than in the accuracy condition. This pattern further supports the hypothesis that attention entropy contributes more explanatory power when memory

integration is emphasized (i.e., when accuracy is prioritized during reading).

Table 5.1: Posterior estimates for the fixed effects of predictors on word reading times in three speed-accuracy manipulation conditions. Predictors with significant effects are bolded.

Accuracy Condition

Parameters	Estimate	95% CrI
intercept	332.07	[-492.62, 1099.99]
visit	4.56	[-329.69, 387.98]
word position	**-0.70**	**[-1.02, -0.38]**
word length	4.24	[-0.17, 8.61]
surprisal	**3.75**	**[0.32, 7.19]**
GDAE	-2.06	[-6.41, 2.42]
word frequency	-2.33	[-6.99, 2.46]
surprisal × GDAE	-0.07	[-3.04, 2.96]
(previous word) word length	**7.02**	**[2.95, 11.07]**
(previous word) surprisal	1.29	[-2.03, 4.61]
(previous word) GDAE	**7.21**	**[3.18, 11.33]**
(previous word) word frequency	-10.33	[-14.43, 6.43]
(previous word) surprisal × (previous word) GDAE	2.01	[-0.87, 4.97]

Balanced Condition

Parameters	Estimate	95% CrI
intercept	216.65	[-425.42, 898.52]
visit	13.82	[-286.03, 307.66]
word position	**-0.48**	**[-0.72, -0.25]**
word length	**16.38**	**[8.62, 24.58]**

surprisal	**3.53**	**[1.10, 6.00]**
GDAE	0.17	[-3.11, 3.53]
word frequency	-2.49	[-6.06, 1.08]
surprisal × GDAE	0.39	[-1.81, 2.61]
(previous word) word length	**3.54**	**[0.46, 6.64]**
(previous word) surprisal	**3.53**	**[1.07, 5.91]**
(previous word) GDAE	**3.03**	**[0.12, 5.92]**
(previous word) word frequency	0.10	[-2.87, 3.00]
(previous word) surprisal × (previous word) GDAE	0.83	[-1.27, 2.94]

Speed Condition

Parameters	Estimate	95% CrI
intercept	330.53	[-150.13, 800.95]
visit	-26.54	[-238.48, 212.36]
word position	0.04	[-0.19, 0.26]
word length	1.84	[-1.08, 4.79]
surprisal	-1.05	[-3.46, 1.36]
GDAE	-1.10	[-4.29, 1.93]
word frequency	-0.17	[-3.35, 2.90]
surprisal × GDAE	0.30	[-1.79, 2.34]
(previous word) word length	-1.24	[-4.27, 1.83]
(previous word) surprisal	**2.74**	**[0.28, 5.10]**
(previous word) GDAE	**3.66**	**[0.74, 6.56]**
(previous word) word frequency	**-4.32**	**[-7.31, -1.54]**
(previous word) surprisal × (previous word) GDAE	1.28	[-0.98, 3.40]

5.6 Discussion

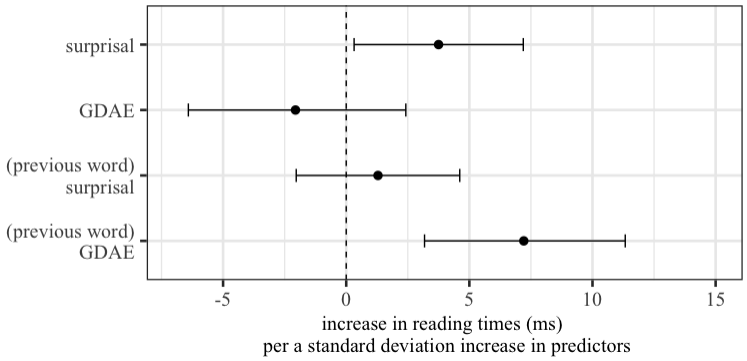
The results from both analyses support the interpretation that attention entropy reflects processing difficulty associated with memory integration during sentence comprehension. Crucially, the effect of attention entropy on reading times was strongest when participants were incentivized to prioritize accuracy, aligning with the view that memory integration becomes more important, and thus more predictive of reading times.

Interestingly, while attention entropy effects appeared in the target region in the Natural Stories Corpus (Futrell et al., 2021) analysis in Chapter 4, they emerged in the spillover region in the current speed–accuracy manipulation experiment. I speculate that this shift in the locus of attention entropy effects may be due to practice effects that led participants to read more quickly, increasing the likelihood of spillover effects. In the present study, participants completed the reading task across multiple sessions, whereas participants in the Natural Stories Corpus study completed a single session and did not return. The repeated sessions and increased familiarity with the task may have contributed to faster reading and reduced time for integrating linguistic input at the target word.

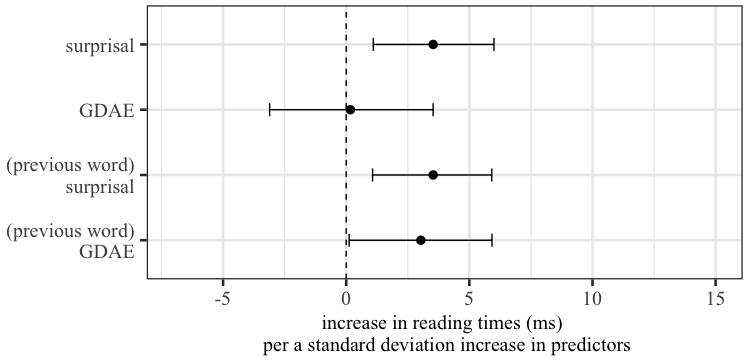
This interpretation is supported by a comparison of average reading times: the mean word reading time in the Natural Stories Corpus is 338.24 ms, whereas the average in the present study was 291.62 ms. This suggests that participants in the current experiment read more quickly overall, possibly pushing memory integration effects into the spillover region.

In sum, the observed spillover effects of attention entropy in this experiment, along with their modulation by the speed-accuracy tradeoff, further support the idea that attention entropy captures a stage of sentence processing distinct from surprisal, one that likely reflects memory retrieval and integration costs.

(a) Coefficient estimates from accuracy-emphasis condition



(b) Coefficient estimates from balanced condition



(c) Coefficient estimates from speed condition

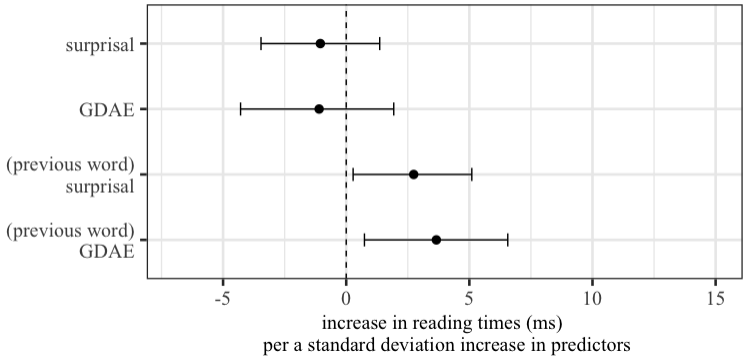


Figure 5.2: Results from Bayesian regression models on predicting reading times under different speed-accuracy tradeoff. The distribution of posterior estimates of the coefficients of standardized predictors from self-paced reading time and eye-tracking data. The lines indicate 95% credible intervals, and the dots indicate means of estimates.

CHAPTER 6

Explaining Psycholinguistic Phenomena Using Attention Entropies

In this chapter, I demonstrate how a range of psycholinguistic phenomena—particularly those associated with memory interference—can be explained using Transformer models' integrative accounts that combine expectation-based and memory-based perspectives. I begin with interference effects in subject–verb agreement and then examine non-agreement interference effects, embedded sentence processing, relative clause processing, and garden path effects. Depending on the type of phenomenon, I use either GDAE or local attention entropy that is computed from a particular attention head specialized for a specific grammar dependency.

6.1 Interference Effects in Subject-verb Agreement

In Section 3.3, I briefly explained how interference effects in subject-verb agreement can be explained with Transformers' attention patterns. In what follows, I will scale up the discussion of how Transformer's attention entropy could provide further explanation about the agreement interference phenomena.

6.1.1 Background

Dependency formation between the subject and the verb can be interfered by distractor nouns held in short-term memory. The degree of this interference depends on the similarity-

level between the retrieval cue (the verb) and potential retrieval candidates (e.g., the subject and intervening distractors) (Lakretz et al., 2021; Wagers et al., 2009; Dillon et al., 2013; Vasishth and Engelmann, 2021).

(2) a. The **slogan** on the poster **was** designed ...

b. The **slogan** on the posters **was** designed ...

c.* The **slogan** on the posters **were** designed ...

d.* The **slogan** on the poster **were** designed ...

These examples in (2) illustrate how subject–verb agreement processing can vary based on two factors: grammaticality and interference. Sentences (2a) and (2b) are grammatical, satisfying subject–verb number agreement. However, they differ in the type of distractor: (2a) includes an interfering distractor (number-compatible with the verb), while (2b) contains a non-interfering distractor (number-incompatible). Similarly, ungrammatical sentences (2c) and (2d) differ in distractor type: (2c) contains an interfering distractor, and (2d) a non-interfering one.

According to cue-based retrieval theory, interfering distractors increase processing difficulty by competing more strongly with the correct subject due to their similarity to the retrieval cue (i.e., the verb). The interference effect is typically measured as the difference in processing difficulty (that can be measure with reading times) between interfering and non-interfering conditions. Interestingly, the manifestation of these effects depends on grammaticality: interference tends to be *inhibitory* (i.e., slower processing) in grammatical conditions (2a–2b), but *facilitatory* (i.e., faster processing) in ungrammatical conditions (2c–2d).

One possible explanation for facilitatory effects in ungrammatical sentences is misretrieval (Logačev and Vasishth, 2016). When the true subject has low cue-match strength and the distractor is highly compatible, comprehenders may mistakenly retrieve the distractor, resulting in an illusion of grammaticality and faster—but incorrect—processing.

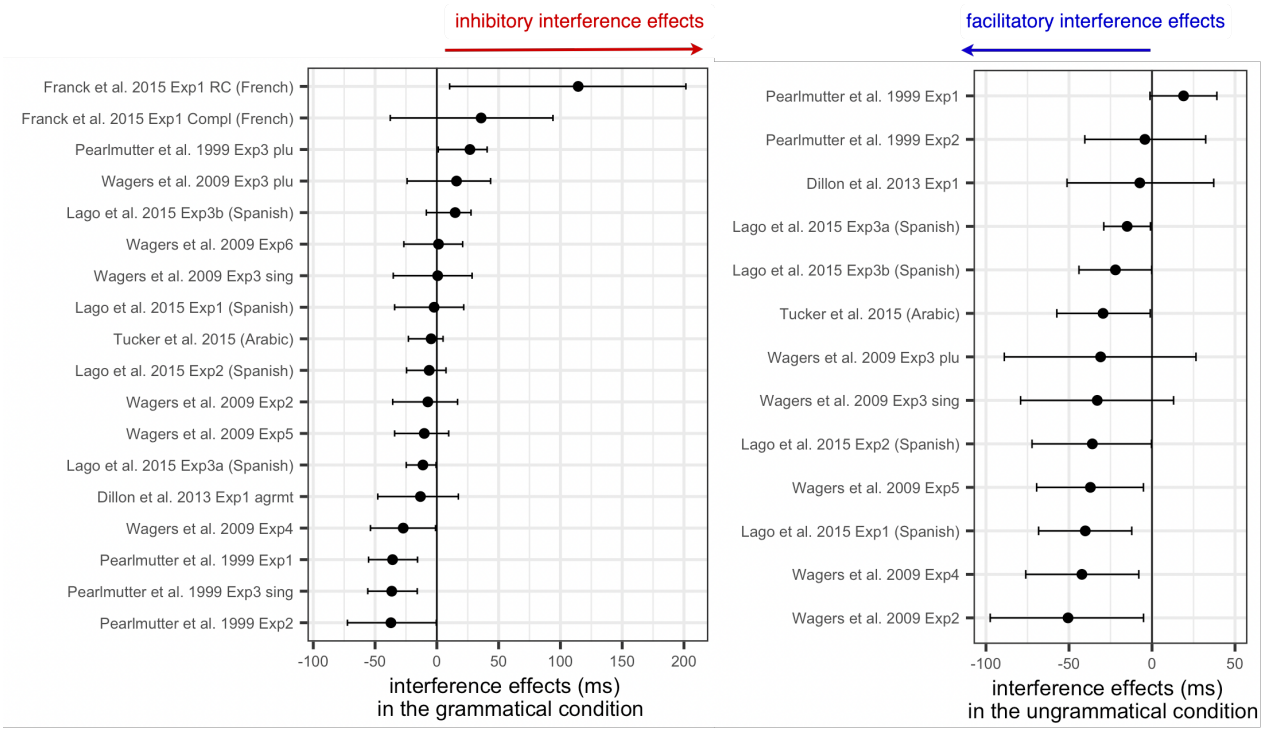


Figure 6.1: Results of the meta-analysis on subject-verb number agreement from Vasishth and Engelmann (2021).

A meta-analysis by Jäger et al. (2017) and Vasishth and Engelmann (2021) found that facilitatory interference effects in ungrammatical sentences are robustly observed in reading time data, while inhibitory effects in grammatical sentences are less reliably detected (see Figure 6.1).

In the following analysis, I investigate how Transformers can capture these interference effects by providing both surprisal-based and memory-based accounts. To this end, I compute metrics from GPT2-small —surprisal, attention entropy, and attention weight directed from the cue to the target. I also visualize attention patterns originating from the cue words (i.e., verbs). For these analyses, I focus on $\text{head}_{4,3}$, which was identified in Section 3.5.1 as specializing in `nsubj` dependency resolution.

6.1.2 Methods

Four metrics were measured with GPT2-small at the cue (the verb): surprisal, attention to target that simply computes the amount of attention paid to the correct target from the cue measured using $\text{head}_{4,3}$, local attention entropy measured at $\text{head}_{4,3}$, and GDAE.

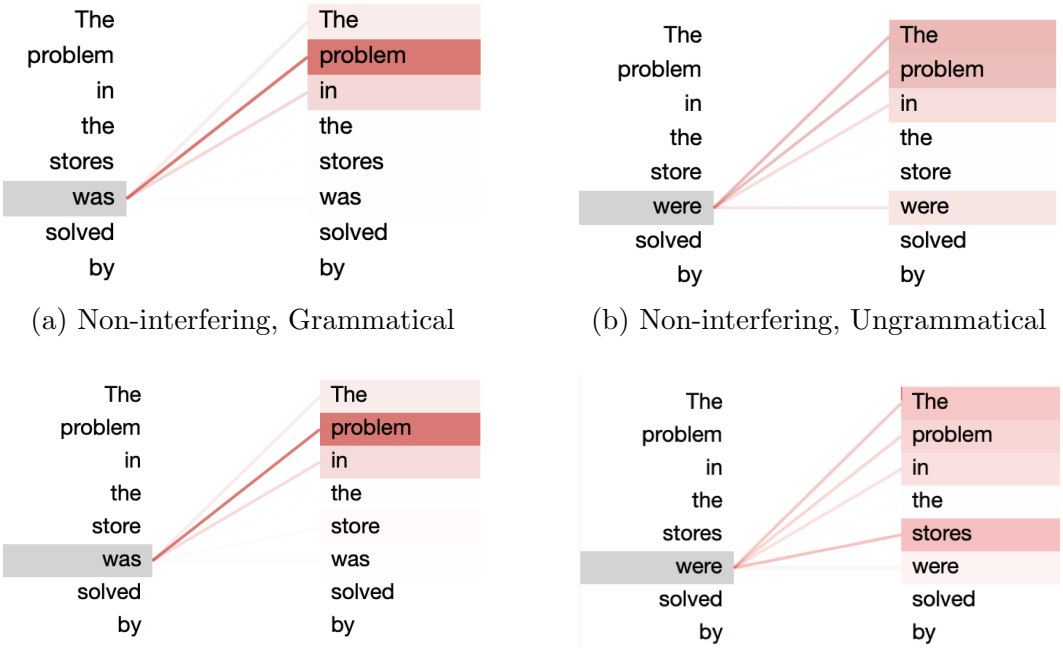
Attention distribution patterns are also visualized using Vig (2019)'s Transformer attention visualization tool.

6.1.3 Materials

Table 6.1: A set of data included for the experiment on subject-verb agreement. (Wagers (2009)'s Exp3 also included sets with plural subjects in the ungrammatical conditions.)

Wagers 2009 Exp 2-3	int	gram	The commentator who the **viewer trusts** ...
	non-int	gram	The commentators who the **viewer trusts** ...
	int	ungram	*The commentators who the **viewer trust** ...
	non-int	ungram	*The commentator who the **viewer trust** ...
Wagers 2009 Exp 4-6	int	gram	The **slogan** on the poster **was** designed ...
	non-int	gram	The **slogan** on the posters **was** designed ...
	int	ungram	*The **slogan** on the posters **were** designed ...
	non-int	ungram	*The **slogan** on the poster **were** designed ...
Dillon 2013 Exp 1 agrmt	int	gram	The **executive** who oversaw the middle manager apparently **was** dishonest ...
	non-int	gram	The **executive** who oversaw the middle managers apparently **was** dishonest ...
	int	ungram	*The **executive** who oversaw the middle managers apparently **were** dishonest ...
	non-int	ungram	* The **executive** who oversaw the middle manager apparently **were** dishonest ...

48 sets of sentences from Wagers et al. (2009)'s Experiments 2-3, 24 sets of sentences from Wagers et al. (2009)'s Experiments 4-7, and 48 sets of sentences from Dillon et al. (2013)'s



(a) Non-interfering, Grammatical

(b) Non-interfering, Ungrammatical

(c) Interfering, Grammatical

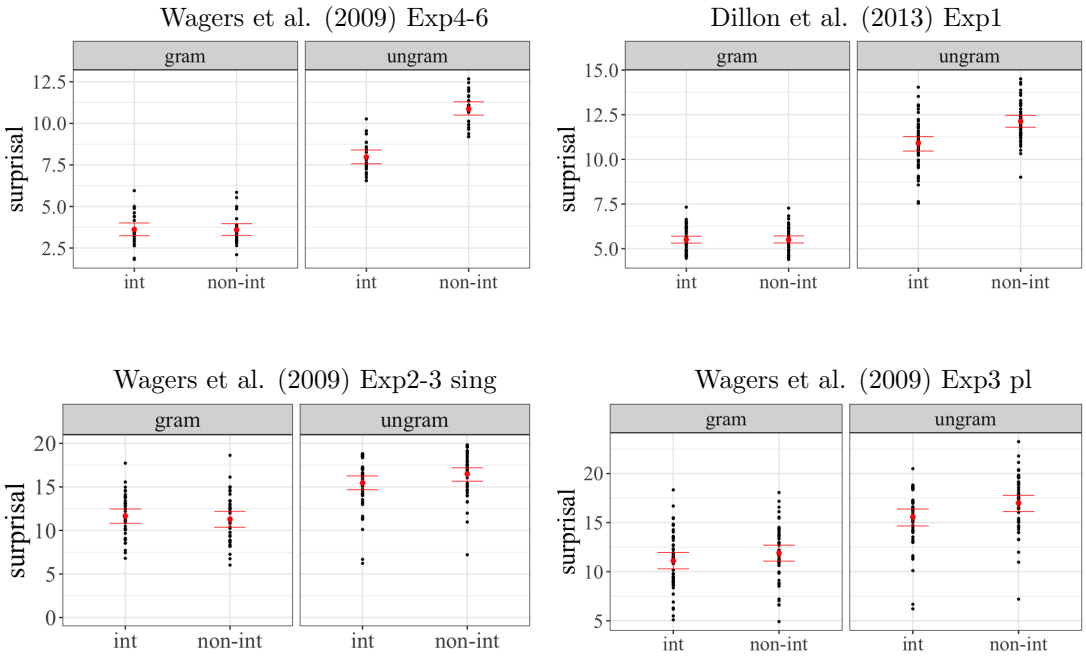
(d) Interfering, Ungrammatical

Figure 6.2: Attention distribution patterns observed at $\text{head}_{4,3}$ by manipulation

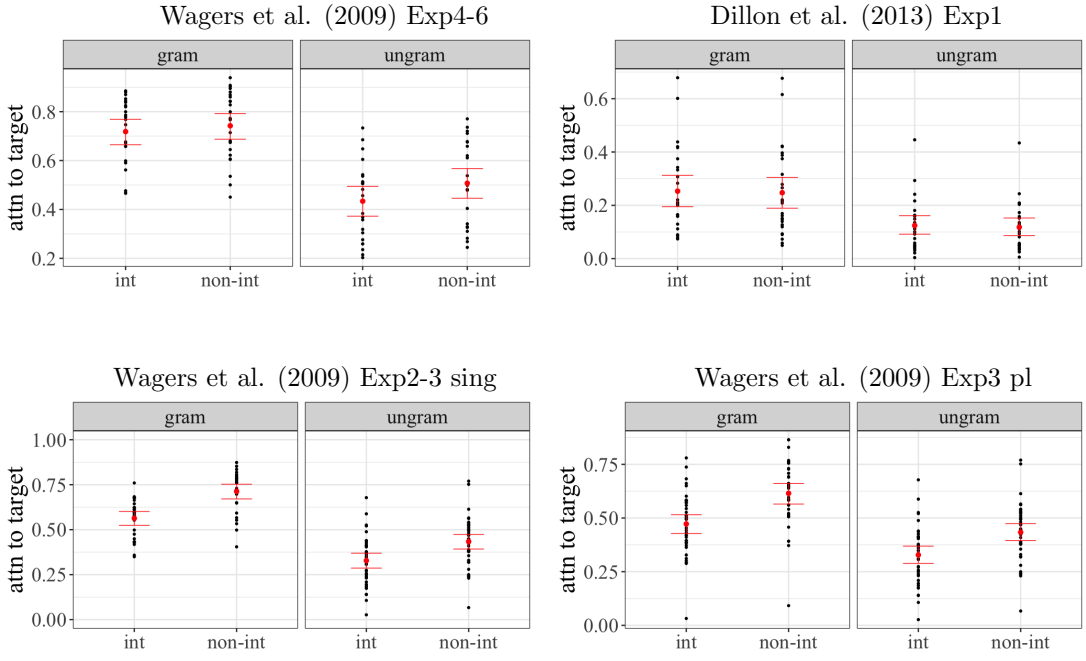
Experiment 1 were included to examine how GPT2-small accounts for the interference effects. All of the sets were in 2 × 2 factorial, having grammaticality and interference as factors. The example sentences are in Table 6.1.

6.1.4 Results

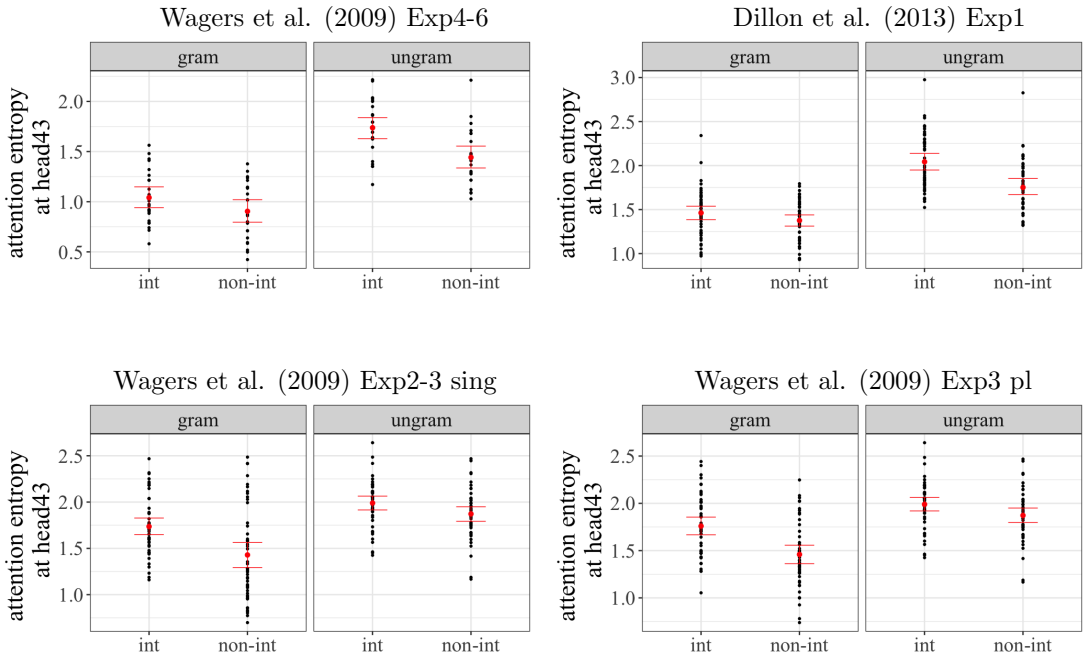
As shown in Figure 6.2, attention distribution patterns observed in head_{43}, are consistent with cue-based retrieval theory. In the grammatical condition (Figures 6.2a and 6.2c), a high proportion of attention is paid to the correct target (*problem*), the distractor in the interfering condition getting a bit more attention compared to the non-interfering condition. Conversely, the ungrammatical condition (Figures 6.2b and 6.2d) shows a different pattern. The correct target gets significantly less attention, and the attention paid to the distractor is even larger in the interfering condition where the distractor has a great amount of compatibility with the cue.



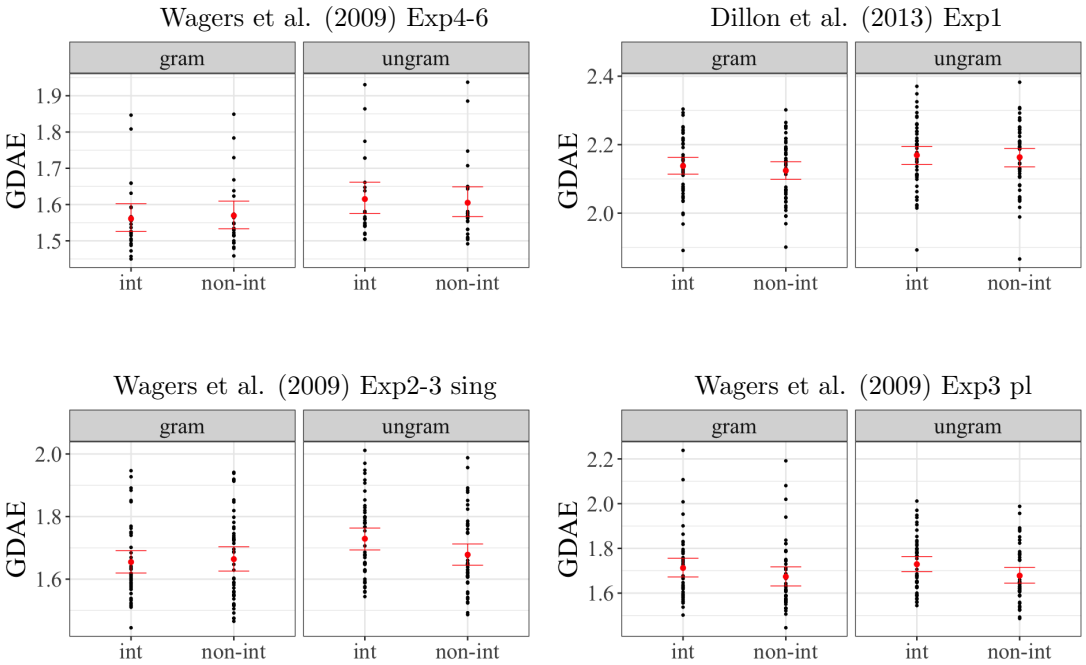
(a) Surprisals



(b) Attention to target measured at $\text{head}_{4,3}$



(c) Local attention entropy measured at $\text{head}_{4,3}$



(d) GDAE

Figure 6.3: At verbs of interest (e.g., *was* in sentences in (2)), four metrics –*surprisals*, *attention to target*, *local attention etnropy measured at* head_{43}, and *GDAE* – were measured. The red dots and lines indicate means and 95% confidence intervals.

	singular subj	plural subj
interfering	80	71
non-interfering	39	51

Table 6.2: Frequency of interfering vs. non-interfering distractors in ungrammatical subject–verb agreement constructions from a randomly sampled subset of the Reddit corpus. Interfering distractors appear approximately twice as often as non-interfering ones in cases involving singular subjects and ungrammatical plural verbs.

The difference among the conditions is also captured with the metrics computed from GPT2-small. The ungrammatical/interfering condition shows the smallest amount of attention to target (Figure 6.3b) and the largest attention entropies (Figures 6.3c and 6.3d), though patterns from aggregate attention entropies are not as clear as local attention entropies.

6.1.5 Discussion

The attention-based metrics and visualization provide explanation about the interference effects in subject-verb number agreement from the perspective of cue-based retrieval. At the same time, it was also observed that surprisals computed with GPT2-small show the same pattern as Vasishth and Engelmann (2021)'s meta analysis (Figure 6.3a): no inhibitory interference effects in the grammatical condition and facilitatory interference effects in the ungrammatical condition.

One possible explanation for the observed facilitatory interference effects is that GPT2-small was exposed to ungrammatical sentences in the training data that have precisely the interference patterns of the ungrammatical sentences in our experiments. To examine such possibility, I analyzed 241 sentences randomly extracted from a Reddit corpus (Chang et al., 2020) whose subjects and verbs do not agree in number, and have either interfering or non-interfering distractors in between. The results shown in Table 6.2 suggest that interfering distractors occur about twice as often as non-interfering distractors in the case of singular subjects with an ungrammatical plural verb, consistent with our expectations that

agreement-attraction errors in production may be evident in unedited corpora. But it seems unlikely that this 2:1 ratio, which corresponds to about a 1 bit difference in surprisal, is sufficient alone to explain the observed surprisal differences. For example, in the Wagers et al. (2009)'s Experiment 4–6, I observed about a 3 bit difference in surprisal, a 2 bit or 4x difference in probability relative to what would be expected on the basis of the corpus counts. More extensive corpus analysis is necessary to confidently rule out this explanation.

6.2 Non-agreement Interference Effects

6.2.1 Background

Interference during subject–verb dependency formation can occur even in the absence of overt grammatical features like number agreement. That is, simply having multiple items with similar linguistic features stored in short-term memory can disrupt retrieval processes, leading to increased processing difficulty. This phenomenon is known as interference effects in non-agreement (Van Dyke and Lewis, 2003; Van Dyke, 2007; Vasishth and Engelmann, 2021).

Such interference effects can be illustrated using the sentence sets in (3) from Van Dyke and Lewis (2003). In these sentences, the retrieval cue *was* must establish two separate dependency relations to be correctly interpreted — one with the main verb in the sentence *forgot* and one with the subject *letter*. During the formation of these dependencies, interference arises due to syntactic complexity introduced by distractors that are partially compatible with the retrieval cues.

Specifically, sentence (3a) is expected to cause no retrieval interference, as it contains no elements that disrupt the formation of the required dependencies. In contrast, sentences (3b) and (3c) introduce increasing levels of interference due to the presence of syntactic distractors. If dependency formation difficulty were driven purely by the distance between the cue and the target, the processing difficulty in (3b) and (3c) would be expected to be

similar. However, previous studies (Van Dyke and Lewis, 2003; Vasishth and Engelmann, 2021) suggest that (3c) induces greater processing difficulty, which is attributed to its higher cue–distractor similarity.

(3) a. Short distance (Short)

The executive assistant **forgot$_{\textbf{target}}$** [that] the letter **was$_{\textbf{cue}}$** waiting for a signature.

The executive assistant forgot [that] the **letter$_{\textbf{target}}$** **was$_{\textbf{cue}}$** waiting for a signature.

b. Long distance & Low interference (Long/Low)

The executive assistant **forgot$_{\textbf{target}}$** [that] the letter which had$_{\text{d1}}$ fallen on the floor **was$_{\textbf{cue}}$** waiting for a signature.

The executive assistant forgot [that] the **letter$_{\textbf{target}}$** which$_{\text{d1}}$ had fallen on the floor $_{\text{d2}}$ **was$_{\textbf{cue}}$** waiting for a signature.

c. Long distance & High interference (Long/High)

The executive assistant **forgot$_{\textbf{target}}$** [that] the letter which revealed$_{\text{d1}}$ that the mayor was$_{\text{d2}}$ responsible **was$_{\textbf{cue}}$** waiting for a signature.

The executive assistant forgot [that] the **letter$_{\textbf{target}}$** which revealed that the mayor $_{\text{d1}}$ was responsible **was$_{\textbf{cue}}$** waiting for a signature.

Van Dyke and Lewis (2003) explains these interference effects using similarity-based interference within the cue-based retrieval theory. In (3b) and (3c), *'was'* triggers two retrieval cues that needs to be integrated with distinct targets: (1) a verb that takes a sentential complement as its constituent, and (2) a noun whose number is singular taking the nominal position. Finding the target for the first cue, (3b) has only one distractor (*'had'*) that partially matches the cue since it has the property of being a verb. Conversely, (3c) has two

distractors (*'revealed'* and *'was'*), one (*'revealed'*) being fully compatible with the retrieval cue and thus generating more confusion while forming the dependency relation than (3b) Likewise, for the second cue, distractors in (3c) show greater compatibility with the cue, especially the distractor (*'mayor'*) being more compatible with the cue than distractors in (3b) as *mayor* is nominal while *floor* is a prepositional noun.

Notice that the interference in this example interacts with garden-path effects in that *forgot* is likely to be interpreted as a verb that takes an NP complement rather than a globally correct interpretation, which is a verb that takes a sentential complement. Unambiguous sentences whose main verb (*'forgot'*) is followed by *that* avoid such conflation, allowing the comparison between the cases where garden path effects are exerted and not. A more detailed discussion of garden-path effects, including their explanation via cue-based retrieval theory and Transformer models, is deferred to Section 6.5. For now, I focus on interference effects induced by distance and syntactic complexity in both ambiguous and unambiguous sentences. In particular, I examine how attention entropy and surprisal make distinct predictions regarding interference in interaction with ambiguity, showing the distinct roles of these two metrics in explaining psycholinguistic phenomena.

Since greater processing difficulty is expected when similarity-based interference is high, as predicted by cue-based retrieval theory, attention entropy should be highest in the *Long & High* condition. This increased entropy effect should persist regardless of ambiguity, as similarity-based interference according to the cue-based retrieval theory is independent of the presence of a disambiguating device (i.e., *that* in (3)). In contrast, from an expectation-based perspective, processing difficulty in the *Long & High* condition would be not pronounced with the disambiguating device, as it helps to guide to the correct interpretation, reducing the degree of unexpectedness even when similarity-based interference is high. Accordingly, surprisal effects in the *Long & High* condition are expected to be significant only in the ambiguous condition.

6.2.2 Methods

Interference effects in the example sentences in the present study are accounted for with two different dependency relations: one for the relation between the subject (*letter*) and the cue (*was*) and the other for the relation between the matrix verb (*forgot*) and the embedded verb (*was*). Thus, in addition to $\text{head}_{4,3}$, which was found to be responsible for `nsubj` dependency relation, I also investigated attention-based metrics from head_{42}, which is found to be specialized for `ccomp` dependency relation that associates verb and its clausal complement. (See Appendix A. for the detailed process of syntactic heads identification.) As a result, I computed six metrics to examine interference effects in subject verb non-agreement: surprisal, GDAE, attention entropy at $\text{head}_{4,3}$, attention entropy at head_{42}, attention to target (subject of the verb) at $\text{head}_{4,3}$ and attention to target (matrix verb) at head_{42} and surprisal.

Attention distribution patterns from $\text{head}_{4,3}$ and head_{42} are also visualized using Vig (2019)'s Transformer attention visualization tool.

6.2.3 Materials

36 sets of sentences from Van Dyke and Lewis (2003)'s Experiments 3-4 are included. In order to prevent any confound effects from word position, I manipulated sentences to have the critical word (i.e., *was* in (3)) at the same position for *long_high* and *long_low* conditions by having an adverbial word in sentences ('*yesterday*'). The example sentences after controlling for the word position are in Table 6.3.

6.2.4 Results

Figure 6.4 (from the head for the subject-verb relation; `nsubj`) and Figure 6.5 (from the head for the matrix verb and embedded verb relation; `ccomp`) illustrate the attention distribution from the cue (*was*). The most diffuse attention distribution is observed in the *Long* & *High*

Table 6.3: An example set of materials used for the experiment on interference effects of subject-verb non-agreement

Short	The executive assistant **forgot**$_{\text{target}}$ [that] the letter **was**$_{\text{cue}}$ waiting for a signature.
	The executive assistant forgot [that] the **letter**$_{\text{target}}$ **was**$_{\text{cue}}$ waiting for a signature.
Long & Low	The executive assistant **forgot**$_{\text{target}}$ [that] the letter which had$_{\text{d1}}$ fallen on the floor yesterday **was**$_{\text{cue}}$ waiting for a signature.
	The executive assistant forgot [that] the **letter**$_{\text{target}}$ which$_{\text{d1}}$ had fallen on the floor$_{\text{d2}}$ yesterday **was**$_{\text{cue}}$ waiting for a signature.
Long & High	The executive assistant **forgot**$_{\text{target}}$ [that] the letter which revealed$_{\text{d1}}$ that the mayor was$_{\text{d2}}$ responsible **was**$_{\text{cue}}$ waiting for a signature.
	The executive assistant forgot [that] the **letter**$_{\text{target}}$ which revealed that the mayor$_{\text{d1}}$ was responsible **was**$_{\text{cue}}$ waiting for a signature.

condition, in both ambiguous and unambiguous cases, aligning with predictions from the cue-based retrieval theory. Figure 6.4 indicates that interference increases as more distracting noun phrases intervene between the verb (*was*) and its subject (*letter*).

However, Figure 6.5 does not provide a clear depiction of interference effects, as head$_{42}$ exhibits a strong bias toward assigning the highest attention to the first token. Further investigation is needed to better understand the role of head$_{42}$ in processing the relationship between matrix verbs and embedded verbs. Despite this limitation, head$_{42}$ is still included in this analysis, as its attention entropy computations provide additional evidence for inter-

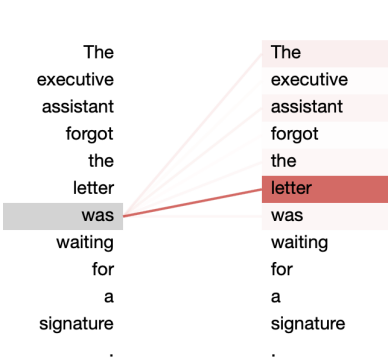
ference effects.

The computed metrics align with predictions from expectation-based and memory-based theories. As shown in Figures 6.6d and 6.6f, greater attention entropy is observed in conditions with high interference. Additionally, Figure 6.6c shows that attention directed to the correct target is highest in the *Short* condition and lowest in the *Long* & *High* condition. Figure 6.6e indicates lower attention to the target in the *Short* condition, although no significant difference is observed between the *Long* & *High* and *Long* & *Low*. Finally, Figure 6.6a reveals that interference effects do not manifest in surprisal values for unambiguous sentences but are present in the ambiguous condition.

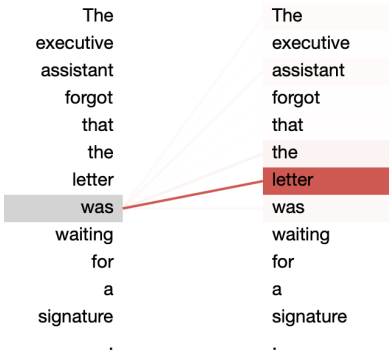
6.2.5 Discussion

The results demonstrate how memory retrieval interference effects in non-agreement can be explained through the Transformer's attention mechanism. Although GDAEs do not distinguish between the *Long* & *High* and *Long* & *Low* conditions, grammar-specific heads ($\text{head}_{4,3}$ and head_{42}) clearly reveal memory interference effects. These effects are observed in both *ambiguous* and *unambiguous* conditions.

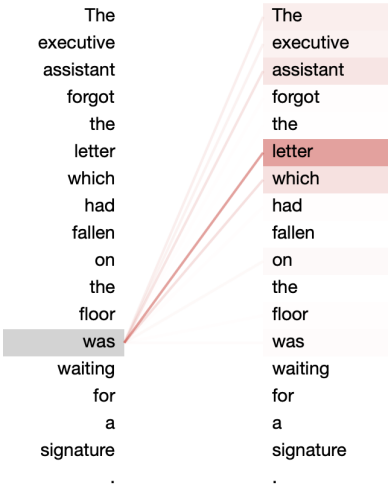
The interference effects in the *Long* & *High* unambiguous condition are not captured by surprisal. This is expected, as surprisal measures predictability rather than memory interference, and the presence of a disambiguating device (*that*) in the unambiguous condition facilitates the globally correct interpretation of the sentence. The fact that interference effects in the unambiguous condition are only reflected in attention entropy highlights the distinct roles of surprisal and attention entropy in explaining sentence processing patterns.



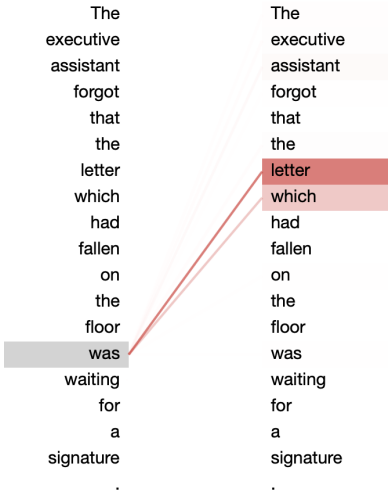
(a) Short, Ambiguous



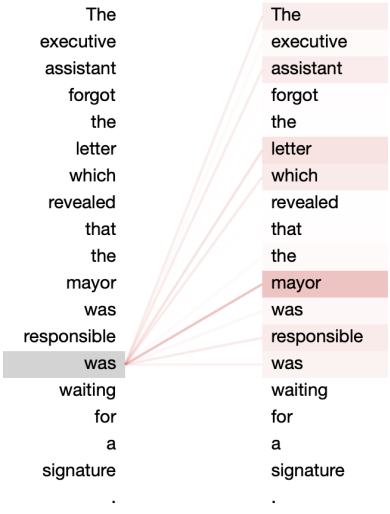
(b) Short, Unambiguous



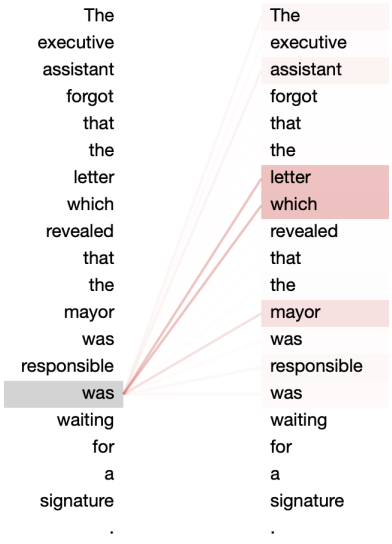
(c) Long-Low, Ambiguous



(d) Long-Low, Unambiguous

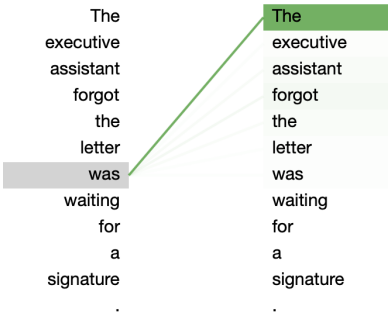


(e) Long-High, Ambiguous

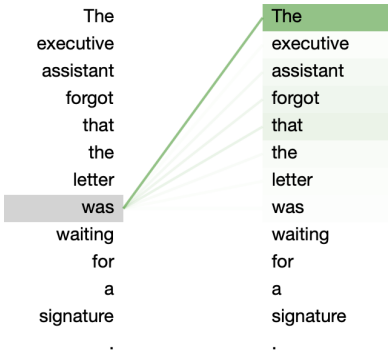


(f) Long-High, Unambiguous

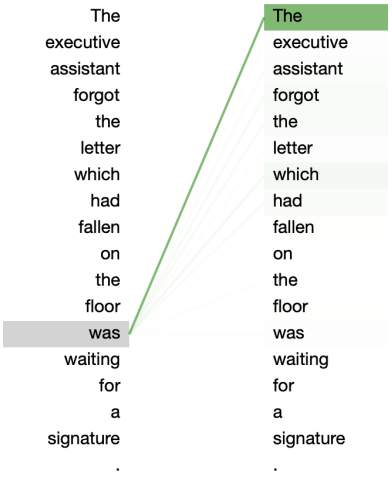
Figure 6.4: Attention distribution patterns observed at head_{43}, specialized for subject-verb relations. The attention distribution is the most diffuse in *Long & High* condition, regardless of ambiguity.



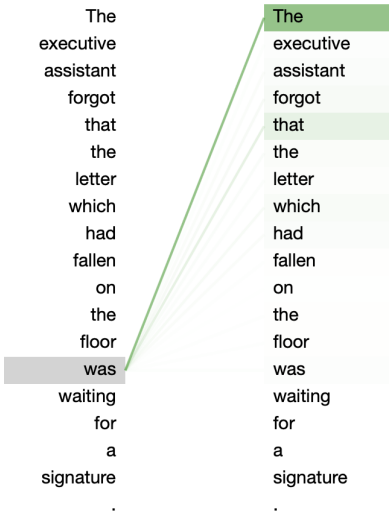
(a) Short, Ambiguous



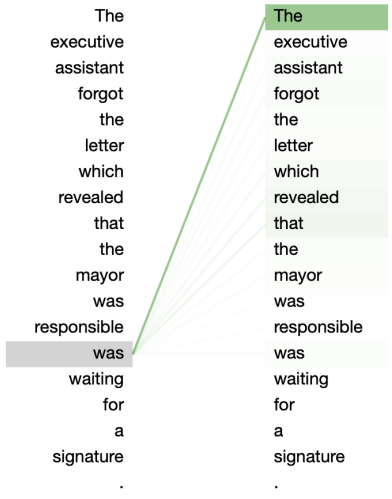
(b) Short, Unambiguous



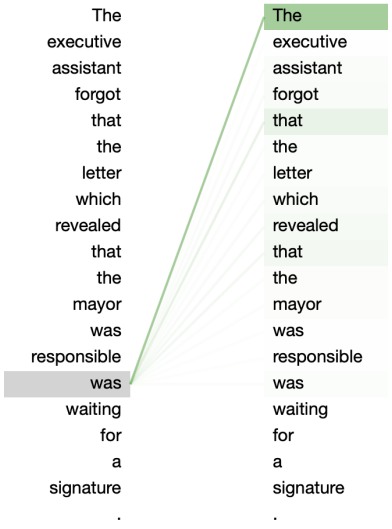
(c) Long-Low, Ambiguous



(d) Long-Low, Unambiguous

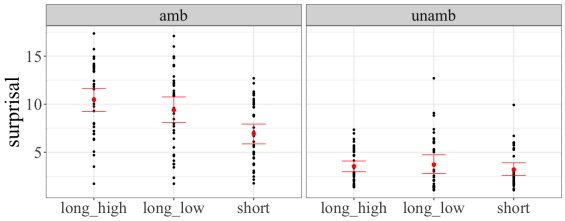


(e) Long-High, Ambiguous

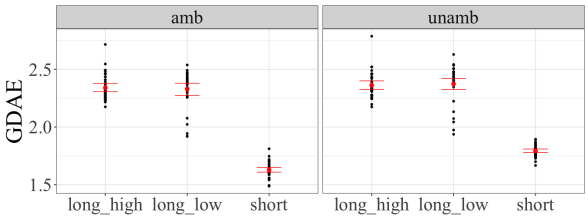


(f) Long-High, Unambiguous

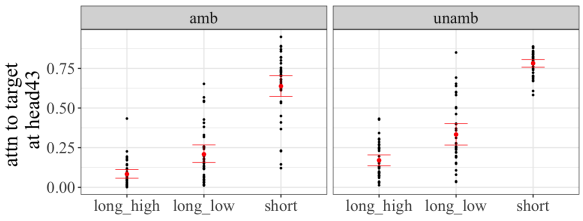
Figure 6.5: Attention distribution patterns observed at head_{42}, specialized for clausal complement relations (*ccomp*). Even though the contrast is not clear due to the attention head's bias to the first token, the attention distribution is the most diffuse in *Long & High* condition regardless of ambiguity.



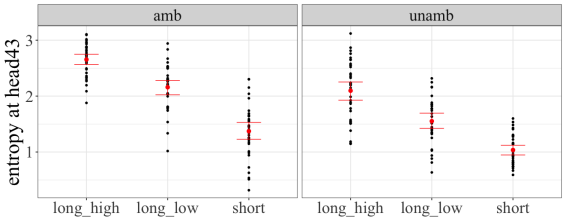
(a) Surprisal



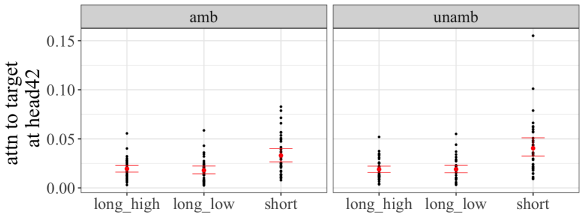
(b) GDAE



(c) Attention allocated to correct subject noun by head_{43}



(d) Attention entropy at head_{43}



(e) Attention allocated to correct subject noun by head_{42}

mm
(f)
At-
ten-
tion
en-
tropy
at
head_{42}

Figure 6.6: *Surprisals* and attention-based metrics are measured at the verb of interest in the materials (e.g., at *was* in examples sentences in (3)).

6.3 Self-embedded Sentences

6.3.1 Background

Center-embedded constructions have constituent structures of the form

(4) Center embedding: $[_{\alpha} \ldots [_{\beta} \ldots] \ldots]$

where the constituent β is embedded within constituent α with material on either side. If α and β are the same type of constituent (under some theory of types) then the construction is an instance of *self-embedding*. In contrast, right (or left) embedding (or branching) involves structures of the form

(5) Right embedding: $[_{\alpha} \ldots [_{\beta} \ldots]]$

where the material in β closes both the constituent β as well as α. Intuitively, the memory load difference is clear: in (4), after β is processed a memory of the first part of α is required to compute further relations within α. Formally, arbitrary center embeddings are precisely the structures that render grammars outside the scope of finite state (limited memory) automata (Chomsky and Miller, 1968).

In (5a) for example, a constituent *that the judge sentences* is center-embedded in the constituent *that the inmates . . . played* which is center embedded inside the constituent *the game. . . involved bats and balls.* Such *double* center embeddings are famously difficult to process and most people reject them as ungrammatical on first listening or reading (Miller and Isard, 1964; Hakes et al., 1976)

(6) a. *Center embedding*

The $\text{game}_{\text{N-}\{\text{LEVEL1}\}}$ [that the $\text{inmates}_{\text{N-level2}}$ [that the $\text{judge}_{\text{N-level}}$ $\text{sentenced}_{\text{V-level3}}$] $\text{played}_{\text{V-level2}}$] $\text{involved}_{\text{V-level1}}$ bats and balls.

b. *Right branching*

The judge$_{\text{N-level3}}$ sentenced$_{\text{V-level3}}$ the inmates$_{\text{N-level2}}$ who played$_{\text{N-level2}}$ the game$_{\text{N-level1}}$ that involved$_{\text{V-level1}}$ bats and balls.

The right-branching counterpart (6b) is processed easily even though the number of items to process is exactly the same in both types of sentences (Lewis, 1996; Yngve, 1960)). Note that the nouns and verbs have been labeled with level numbers 1, 2 and 3, where 3 is the most embedded level. This descriptive level reference will be used in the analyses below.

The study of Lewis and Vasishth (2005) showed that the cue-based retrieval theory can successfully account for the differential processing difficulty of center-embedded and right-branching sentences. Specifically, it showed that the processing time does not dramatically increase for deep right branching sentences at embedded verbs contrary to center embedded sentences. In addition, their model also simulated misanalyses of dependencies in center-embedded sentences that result in the elimination of the correct candidate for subsequent retrievals. In other words, failure to attaching the retrieval verb to the correct target can lead to the final verb being left without its correct unattached subject, generating further difficulty. For instance, when the level2 verb (*played* in (6a) is incorrectly attached to the level 1 noun (*game* in (6a)) rather than the correct level 2 noun (*inmates*), the correct dependent is not available for the final verb.

In the following analyses I examine attention metrics and surprisal at the verbs in the center and right-branching sentences from Stolz (1967). I also directly visualize the attention patterns of a single attention head (head$_{43}$) that is specialized for `nsubj` dependency. Although the current analyses cannot provide direct evidence of mis-retrievals or relations established in error, I provide indirect evidence of the possibility of such errors in the attention patterns associated with subject-verb head.

6.3.2 Methods

Four metrics were measured with GPT2-small at the cue (the verb): surprisal, attention to target that simply computes the amount of attention paid to the correct target from the cue measured using $\text{head}_{4,3}$, local attention entropy measured using $\text{head}_{4,3}$, and GDAE.

Attention distribution patterns are also visualized using Vig (2019)'s Transformer attention visualization tool. Additionally, in order to see whether early misattachments of an embedded verb and its subject might be possible in the center-embedded sentences, I also measured how much attention is paid to each noun in embedded sentences from all levels of verbs.

6.3.3 Materials

Table 6.4: An example set of materials used for the experiment on embedded sentence processing

Center-embedded	LEVEL1	The **game**$_{\text{target}}$ that the inmates that the judge sentenced played **involved**$_{\text{cue}}$ bats and balls.
	LEVEL2	The game that the **inmates**$_{\text{target}}$ that the judge sentenced **played**$_{\text{cue}}$ involved bats and balls.
	LEVEL3	The game that the inmates that the **judge**$_{\text{target}}$ **sentenced**$_{\text{cue}}$ played involved bats and balls.
Right-branching	LEVEL1	The judge sentenced the inmates who played the **game**$_{\text{target}}$ that **involved**$_{\text{cue}}$ bats and balls.
	LEVEL2	Yesterday evening, the judge sentenced the **inmates**$_{\text{target}}$ who **played**$_{\text{cue}}$ the game that ...
	LEVEL3	Before the sun went down, the **judge**$_{\text{target}}$ **sentenced**$_{\text{cue}}$ the inmates who ...

Fifteen sets of sentences from Stolz (1967) were included, with the same structure as in Table 6.4. In order to prevent confound effects from the word position, adverbial phrases are added at the beginning of right branching sentences at levels 2 and 3 (e.g., *'Yesterday evening'* or *'Before the sun went down'*). By doing so, I ensured that verbs of interest appear at the same position for both center-embedded and right-branching sentences.

6.3.4 Results

Consider first the visualization in Figure 6.7 of attention patterns at head_{43} for level 2-3 sentences. There is a striking contrast between center and right branching sentences in how sharply focused attention is allocated to the correct subject dependent. This is so even though the lexical items and thus semantic constraints are identical.

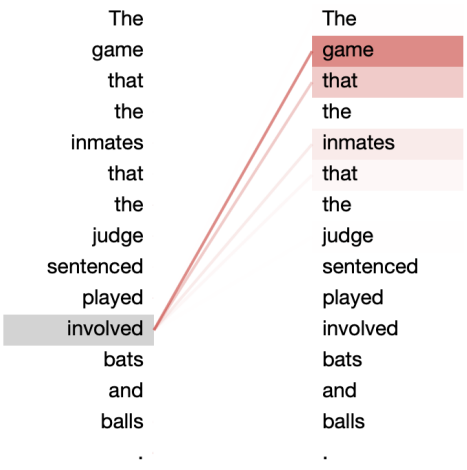
Figure 6.8 visualizes the amount of attention allocated to each noun from the level 2 (middle verb) and level 3 verbs. For example, in Figure 6.8a that in right branching sentences, the level 2 verb is allocating by far the most attention to the correct level 2 subject noun. But in the center-embedded sentences, the level 2 verb is allocating attention more uniformly across the three nouns, with the greatest amount of attention allocated to the noun that is actually the subject of the level 1 (main) verb.

Finally, Figure 6.9 shows four different quantities computed at each of the three verbs: surprisal, GDAE, the attention paid to the correct subject target by attention head_{43}, and the local attention entropy computed at attention head_{43}. Both attention entropy metrics are higher at the second and third verbs in center-embedded sentences, and attention to the correct subject noun is lower for the second and third verb in center-embedded sentences.

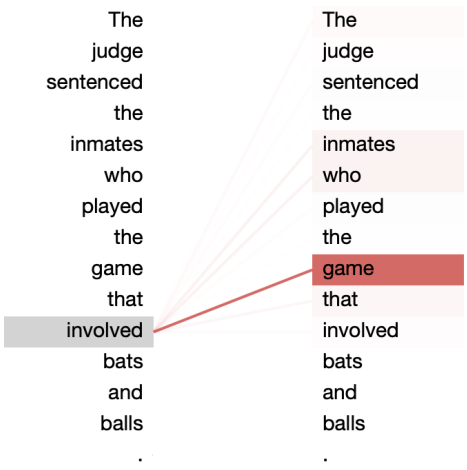
6.3.5 Discussion

These patterns are well aligned with the memory retrieval interference-based explanation about processing difficulty of center-embedded sentences (Lewis and Vasishth, 2005), which has a role for both increased similarity-based interference as well as incorrect attachments

of subjects and verbs. A plausible explanation of the increased surprisal at the final verb in the center embedded constructions is that the language model has computed a state in which the most likely subject dependent of the middle verb is the first noun phrase—making the continuation with a third verb more surprising. This also provides an account of the grammatical illusion identified by Janet Fodor many years ago: dropping the middle verb in double center embeddings leads to increased acceptability (Christiansen and MacDonald, 2009; Gibson and Thomas, 1996; Huang and Phillips, 2021).



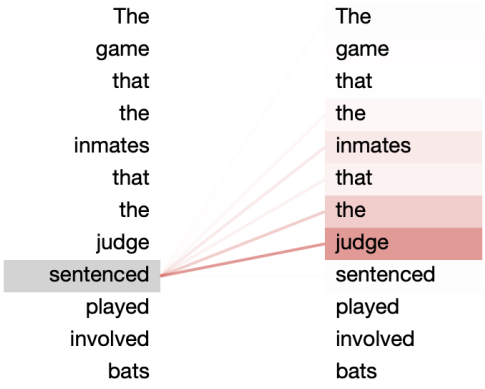
(a) Center-embedding, level 1 verb



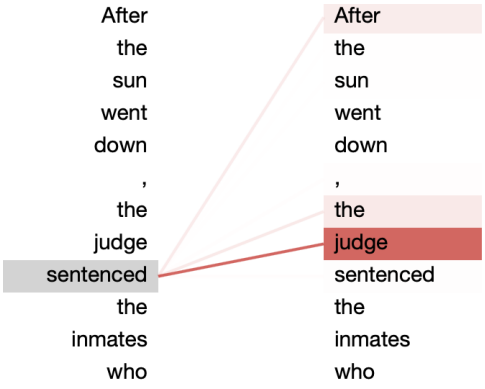
(b) Right-branching, level 1 verb

(c) Center-embedding, level 2 verb

(d) Right-branching, level 2 verb

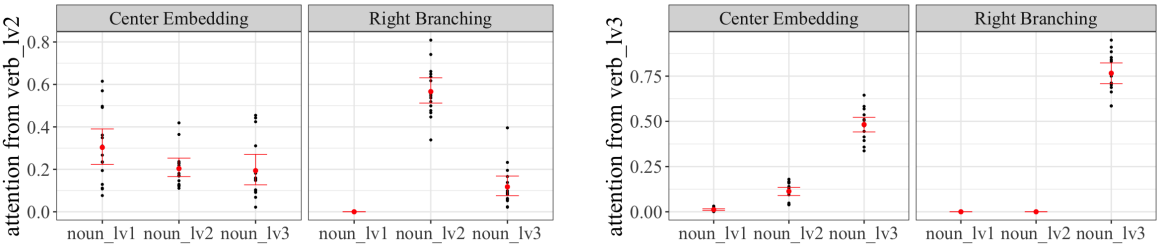


(e) Center-embedding, level 3 verb



(f) Right-branching, level 3 verb

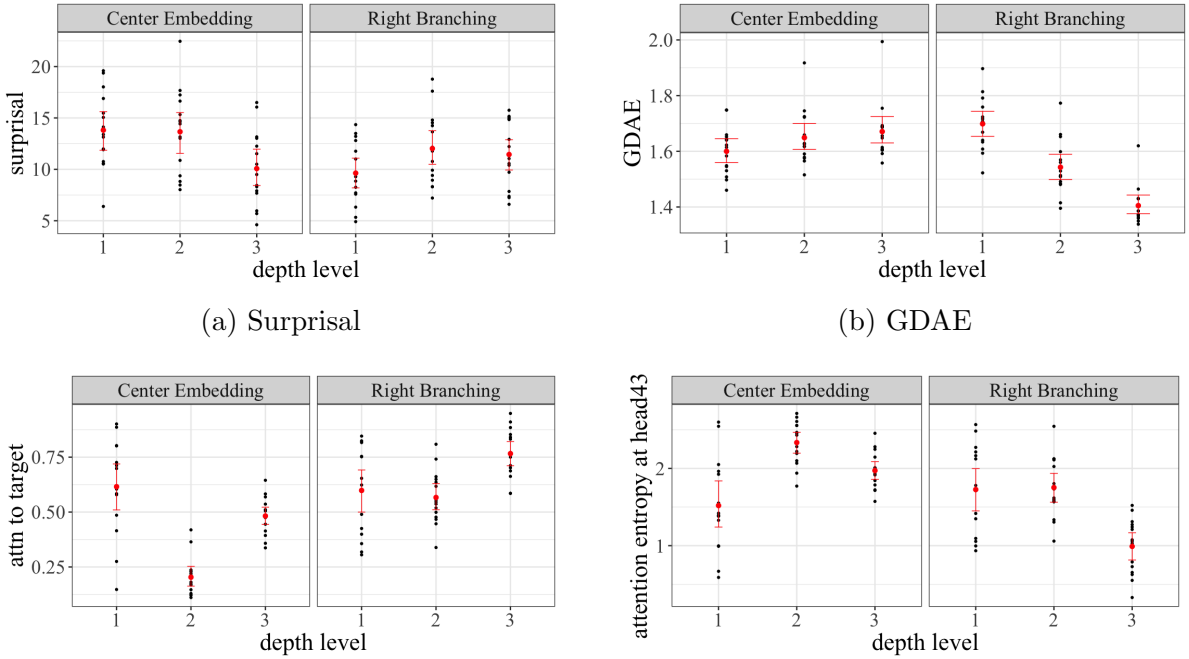
Figure 6.7: Attention distribution patterns observed at head_{43}, specialized for subject-verb relations. There is a striking contrast between right branching and center embedded structures in the degree to which attention is sharply focused attention on the correct subject at the two innermost verbs.



(a) Attention allocated from level 2 verb

(b) Attention allocated from level 3 verb

Figure 6.8: Attention allocated to nouns from the level 2 and level 3 verbs (*played* and *sentenced* respectively, in Table 6.4) by the self-attention head (head_{43}) that is specialized for building subject-verb relations. Attention is sharply allocated to the correct subject noun in the right branching sentences, but in the center-embedded sentences, the middle level 2 verb (*played* in Table 6.4) is diffusely allocating attention, with most attention on the incorrect subject (*game* in Table 6.4).



(a) Surprisal

(b) GDAE

(c) Attention allocated to correct subject noun by head_{43}

(d) Attention entropy at head_{43}

Figure 6.9: At verbs at three levels four metrics— *surprisals*, *GDAE*, *attention allocated to correct subject noun in* head_{43}, and *local attention entropy computed at* head_{43}, specialized for subejct-verb relations. Depth levels 1-3 indicate *involved*, *played*, and *sentenced* in Table 6.4 respectively.

6.4 Relative Clause Processing

6.4.1 Background

As Levy and Gibson (2013) have pointed out, the locus of difficulty in object-extracted clauses is useful in distinguishing expectation-based and memory-based theories. Given that subject-extracted relative clauses (SRCs) are more frequent than object-extracted clauses (ORCs), expectation-based theories predict that language comprehenders may anticipate a verb after processing *that*. Thus, the locus of the difficulty of ORCs predicted by expectation-based theories is the onset of the noun phrase (*the* in (7b)) since it is the place where the expectation towards a verb is unrealized, and surprisals computed from early PCFG-based models as well as recent large language models (including the analyses provided below) align with this intuitive prediction. But importantly, surprisal does *not* predict an increase in difficulty at the embedded verb of the ORC. In contrast, memory-based theories predict the locus of ORCs' processing difficulty is the verb in the relative clauses for a variety of reasons; for retrieval interference accounts the embedded clause verb (*attack* in (7b)) has two candidates as its dependent subject in ORCs (*reporter* and *senator* in (7b) while only one candidate in SRC (*reporter* in (7a), and two syntactic relations must be computed.

(7) a. *Subject Relative Clause (SRC)*

t The reporter$_{\text{i}}$ that t_{i} **attacked$_{\text{rcv}}$ the$_{\text{rcn}}$** senator admitted the error.

b. *Object Relative Clause (ORC)*

The reporter$_{\text{i}}$ that **the$_{\text{rcn}}$** senator **attacked$_{\text{rcv}}$** t_{i} admitted the error.

Most reading time studies have shown that the significantly greater processing difficulty differential between ORCs and SRCs is indeed found at the relative clause verb (Grodner and Gibson, 2005; Levy et al., 2013) , which is aligned with the prediction of memory-based theories. But Staub (2010) also found increased difficulty at the RC noun onset of the ORC,

where there were increased regressions. The maze task results of Vani et al. (2021) also showed that slowdowns were observed in the RC noun onset.

In the following analysis, both surprisal and GDAE will be examined at critical words (embedded verbs and noun onsets) in the SRC and ORC materials (Staub, 2010).

6.4.2 Methods

For each of the critical sentences from Staub (2010), I used GPT2-small to compute surprisal and GDAE at two key regions: the onset of the relative clause noun and the relative clause verb (see example (7) above). Local attention entropy was not included in this analysis, as the two regions involve different grammatical dependencies and would therefore require different attention heads. Using different heads would complicate interpretation and make it difficult to directly compare local attention entropy effects across regions.

6.4.3 Materials

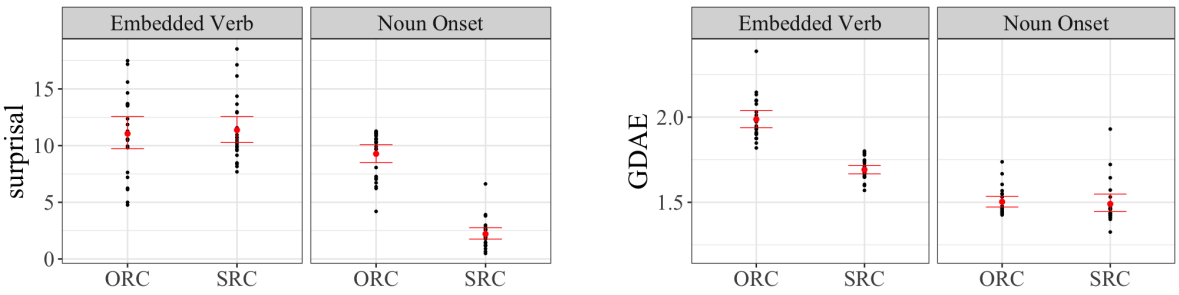
24 sets of sentences from Staub (2010) were included. In order to prevent confound effects from the position of the critical word, the example sentences were manipulated to have an adverbial phrase at the beginning of the sentences (e.g., *'yesterday'* or *'yesterday at noon'*). Different manipulations were required to control for the position of the relative clause noun onset and for the position of the relative clause verb. The manipulated structures are as in Table 6.5.

6.4.4 Results

As shown in Figure 6.10, greater surprisals of ORCs are observed at the onset of relative clause noun, but not the embedded verbs, consistent with earlier surprisal analyses (Levy and Gibson, 2013). But greater GDAEs of ORCs are observed at the relative clause verb, consistent with the assumption that greater retrieval interference occurs at the embedded

Table 6.5: An example set of manipulated materials used for the experiment on relative clause processing

	Materials used to compute difficulty metrics at the noun onset
SRC	Yesterday, the bus driver who followed *the* kids wondered about the location of a hotel.
ORC	Yesterday at noon, the bus driver who *the* kids followed wondered about the location of a hotel.
	Materials used to compute difficulty metrics at the embedded verb
SRC	Yesterday at noon, the bus driver who *followed* the kids wondered about the location of a hotel.
ORC	Yesterday, the bus driver who the kids *followed* wondered about the location of a hotel.



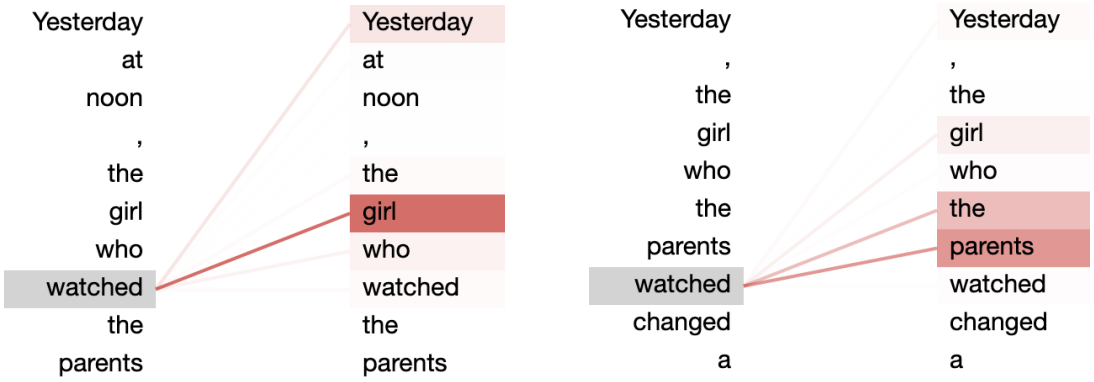
(a) Surprisal

(b) GDAE

Figure 6.10: *Surprisals* and *GDAEs* are measured at the noun onsets and at the embedded verbs (e.g., at *the* and *followed* in Table 6.5) in SRC and ORC sentences.

ORC verb.

Figure 6.11 illustrates how the attention at $\text{head}_{4,3}$ from the embedded verb is more diffuse in an object relative clause than in a subject relative clause. $\text{Head}_{4,3}$ is used because it is most likely to be one of the heads functionally important at the verb site because it computes subject-verb relations.



(a) Subject Relative Clause

(b) Object Relative Clause

Figure 6.11: Visualizing attention patterns at the embedded verb (watched) of object and subject relative clauses. Shown are the attention patterns of $\text{head}_{4,3}$ which is found to distribute attentions according to subject-verb relations. The distribution is more diffuse in the object relative clause at the embedded verb because attention is distributed to two noun phrases (girl, the, parents).

6.4.5 Discussion

The analysis of relative clause processing supports that attention entropy complements surprisal in explaining sentence processing difficulty by incorporating memory retrieval-based sentence processing difficulties. While surprisal accounts for increased difficulty at the noun onset in object relative clauses (ORCs), attention entropy explains this difficulty observed at the embedded verbs, consistent with cue-based retrieval theory. This divergence suggests that surprisal and attention entropy reflect distinct aspects of sentence processing, showing the promise of Transformers as integrative sentence processing models.

6.5 Garden Path Effects

6.5.1 Background

Garden path effects refer to the processing difficulty that arises when an initially preferred interpretation of an ambiguous phrase conflicts with the sentence's global structure. These

effects are caused by various sources of syntactic ambiguity. One common example involves ambiguity caused by a phrase that can be interpreted either as a noun phrase (NP) complement or as part of a sentential (S) complement—commonly referred to as NP/S ambiguity. In sentence (8a), the noun *walls* is initially interpreted as the direct object of the verb *maintained*. However, this local interpretation is inconsistent with the global structure, in which the word *walls* functions as the subject of the embedded clause.

Another type of garden path effect occurs with NP/Z ambiguity, where a verb can be either transitive or intransitive. As illustrated in example (9a), the verb *phoned* can take a direct object or not. When followed immediately by a noun, it is often interpreted as transitive, leading to a locally plausible but globally incorrect interpretation of the sentence structure.

(8) **NP/S Ambiguity**

a. *Ambiguous*

The worker maintained the **walls$_{\textbf{target}}$** **fell$_{\textbf{cue}}$** down in a heap before he arrived.

b. *Unambiguous*

The worker maintained that the **walls$_{\textbf{target}}$** **fell$_{\textbf{cue}}$** down in a heap before he arrived.

(9) **NP/Z Ambiguity**

a. *Ambiguous*

Even though the girl phoned the **instructor$_{\textbf{target}}$** **was$_{\textbf{cue}}$** very upset with her for missing a lesson.

b. *Unambiguous*

Even though the girl phoned, the **instructor$_{\textbf{target}}$** **was$_{\textbf{cue}}$** very upset with her for missing a lesson.

Garden path effects are typically measured by comparing behavioral (e.g., reading times) or neurological responses (e.g., EEG or fMRI signals) at critical regions. The critical regions refer to the points in an ambiguous sentence where the initially preferred interpretation is disconfirmed and reanalysis is required (e.g., verbs like *fell* in (8) or *was* in (9)). The magnitude of garden path effects varies depending on the type of ambiguity. Prior studies have found that NP/Z ambiguity elicits stronger garden path effects than NP/S ambiguity (Pritchett, 1988; Sturt et al., 1999).

Such differences in the magnitude of garden path effects have not been explained using surprisals computed from neural-network-based language models (Van Schijndel and Linzen, 2021; Huang et al., 2024). In what follows, I test whether attention-based metrics can account for these differences in garden path effect magnitudes, drawing on explanations grounded in cue-based retrieval theory.

Within the cue-based retrieval theory, I presume the amount of garden path effects can be explained by the degree of difference between the initial incorrect interpretation of the correct target and the interpretation that is prompted by the retrieval cue – the more local/incorrect initial interpretation of the target differs from what the cue is prompted to retrieve, the larger the garden path effects become. This is because a greater mismatch increases the processing demands during reanalysis.

This presumption can explain the magnitude difference between NP/S and NP/Z ambiguities. The disambiguator *fell* in (8a) prompts a retrieval that searches for its target with the following critical properties: [+NOUN, +NOM, + Governed by *maintain*], which required the target to be a subject nominal noun and be governed by the matrix verb *maintained*. The initial interpretation of the correct retrieval target *walls* contains properties: [+NOUN, +ACC, +Governed by *maintain*]. Given that the only property that the fails to satisfy is the property to be nominal (+NOM), it might cause relatively small amount of garden path effects.

In the case of the disambiguator *was* in (9a), it prompts retrieval of the target that is

compatible with the properties [+NOUN, +NOM, -Governed by *phoned*], and this should be satisfied by the target *instructor* in order for the globally correct interpretation. However, the initial interpretation of *instructor* barely matches the properties that are searched for by the cue because it is incorrectly interpreted to have the properties [+NOUN, +ACC, + Governed by *phone*]. Consequently, such incompatibility between the initial interpretation of the target and the cue in NP/Z is larger than in NP/S, which results in larger garden path effects compared to NP/S.

6.5.2 Methods

Four metrics were measured with GPT2-small at the disambiguating word (e.g., *fell* in (8a) and *was* in (9a)): surprisal, attention to target using $\text{head}_{4,3}$, local attention entropy measured using $\text{head}_{4,3}$, and GDAE.

Attention distribution patterns are also visualized using Vig (2019)'s Transformer attention visualization tool.

6.5.3 Matreials

20 sets of sentences for NP/S ambiguity and for NP/Z ambiguity that have the same structure as Table 6.6 are used. All materials are from Grodner et al. (2003). In order to prevent any confound effects from word position, I manipulated sentences to have the critical word (i.e., *fell* in (8) and *was* in (9)) at the same position for ambiguous and unambiguous conditions by having adverbial phrases at the beginning of sentences.

6.5.4 Results

Attention distribution patterns in Figure 6.12 illustrate how garden path effects are explained from the perspective of the cue-based retrieval theory: the less similarity between the initial interpretation of the target and the cue (also the disambiguator) causes more diffuse attention

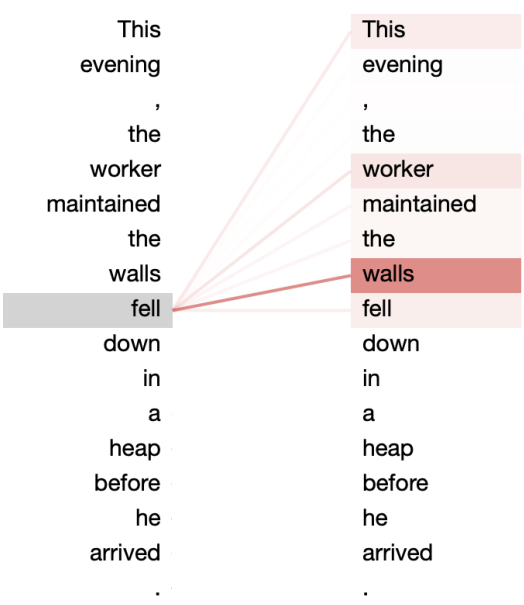
Table 6.6: An example set of manipulated materials used for the experiment on relative clause processing

	Materials used to compute difficulty metrics in NP/S sentence constructions
Ambiguous	This evening, the worker maintained the **walls$_{\text{target}}$ fell$_{\text{cue}}$** down in a heap before he arrived.
Unambiguous	Today, the worker maintained that the **walls$_{\text{target}}$ fell$_{\text{cue}}$** down in a heap before he arrived.
	Materials used to compute difficulty metrics in NP/Z sentence constructions
Ambiguous	This evening, even though the girl phoned the **instructor$_{\text{target}}$ was$_{\text{cue}}$** very upset with her for missing a lesson.
Unambiguous	Today, even though the girl phoned, the **instructor$_{\text{target}}$ was$_{\text{cue}}$** very upset with her for missing a lesson.

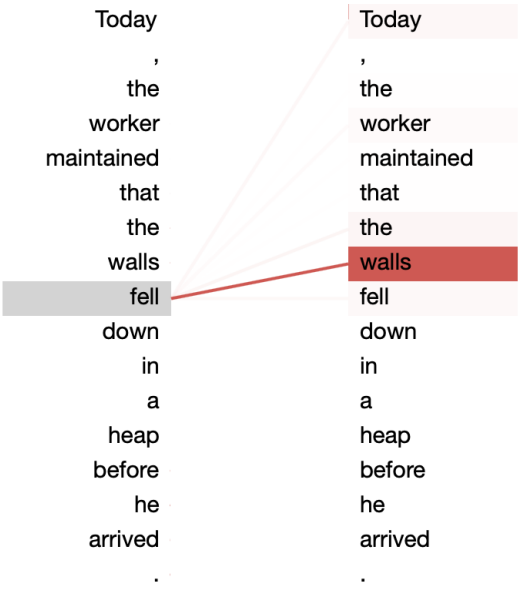
distribution. The NP/Z ambiguous condition shown in Figure 6.12c illustrates the greater amount of attention paid to the distractor (*girl*) than the amount of attention paid to the target (*instructor*) accounts for how the greater garden path effects can be caused in NP/Z condition.

Figure 6.13 shows the GPT2-based metrics computed at disambiguating cue verbs. Surprisals are greater in the ambiguous conditions both in NP/S and NP/Z. However, the difference in garden path effects by the type of ambiguity is not clearly shown in surprisals, similar to previous findings.

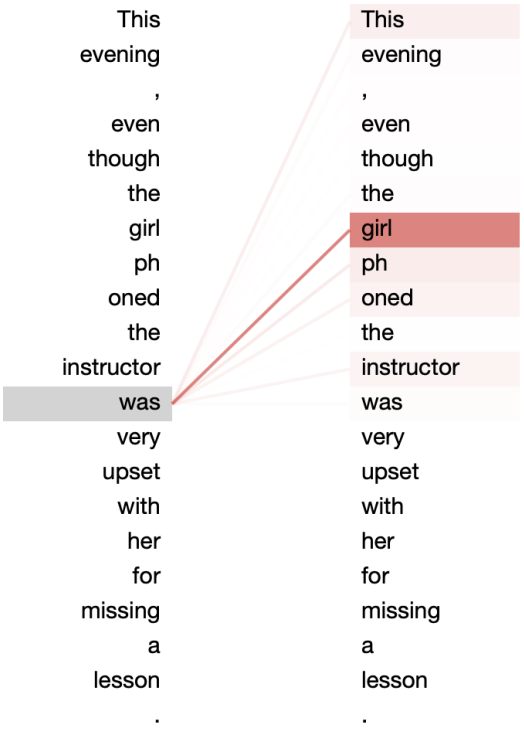
In contrast, attention-based metrics capture no or little of the ambiguity effects. Specifically, no significant garden path effects in NP/S construction explained with aggregate and local attention entropies. However, the magnitude difference of garden path effects by the type of ambiguity are explained with attention-related metrics. As shown in Figure 6.12c, the amount of attention paid to the correct target in the ambiguous condition is significantly



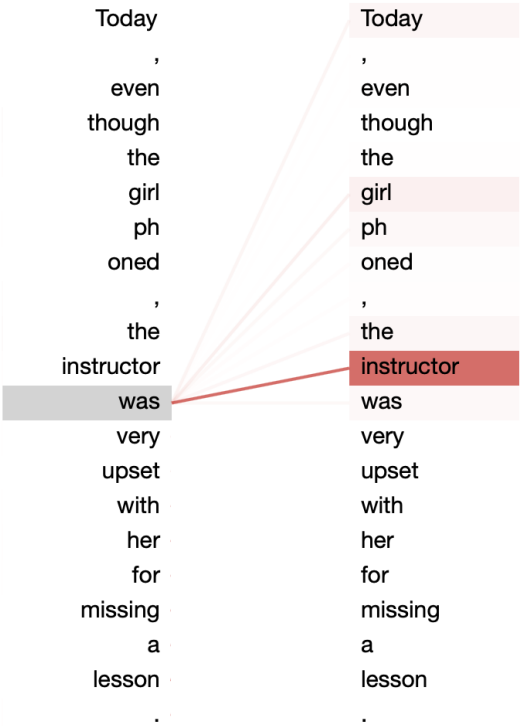
(a) NP/S, Ambiguous



(b) NP/S, Unambiguous

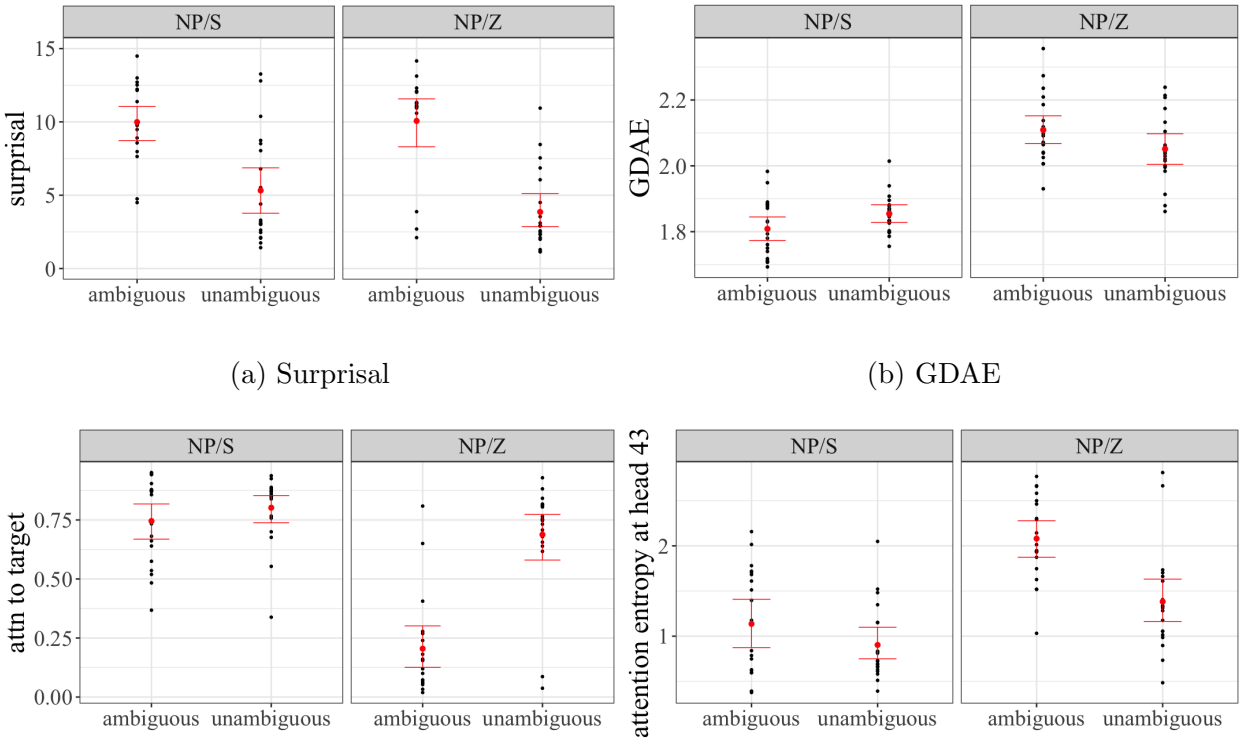


(c) NP/Z, Ambiguous



(d) NP/Z, Unambiguous

Figure 6.12: Visualizing attention patterns at disambiguating words. Shown are the attention patterns of $\text{head}_{4,3}$ which is found to distribute attentions sccording to subject-verb relations.



(a) Surprisal

(b) GDAE

(c) Attention allocated to correct target noun by $head_{43}$

(d) Attention entropy at $head_{43}$

Figure 6.13: At disambiguating verbs four metrics—*surprisals*, *GDAE*, *attention allocated to correct target in* $head_{43}$, and *local attention entropy computed at* $head_{43}$

greater in NP/S condition. Also, Figure 6.12d shows that attention entropies in the ambiguity condition estimated with $head_{4,3}$ are much greater in NP/Z than in NP/S, showing the different magnitude in the garden path effects could be explained with Transformers' attention-based metrics.

6.5.5 Discussion

As expected, surprisal values were higher for ambiguous sentences compared to their unambiguous counterparts, reflecting the increased processing difficulty during ambiguity resolution. However, consistent with prior research (Van Schijndel and Linzen, 2021; Huang et al., 2024), surprisals did not differentiate the magnitude of garden path effects between NP/S

and NP/Z constructions.

In contrast, attention-based metrics revealed a clear distinction by the type of ambiguity, even though they did not fully capture garden path effects. Specifically, attention entropy measured with $\text{head}_{4,3}$ was notably higher in NP/Z ambiguities, aligning with the greater reanalysis difficulty observed in human data. Furthermore, attention distribution patterns, as shown in Figure 6.12 and Figure 6.13c, demonstrated that in NP/Z cases, distractors received more attention than the correct targets, suggesting increased retrieval interference in NP/Z conditions.

These findings support the notion that Transformers' attention-based measures offer a complementary explanation for sentence processing difficulty, extending beyond what expectation-based metrics alone can account for. However, the inability of attention-based metrics to explain garden path effects, particularly in NP/S constructions, motivates further investigation.

CHAPTER 7

Summary and General Discussion

I have demonstrated that Transformers are a promising new foundation for integrative models of human sentence processing that seamlessly combine expectation-based and memory-based perspectives. In particular, Transformers can be viewed as predictive cue-based retrieval models that make next-word prediction through internal memory retrieval processes. In this chapter, I provide a summary of the main contributions and discuss directions that must be followed in future research.

7.1 Summary of Contributions

I illustrated that the computational architecture of the Transformer, particularly its key-query-value dot-product attention mechanism, embodies the core assumptions of cue-based retrieval models. Specifically, they share a common functional motivation: to provide a memory mechanism that is capable of handling long-distance dependencies in natural language, given sufficiently discriminating cues and distinct memory representations (Lewis and Vasishth, 2005; Lewis et al., 2006; Vaswani et al., 2017). Based on this functional similarity, in Chapter 3, I introduced a novel metric — attention entropy — which quantifies similarity-based interference effects using attention values from Transformer models. I also explained the process to select attention heads based on their grammatical roles, which can be used to compute local or global attention entropies.

In Chapter 4, I provided empirical evidence that attention entropy provides additional

explanation for sentence processing, independently of surprisal. To this end, I fit a Bayesian mixed-effects model to predict word-level reading times using data from the Natural Stories self-paced reading corpus (Futrell et al., 2021) and the GECO eye-tracking corpus (Cop et al., 2017). Predictors included attention entropy, surprisal, and a range of psycholinguistic variables such as word frequency, length, and position. The results showed that attention entropy explained additional variance in sentence processing difficulty beyond what was explained by surprisal. Interestingly, the locus of attention entropy effects differed across reading measures: in self-paced reading and go-past times, effects appeared at the target word; in eye-tracking first-fixation times, effects emerged in the spillover region. These patterns are consistent with the interpretation of attention entropy as a measure of memory integration cost.

In Chapter 5, I examined how the effect of attention entropy on reading times varied under different speed–accuracy trade-off conditions. The results revealed that the influence of attention entropy increased when accuracy was emphasized over speed, further supporting its role as an estimate of memory retrieval cost in explaining sentence processing times.

Building on this evidence, I applied Transformer-based integrative accounts to a set of psycholinguistic phenomena in Chapter 6: interference effects in subject–verb agreement, interference effects in non-agreement, processing of center-embedded vs. right-branching sentences, relative clause processing (subject vs. object relative clauses), and garden path effects. Using both qualitative and quantitative analyses, I demonstrated that these phenomena can be explained through the Transformer's integrative accounts using surprisals and attention-based metrics.

The results from the analyses suggest that Transformers offer a novel integrative account of psycholinguistic phenomena that may offer an alternative or complement to other approaches, such as entropy-reduction based accounts of asymmetries between subject and object relative clauses (Chen and Hale, 2021). Taken together, the findings and analyses in this thesis show that attention entropies provide empirical coverage in human sentence pro-

cessing that goes beyond what surprisal can account for. Combined with an understanding of the attention mechanism in Transformers, these findings support the idea that Transformers can provide mechanistic models of sentence processing that generate next-word predictions through cue-based memory retrieval operations.

7.2 Limitations and Future Directions

This thesis opens several avenues for future research that can deepen and broaden the use of Transformer-based models as integrative sentence processing models. In what follows, I outline key limitations and the directions for extending the current work by addressing the limitations.

These directions include refining methods for selecting attention heads to compute aggregate attention entropies, incorporating noisy memory representation alongside retrieval-based interference, developing models that account for the temporal dynamics of memory access, and improving generalizability across both naturalistic and controlled experimental settings.

7.2.1 Exploring Alternative Attention Head Selection Methods

As shown in Section 4.2, the selection of attention heads based on their grammatical function increased the predictive power of aggregate attention entropies for sentence processing times, compared to the aggregate attention entropies computed from the entire set of attention heads. However, the current selection method is not yet complete and there could be multiple methods to select attention heads to compute aggregate attention entropies. For example, Oh and Schuler (2022) proposed other types of Transformers' attention-based reading time estimation, based on entropy or distance that are computed from the attention distribution, and they selected a subset of attention heads by including heads only from the top-most layer of the GPT2-small model. This could be one of the approaches among many others, which

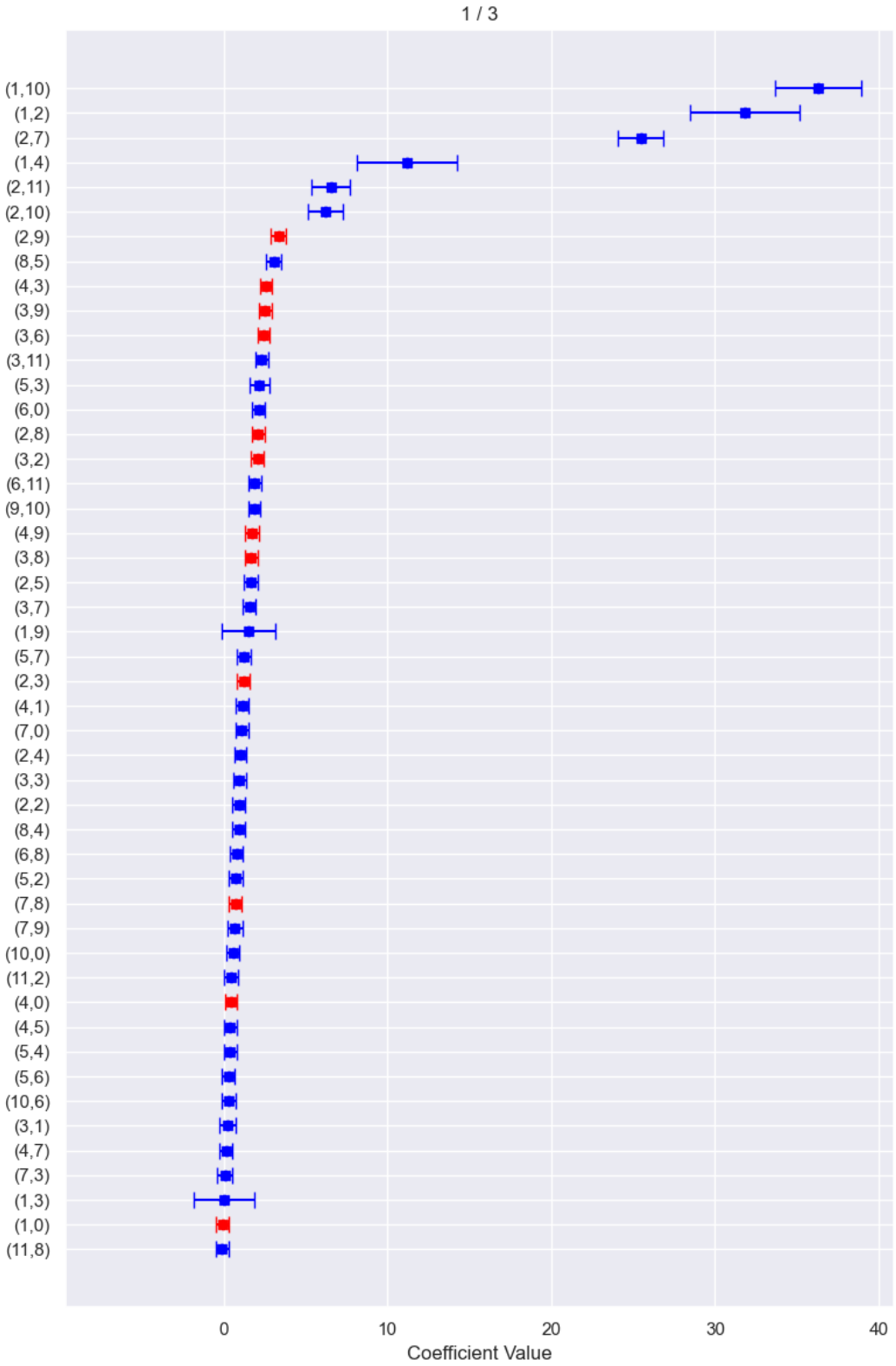
could be complementary to one another. This becomes manifest especially with the fact that none of the attention heads from the top-most layer was included in the 20 grammar-dependency-based set of attention heads[1].

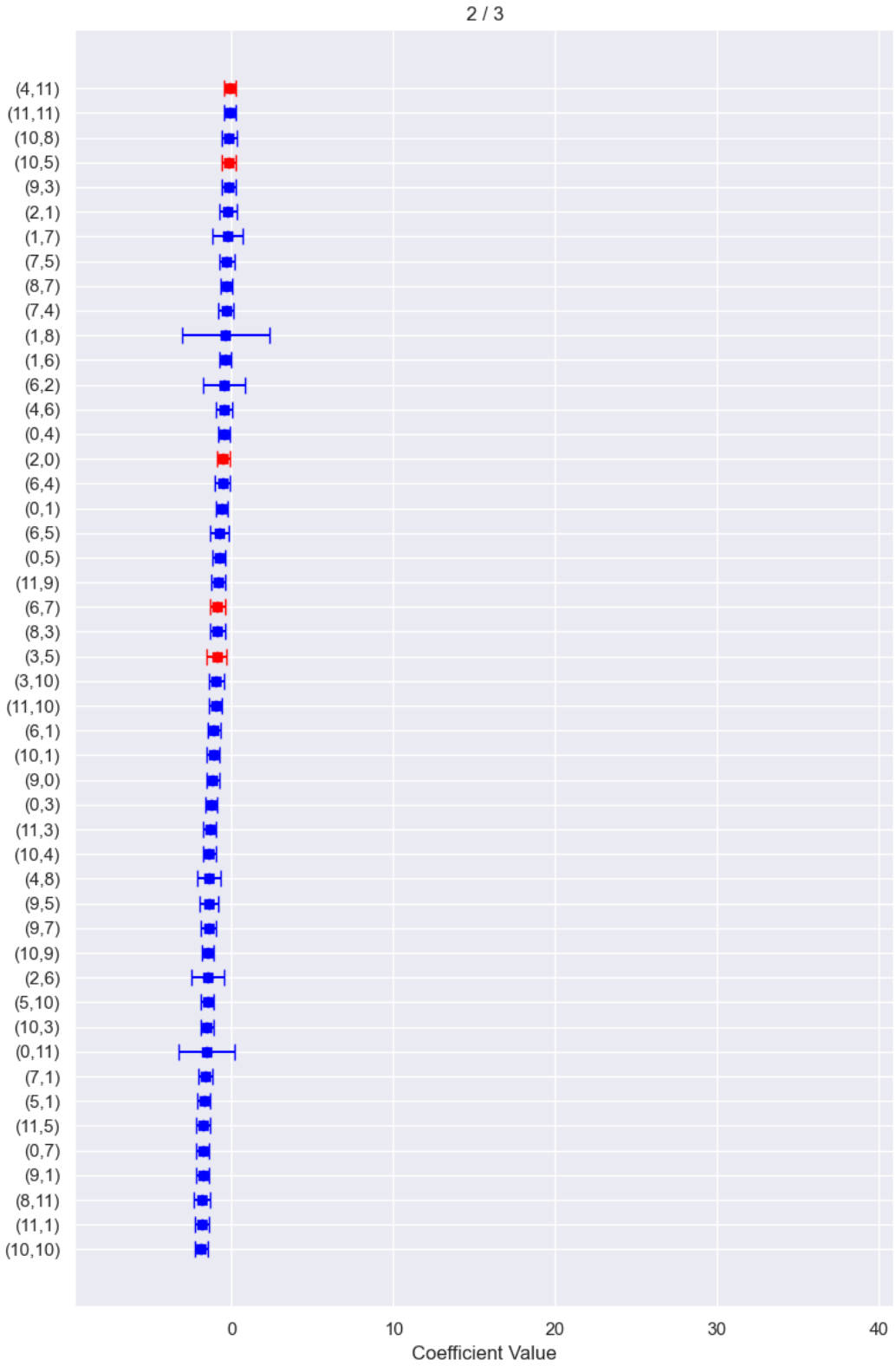
To see each individual attention head's power in explaining sentence reading times, I fit 144 frequentist models with the Equation 7.1 with attention entropies computed from attention entropies from all individual 144 attention heads in GPT2-small. All methods and material were the same as the method in Section 4.2 except that attention entropies are computed using a single attention head (i.e., local attention entropies). To see whether the local attention entropies from selected attention heads better explain human reading times than entropies from the non-selected attention heads, I compared the coefficients from models.

$$\begin{aligned} ReadingTimes \sim lm(Surprisal + LocalAttnEntropy + Frequency + WordLength + \\ SpilloverSurprisal + SpilloverLocalAttnEntropy + SpilloverFrequency + \\ SpilloverWordLength + Position) \end{aligned} \tag{7.1}$$

Figure 7.1 shows the results ordered by the magnitude of coefficients. Even though the entropies from the selected attention heads tend to have higher coefficients, it does not seem to be always true. For instance, the increase of the attention entropies computed from $\text{head}_{1,1}$, $\text{head}_{0,6}$ and $\text{head}_{4,2}$ have even negative coefficients. Additionally, coefficients from the attention heads selected from the top-most layers as in Oh and Schuler (2022) do not have high coefficients. Another interesting finding is that some attention heads that are not found to serve grammatical function show very high coefficients (e.g., $\text{head}_{1,10}$, $\text{head}_{1,2}$,

[1]This is consistent with findings that intermediate layers are computing representations of sentence structure (Hoover et al., 2019; Zini and Awad, 2022).


1 / 3
(1,10)
(1,2)
(2,7)
(1,4)
(2,11)
(2,10)
(2,9)
(8,5)
(4,3)
(3,9)
(3,6)
(3,11)
(5,3)
(6,0)
(2,8)
(3,2)
(6,11)
(9,10)
(4,9)
(3,8)
(2,5)
(3,7)
(1,9)
(5,7)
(2,3)
(4,1)
(7,0)
(2,4)
(3,3)
(2,2)
(8,4)
(6,8)
(5,2)
(7,8)
(7,9)
(10,0)
(11,2)
(4,0)
(4,5)
(5,4)
(5,6)
(10,6)
(3,1)
(4,7)
(7,3)
(1,3)
(1,0)
(11,8)
0
10
20
30
40
Coefficient Value


2 / 3
(4,11)
(11,11)
(10,8)
(10,5)
(9,3)
(2,1)
(1,7)
(7,5)
(8,7)
(7,4)
(1,8)
(1,6)
(6,2)
(4,6)
(0,4)
(2,0)
(6,4)
(0,1)
(6,5)
(0,5)
(11,9)
(6,7)
(8,3)
(3,5)
(3,10)
(11,10)
(6,1)
(10,1)
(9,0)
(0,3)
(11,3)
(10,4)
(4,8)
(9,5)
(9,7)
(10,9)
(2,6)
(5,10)
(10,3)
(0,11)
(7,1)
(5,1)
(11,5)
(0,7)
(9,1)
(8,11)
(11,1)
(10,10)
0
10
20
30
40
Coefficient Value

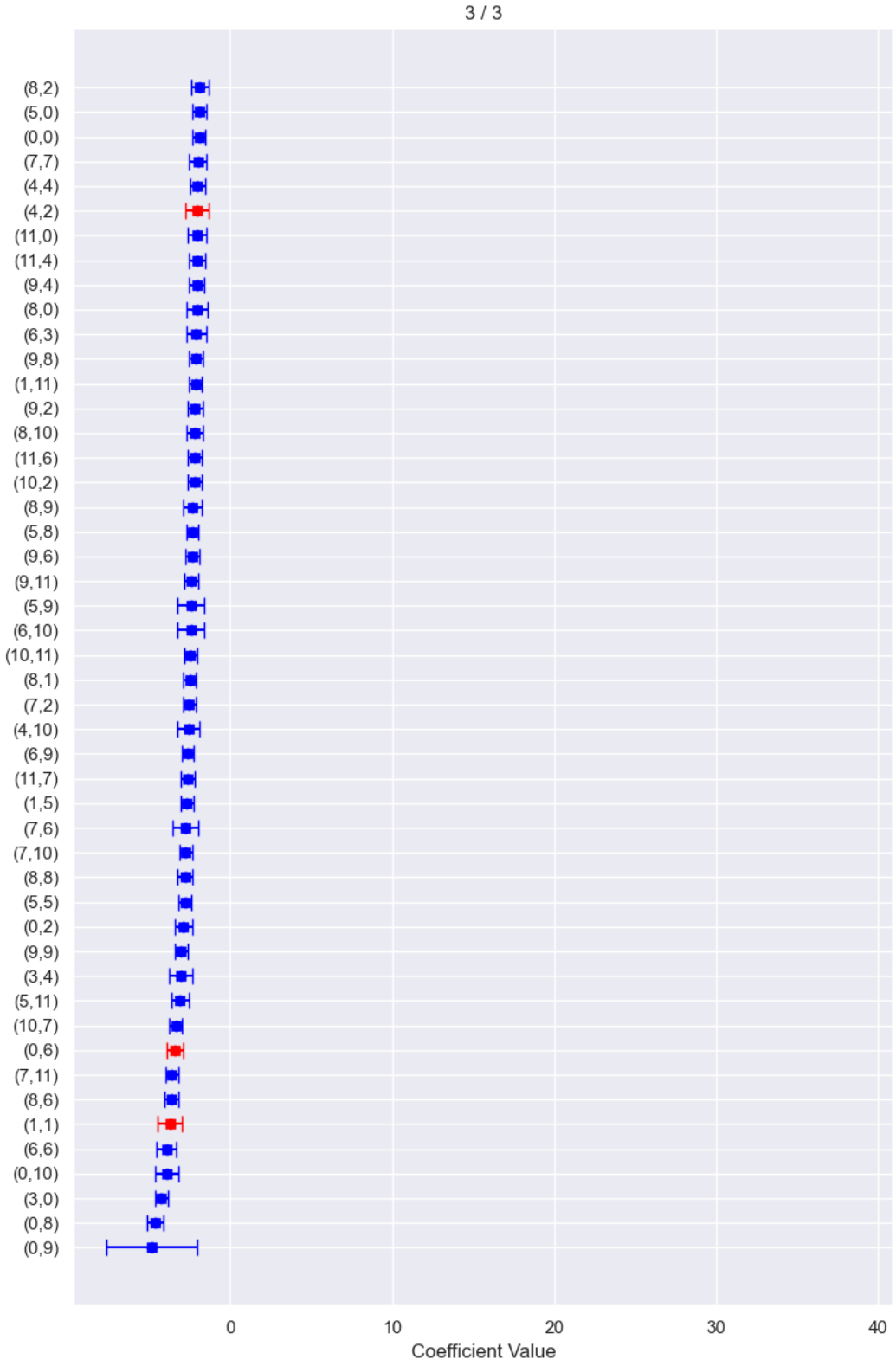


Figure 7.1: Caterpillar plots of attention entropy coefficient estimates with 95% confidence intervals for. Each point represents each head positioned at (layer, head). Red markers indicate attention heads that were selected as ones that process grammar dependencies. Horizontal error bars represent the 95% confidence intervals.

$\text{head}_{2,7}$, $\text{head}_{1,4}$, $\text{head}_{2,11}$, $\text{head}_{2,10}$). These results together suggest that further exploration of attention head selection is necessary.

7.2.2 Incorporating Noisy Memory

Another important direction for extending the work in this thesis is to integrate a more complete representation of working memory costs into the Transformer-based framework. As briefly discussed in Chapter 2, there are two distinct particular components of working memory cost in sentence processing, according to Gibson (1998): *memory cost*, which reflects the effort required to maintain elements in memory over a distance, and *integration cost*, which reflects the difficulty of retrieving and integrating those elements when needed.

The attention entropy metric developed in this thesis captures the latter—*integration cost*—by quantifying memory retrieval interference using attention values. However, to build a more comprehensive and cognitively grounded model that combines expectation-based and memory-based accounts of processing, it is important to incorporate *memory cost* as well.

Although *memory cost* was not modeled in this thesis, doing so within the Transformer framework is likely feasible. One promising approach is to incorporate the idea of lossy or noisy memory representations, such as in the lossy-context surprisal model (Futrell et al., 2020) or in resource-rational models of memory decay (Hahn et al., 2022). These approaches assume that memory representations degrade over time or with interference, leading to less reliable access to prior context. In a Transformer, this could be implemented by perturbing or compressing past representations as a function of their distance from the current word or their contextual relevance.

By combining attention entropy (as a measure of integration cost) with mechanisms for lossy memory (as a proxy for memory cost), future models could more fully simulate the dual pressures that shape sentence processing. Such a hybrid approach would offer a computational realization of the core insights from Dependency Locality Theory within a modern neural architecture, advancing our understanding of the interplay between memory con-

straints and predictive processing in real-time language comprehension.

7.2.3 Integrating Explicit Memory Dynamics

The similarities between cue-based retrieval parsers and the Transformer's attention mechanism enable the interpretation of Transformers as novel mechanistic models of sentence processing, conceptualizing sentence processing as the generation of predictions through cue-based memory retrieval. However, this interpretation is limited in its ability to explain why increased attention entropy leads to longer reading times. Furthermore, the analyses presented here fall short of demonstrating exactly how patterns of attention distribution influence next-token prediction.

One potential direction for addressing this limitation is to incorporate an explicit computational model of the *dynamics* of memory access. Transformer models compute all attention operations in a single feed-forward pass at each word; they do not inherently offer a mechanistic account of how memory unfolds over time. By introducing noise and an internal sequential evidence accumulation mechanism, it would be possible to simulate how the attention distribution dynamically affects prediction.

For example, multiple memory retrievals could be integrated over time until a certain quality threshold is reached (Shadlen and Shohamy, 2016). This threshold may be related to attention entropy. Such an approach would be conceptually similar to models that perform sequential sampling of noisy perceptual evidence (e.g., word identity) until a belief threshold is met (Norris, 2006; Wald and Wolfowitz, 1948). This could provide a mechanistic basis for observed linear effects of surprisal. In this way, a Transformer could be extended into a bounded rational comprehender—bounded by perceptual and memory noise, and rationally adapted to those bounds and to speed-accuracy tradeoffs imposed by task demands (Lewis et al., 2014).

7.2.4 Addressing Challenges in Reading Time Estimation with Fixed Parameters

Another limitation is the lack of a single model, with fixed parameters, that can explain quantitative reading-time contrasts across both naturalistic corpora (e.g., Natural Stories and GECO in Chapter 4) and controlled experimental materials (in Chapter 6). This challenge is not unique to the attention entropy framework, but remains a broader issue in the field (Huang et al., 2024).

One potential reason for this difficulty lies in the modeling framework used to estimate these effects. In Chapters 4 and 5, I used linear mixed-effects models, which assume linearity, homoscedasticity, and other statistical properties. However, human behavioral responses—such as reading times—can vary non-linearly across time and stimuli. This mismatch can make it difficult for such models to accurately capture the full scope of variation in naturalistic reading behavior.

Future work should therefore take greater care in selecting model structures. More flexible alternatives—such as continuous-time deconvolutional regression (Shain and Schuler, 2018; Shain, 2021)—could offer a promising solution. These models are less constrained by assumptions of linearity and may better capture the temporal dynamics of language comprehension.

7.3 Conclusion

The findings in this thesis offer compelling evidence that Transformer models provide a promising foundation for richer models of human sentence processing that are both cognitively motivated and linguistically competent. In particular, the Transformer's attention mechanism can be functionally interpreted as a predictive cue-based retrieval model, making it a useful tool for modeling working memory-related aspects of language processing while

also capturing aspects of expectation-based accounts.

Based on the properties of Transformer's attention mechanism as a cue-based retrieval model, I proposed a novel *attention entropy* metric for quantifying similarity-based interference effects in sentence processing. Through empirical analyses of naturalistic reading time corpora, I showed that attention entropy accounts for variance in sentence processing times that cannot be explained by surprisal alone. Moreover, its predictive power increases under conditions that emphasize comprehension accuracy over speed, further supporting its interpretation as a signal of memory retrieval cost.

To show how the combination of surprisal and attention entropy provides an integrated account of psycholinguistic phenomena, I applied this framework to a diverse set of cases involving memory interference effects—such as subject–verb agreement interference, non-agreement attraction effects, self-embedded sentence processing, relative clause processing, and garden-path processing. These case studies further illustrated how expectation-based and memory-based accounts can be integrated with Transformers in order to account for human sentence processing.

By linking the internal processing mechanisms in Transformers that reflect working memory operations with the probabilistic distributions from which the model generates its predictions, this work opens new directions for building integrative models that combine expectation-based and memory-based accounts to explain human sentence processing.

APPENDIX A

Identification of Heads Specialized for a Grammar Dependency

In Section 3.5.1, I introduced a method to select attention heads that process a certain type of grammar dependency following the method introduced in Voita et al. (2019). Table A.1 includes the entire results of the selection based on the criteria explained in Section 3.5.1, although I only used attention heads with the highest accuracy for each grammar dependency when computing Grammar Dependeny Attention Entropies.

Table A.1: Results from the head selection analysis

Dependency Type	Frequency-based Accuracy	Attention-based Accuracy
acl	2, 40.64%	(2,9), 70.18%
		(3,7), 53.37%
acl:relcl	2, 29.65%	(6,7), 42.74%
		(6,1), 36.64%
		(7,8), 35.64%
		(4,6), 35.14%
		(2,9), 35.10%
		(4,3), 34.87%

Table A.1 (continued)

Dependency Type	Frequency-based Accuracy	Attention-based Accuracy
		(3,7), 34.06%
advcl	2, 10.76%	(3,5), 20.42%
		(0,0), 17.35%
		(3,4), 17.31%
		(4,9), 16.78%
		(6,3), 16.64%
		(4,6), 14.81%
		(5,9), 14.51%
		(3,9), 14.43%
		(0,6), 14.43%
		(4,8), 14.32%
		(1,5), 12.32%
advmod	-1, 42.55%	(4,11), 53.81%
		(3,11), 52.77%
		(6,8), 50.15%
		(5,6), 49.52%
amod	-1, 78.72%	None
appos	2, 31.05%	None
aux	-1, 52.39%	(3,8), 74.04%
		(3,9), 69.71%
		(1,0), 62.43%
		(2,2), 61.04%
		(2,4), 59.95%

Table A.1 (continued)

Dependency Type	Frequency-based Accuracy	Attention-based Accuracy
		(2,8), 58.07%
aux:pass	-1, 88.00%	(2,8), 97.09%
case	-2, 41.50%	(2,0), 85.98%
		(3,8), 82.54%
		(4,0), 75.53%
		(2,8), 73.23%
		(5,4), 68.10%
		(2,5), 53.73%
		(2,2), 52.20%
		(1,2), 51.92%
		(8,7), 50.79%
		(3,3), 46.97%
		(6,8), 46.44%
		(6,0), 46.35%
		(7,4), 46.23%
		(6,11), 46.02%
cc	-1, 42.74%	(3,8), 65.92%
		(4,1), 65.40%
		(2,5), 61.12%
		(2,4), 56.28%
		(2,0), 55.94%
		(1,1), 55.28%
		(2,6), 49.83%

Table A.1 (continued)

Dependency Type	Frequency-based Accuracy	Attention-based Accuracy
cc:preconj	-1, 37.40%	(10,5), 74.05%
		(4,1), 73.28%
		(8,0), 63.36%
		(9,10), 62.60%
		(3,11), 53.44%
		(6,5), 53.44%
		(3,1), 48.09%
		(6,11), 44.27%
		(8,2), 41.22%
ccomp	3, 17.53%	(4,2), 29.65%
		(5,3), 27.24%
		(4,1), 22.63%
		(4,6), 21.69%
		(6,2), 21.38%
		(0,0), 20.93%
		(1,5), 19.63%
compound	-1, 86.22%	None
compound:prt	1, 84.65%	(1,0), 95.27%
		(2,4), 94.98%
		(2,8), 94.59%
		(3,9), 93.34%
conj	2, 30.44%	None
cop	-1, 36.78%	(2,8), 82.98%

Table A.1 (continued)

Dependency Type	Frequency-based Accuracy	Attention-based Accuracy
		(3,9), 73.14%
		(3,8), 63.91%
		(4,0), 61.25%
		(2,5), 59.12%
		(1,1), 50.36%
		(6,0), 48.79%
		(5,4), 46.66%
		(8,7), 44.62%
		(7,0), 42.75%
csubj	2, 49.81%	(3,2), 56.81%
det	-1, 63.22%	(2,3), 81.92%
		(3,2), 80.18%
det:predet	-2, 70.66%	None
discourse	-3, 19.53%	None
expl	-1, 65.51%	None
fixed	1, 89.55%	None
flat	1, 92.09%	None
goeswith	-1, 100%	None
iobj	1, 88.08%	(4,0), 97.93%
		(5,4), 97.93%
list	2, 30.00%	(0,6), 41.67%

Table A.1 (continued)

Dependency Type	Frequency-based Accuracy	Attention-based Accuracy
mark	-1, 47.33%	(2,0), 63.09%
		(3,8), 62.42%
nmod	3, 38.20%	None
nmod:npmod	-1, 73.15%	None
nmod:poss	-1, 62.55%	(3,6), 75.08%
		(2,3), 73.95%
		(3,2), 70.36%
nmod:tmod	2, 30.00%	(7,8), 50.00%
		(9,3), 50.00%
		(6,0), 47.50%
		(10,9), 47.50%
		(2,9), 47.50%
		(9,0), 47.50%
		(11,11), 45.00%
		(8,7), 45.00%
		(11,8), 45.00%
		(6,7), 45.00%
		(6,1), 42.50%
		(11,3), 42.50%
		(7,0), 42.50%
		(4,3), 42.50%
		(9,10), 42.50%
		(5,7), 40.00%

Table A.1 (continued)

Dependency Type	Frequency-based Accuracy	Attention-based Accuracy
		(8,5), 40.00%
		(3,7), 40.00%
		(5,4), 37.50%
		(3,5), 37.50%
		(11,10), 37.50%
		(0,6), 37.50%
		(1,7), 35.00%
		(8,4), 35.00%
		(4,6), 35.00%
nsubj	-1, 40.63%	(4,3), 56.85%
		(6,0), 47.13%
		(3,6), 46.30%
		(2,9), 46.27%
nsubj:pass	-2, 39.59%	(4,3), 67.63%
		(3,7), 53.90%
		(2,9), 52.12%
nummod	-1, 54.27%	(3,6), 71.54%
		(1,0), 66.04%
		(5,6), 65.88%
		(4,11), 63.30%
		(6,8), 61.46%
		(0,7), 60.47%
obj	2, 37.98%	(2,8), 84.15%

Table A.1 (continued)

Dependency Type	Frequency-based Accuracy	Attention-based Accuracy
		(4,0), 81.57%
		(3,9), 76.11%
		(5,4), 73.74%
		(3,8), 69.99%
		(6,11), 59.23%
		(5,7), 55.01%
		(8,7), 54.01%
		(9,10), 53.31%
		(8,4), 51.21%
		(11,11), 50.38%
		(10,5), 47.59%
		(3,11), 47.19%
		(3,3), 45.26%
		(7,4), 44.93%
		(11,10), 44.80%
		(7,8), 44.55%
		(2,9), 44.44%
		(4,6), 43.44%
		(10,11), 45.26%
obl	3, 24.24%	(4,9), 35.17%
		(3,9), 28.98%
obl:npmod	-1, 62.43%	(2,9), 82.01%
		(5,6), 70.90%

Table A.1 (continued)

Dependency Type	Frequency-based Accuracy	Attention-based Accuracy
		(2,4), 29.58%
obl:tmod	2, 16.07%	(4,9), 54.29%
		(6,11), 39.64%
		(7,8), 36.79%
		(3,5), 35.71%
		(2,8), 35.36%
		(5,7), 35.18%
		(9,10), 33.57%
		(11,11), 33.39%
		(5,3), 30.89%
		(10,9), 30.36%
		(3,9), 30.18%
		(11,10), 28.93%
		(7,4), 28.57%
		(8,4), 28.39%
		(4,6), 27.68%
		(6,3), 26.43%
		(10,5), 26.43%
		(4,3), 25.89%
		(3,8), 25.18%
		(7,5), 25.00%
		(3,7), 24.46%
		(7,3), 24.46%
		(0,2), 24.11%

Table A.1 (continued)

Dependency Type	Frequency-based Accuracy	Attention-based Accuracy
		(8,5), 23.75%
		(2,9), 22.68%
		(6,7), 22.50%
		(8,7), 22.50%
		(3,11), 22.14%
		(10,11), 21.61%
		(0,0), 21.25%
		(11,8), 20.89%
		(7,0), 20.71%
		(7,9), 20.71%
		(0,6), 20.54%
		(11,3), 20.00%
		(11,7), 19.29%
		(1,7), 18.57%
		(6,8), 18.21%
		(9,3), 18.21%
		(4,0), 18.04%
parataxis	3, 13.64%	(3,5), 20.00%
		(0,0), 17.70%
		(3,4), 16.83%
		(5,9), 15.86%
		(0,6), 15.86%
punctuation	-1, 16.57%	(1,1), 37.02%
		(2,6), 36.05%

Table A.1 (continued)

Dependency Type	Frequency-based Accuracy	Attention-based Accuracy
		(2,5), 35.95%
		(9,11), 35.58%
		(10,2), 33.02%
		(10,2), 31.84%
		(10,4), 31.03%
		(4,11), 30.29%
		(11,2), 29.91%
		(1,3), 29.33%
		(3,7), 28.99%
		(7,11), 28.42%
		(1,7), 28.12%
		(9,3), 27.19%
		(3,3), 26.42%
		(7,0), 25.42%
		(11,4), 25.34%
		(5,6), 24.68%
		(3,1), 24.58%
		(11,11), 24.58%
		(4,1), 24.33%
		(5,4), 24.31%
		(10,10), 24.24%
		(5,3), 24.02%
		(8,5), 23.68%
		(6,1), 23.55%

Table A.1 (continued)

Dependency Type	Frequency-based Accuracy	Attention-based Accuracy
		(8,6), 23.30%
		(8,2), 23.20%
		(11,0), 23.13%
		(6,8), 22.31%
		(3,8), 22.25%
		(11,3), 22.02%
		(4,6), 21.68%
		(4,2), 21.65%
		(8,3), 21.48%
		(8,0), 21.19%
		(5,2), 20.94%
		(1,2), 20.92%
		(7,9), 20.77%
		(2,9), 20.73%
		(11,10), 20.68%
		(11,9), 20.48%
		(10,7), 20.47%
		(8,11), 20.32%
		(9,6), 20.22%
		(6,7), 19.95%
		(5,8), 19.24%
		(8,7), 19.10%
		(7,8), 19.08%
		(9,9), 18.97%

Table A.1 (continued)

Dependency Type	Frequency-based Accuracy	Attention-based Accuracy
		(3,2), 18.80%
		(6,5), 18.79%
		(1,8), 18.54%
		(7,7), 18.27%
root	0, 100.00%	None
vocative	2, 26.32%	None
xcomp	2, 54.90%	(3,9), 75.04%

BIBLIOGRAPHY

Alcott, L. M. (1915). *Louisa May Alcott: A Child's Biography*. Mary Stoyell Stimpson (Excerpt from A Child's Book of American Biography). Accessed from http://www.geocities.com/area51/rampart/2627/fenklpage.html.

Andersen, H. C. (1855). *The Money-Box*. Original Publication. Accessed from various public domain sources.

Baker, C. L. (1995). *English syntax*. Mit Press.

Bird, S. (2006). Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72.

Boyce, V. and Levy, R. (2023). A-maze of natural stories: Comprehension and surprisal in the maze task. *Glossa Psycholinguistics*, 2(1).

Brennan, J. R., Dyer, C., Kuncoro, A., and Hale, J. T. (2020). Localizing syntactic predictions using recurrent neural network grammars. *Neuropsychologia*, 146:107479.

Bürkner, P.-C. (2017). brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, 80:1–28.

Caplan, D. and Waters, G. S. (1990). Short-term memory and language comprehension: A critical review of the neuropsychological literature.

Carpenter, P. (1989). The role of working memory in language comprehension. *Complex information processing: The impact of Herbert Simon*, pages 31–68.

Chang, J. P., Chiam, C., Fu, L., Wang, A. Z., Zhang, J., and Danescu-Niculescu-Mizil, C. (2020). Convokit: A toolkit for the analysis of conversations. *arXiv preprint arXiv:2005.04246*.

Chen, Z. and Hale, J. T. (2021). Quantifying structural and non-structural expectations in relative clause processing. *Cognitive Science*, 45(1):e12927.

Chomsky, N. (1965). Aspects of the theory of syntax.

Chomsky, N. (1970). Remarks on nominalization. *Readings in English transformational grammar/Ginn and Company*.

Chomsky, N. (1980). Rules and representations. *Behavioral and brain sciences*, 3(1):1–15.

Chomsky, N. and Miller, G. A. (1968). Introduction to the formal analysis of natural languages. *Journal of Symbolic Logic*, 33(2).

Christiansen, M. H. and MacDonald, M. C. (2009). A usage-based approach to recursion in sentence processing. *Language Learning*, 59:126–161.

Contreras Kallens, P., Kristensen-McLachlan, R. D., and Christiansen, M. H. (2023). Large language models demonstrate the potential of statistical learning in language. *Cognitive Science*, 47(3):e13256.

Cop, U., Dirix, N., Drieghe, D., and Duyck, W. (2017). Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior research methods*, 49(2):602–615.

Cunnings, I. and Sturt, P. (2018). Retrieval interference and semantic interpretation. *Journal of Memory and Language*, 102:16–27.

Demberg, V. and Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Demberg, V. and Keller, F. (2009). A computational model of prediction in human parsing: Unifying locality and surprisal effects. In *Proceedings of the annual meeting of the cognitive science society*, volume 31.

Dillon, B., Mishler, A., Sloggett, S., and Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, 69(2):85–103.

Ehrlich, S. F. and Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of verbal learning and verbal behavior*, 20(6):641–655.

Engbert, R., Nuthmann, A., Richter, E. M., and Kliegl, R. (2005). Swift: a dynamical model of saccade generation during reading. *Psychological review*, 112(4):777.

Fenkl, H. I. (2000). *The Goblin's Club*. Bo-Leaf Books. Translated by Heinz Insu Fenkl.

Frank, M. C. (2023). Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*, 27(11):990–992.

Futrell, R., Gibson, E., and Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive science*, 44(3):e12814.

Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S. T., and Fedorenko, E. (2021). The natural stories corpus: a reading-time corpus of english texts containing rare syntactic constructions. *Language Resources and Evaluation*, 55(1):63–77.

Futrell, R. and Mahowald, K. (2025). How linguistics learned to stop worrying and love the language models. *arXiv preprint arXiv:2501.17047*.

Futrell, R., Wilcox, E., Morita, T., and Levy, R. (2018). Rnns as psycholinguistic subjects: Syntactic state and grammatical dependency. *arXiv preprint arXiv:1809.01329*.

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., and Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. *arXiv preprint arXiv:1903.03260*.

Gao, R. and Yu, C.-L. (In prep). Humans consider extensive prior contexts during natural reading: An eye-tracking examination with the geco dataset. In preparation.

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

Gibson, E. et al. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, 2000:95–126.

Gibson, E. and Thomas, J. (1996). The processing complexity of english center-embedded and self-embedded structures. In *Proceedings of the NELS 26 workshop on language processing: MIT Working Papers in Linguistics*.

Gibson, E. A. F. (1991). *A computational theory of human linguistic processing: Memory limitations and processing breakdown*. Carnegie Mellon University.

Gordon, P. C., Hendrick, R., and Johnson, M. (2001). Memory interference during language processing. *Journal of experimental psychology: learning, memory, and cognition*, 27(6):1411.

Grodner, D. and Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentenial complexity. *Cognitive science*, 29(2):261–290.

Grodner, D., Gibson, E., Argaman, V., and Babyonyshev, M. (2003). Against repair-based reanalysis in sentence comprehension. *Journal of Psycholinguistic Research*, 32(2):141–166.

Hahn, M., Futrell, R., Levy, R., and Gibson, E. (2022). A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, 119(43):e2122602119.

Hakes, D. T., Evans, J. S., and Brannon, L. L. (1976). Understanding sentences with relative clauses. *Memory & Cognition*, 4(3):283–290.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Hale, J. (2006). Uncertainty about the rest of the sentence. *Cognitive science*, 30(4):643–672.

Hale, J. (2016). Information-theoretical complexity metrics. *Language and Linguistics Compass*, 10(9):397–412.

Hale, J. T., Campanelli, L., Li, J., Bhattasali, S., Pallier, C., and Brennan, J. R. (2022). Neurocomputational models of language processing. *Annual Review of Linguistics*, 8(1):427–446.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hoover, B., Strobelt, H., and Gehrmann, S. (2019). exbert: A visual analysis tool to explore learned representations in transformers models. *arXiv preprint arXiv:1910.05276*.

Hu, J., Gauthier, J., Qian, P., Wilcox, E., and Levy, R. P. (2020). A systematic assessment of syntactic generalization in neural language models. *arXiv preprint arXiv:2005.03692*.

Huang, K.-J., Arehalli, S., Kugemoto, M., Muxica, C., Prasad, G., Dillon, B., and Linzen, T. (2024). Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, 137:104510.

Huang, N. and Phillips, C. (2021). When missing nps make double center-embedding sentences acceptable. *Glossa: a journal of general linguistics*, 6(1).

Jäger, L. A., Engelmann, F., and Vasishth, S. (2017). Similarity-based interference in sentence comprehension: Literature review and bayesian meta-analysis. *Journal of Memory and Language*, 94:316–339.

Kallini, J., Papadimitriou, I., Futrell, R., Mahowald, K., and Potts, C. (2024). Mission: Impossible language models. *arXiv preprint arXiv:2401.06416*.

Kodner, J., Payne, S., and Heinz, J. (2023). Why linguistics will thrive in the 21st century: A reply to piantadosi (2023). *arXiv preprint arXiv:2308.03228*.

Kuribayashi, T., Oseki, Y., Brassard, A., and Inui, K. (2022). Context limitations make neural language models more human-like. *arXiv preprint arXiv:2205.11463*.

Lakretz, Y., Hupkes, D., Vergallito, A., Marelli, M., Baroni, M., and Dehaene, S. (2021). Mechanisms for handling nested dependencies in neural-network language models and humans. *Cognition*, page 104699.

Lan, N., Chemla, E., and Katzir, R. (2024). Large language models and the argument from the poverty of the stimulus. *Linguistic Inquiry*, pages 1–28.

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

Levy, R. (2013). Memory and surprisal in human sentence comprehension.

Levy, R., Fedorenko, E., and Gibson, E. (2013). The syntactic complexity of russian relative clauses. *Journal of memory and language*, 69(4):461–495.

Levy, R. and Gibson, E. (2013). Surprisal, the pdc, and the primary locus of processing difficulty in relative clauses. *Frontiers in Psychology*, 4:229.

Lewis, R. (1998). Working memory in sentence processing: Retroactive and proactive interference in parsing. In *11th Annual CUNY Conference on Human Sentence Processing, New Brunswick, NJ*.

Lewis, R. L. (1993). An architecturally-based theory of human sentence comprehension. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF COMPUTER SCIENCE.

Lewis, R. L. (1996). Interference in short-term memory: The magical number two (or three) in sentence processing. *Journal of psycholinguistic research*, 25(1):93–115.

Lewis, R. L. (2000). Specifying architectures for language processing: Process, control, and memory in parsing and interpretation. *Architectures and mechanisms for language processing*, pages 56–89.

Lewis, R. L., Howes, A., and Singh, S. (2014). Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in cognitive science*, 6(2):279–311.

Lewis, R. L., Shvartsman, M., and Singh, S. (2013). The adaptive nature of eye movements in linguistic tasks: How payoff and architecture shape speed-accuracy trade-offs. *Topics in cognitive science*, 5(3):581–610.

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

Lewis, R. L., Vasishth, S., and Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends in cognitive sciences*, 10(10):447–454.

Lissón, P., Paape, D., Pregla, D., Burchert, F., Stadie, N., and Vasishth, S. (2023). Similarity-based interference in sentence comprehension in aphasia: A computational evaluation of two models of cue-based retrieval. *Computational Brain & Behavior*, 6(3):473–502.

Lissón, P., Pregla, D., Nicenboim, B., Paape, D., Van het Nederend, M. L., Burchert, F., Stadie, N., Caplan, D., and Vasishth, S. (2021). A computational evaluation of two models of retrieval processes in sentence processing in aphasia. *Cognitive Science*, 45(4):e12956.

Logačev, P. and Vasishth, S. (2016). A multiple-channel model of task-dependent ambiguity resolution in sentence comprehension. *Cognitive Science*, 40(2):266–298.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in cognitive sciences*.

Mancheva, L., Reichle, E. D., Lemaire, B., Valdois, S., Ecalle, J., and Guérin-Dugué, A. (2015). An analysis of reading skill development using ez reader. *Journal of Cognitive Psychology*, 27(5):657–676.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Marcus, M. P. (1979). An overview of a theory of syntactic recognition for natural language.

McElree, B., Foraker, S., and Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of memory and language*, 48(1):67–91.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., and Brockman, W. (2011). The google books team. *Pickett, JP, Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, MA, & Aiden, EL*, pages 176–182.

Miller, G. A. and Isard, S. (1964). Free recall of self-embedded english sentences. *Information and control*, 7(3):292–303.

Millière, R. (2024a). Language models as models of language. *arXiv preprint arXiv:2408.07144*.

Millière, R. (2024b). Philosophy of cognitive science in the age of deep learning. *Wiley Interdisciplinary Reviews: Cognitive Science*, 15(5):e1684.

Norris, D. (2006). The bayesian reader: explaining word recognition as an optimal bayesian decision process. *Psychological review*, 113(2):327.

Norris, D. (2009). Putting it all together: a unified account of word recognition and reaction-time distributions. *Psychological Review*, 116(1):207.

Oh, B.-D., Clark, C., and Schuler, W. (2022). Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, 5:777963.

Oh, B.-D. and Schuler, W. (2022). Entropy-and distance-based predictors from gpt-2 attention patterns predict reading times over and above gpt-2 surprisal. *arXiv preprint arXiv:2212.11185*.

Oh, B.-D. and Schuler, W. (2023a). Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. *arXiv preprint arXiv:2304.11389*.

Oh, B.-D. and Schuler, W. (2023b). Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Piantadosi, S. T. (2023). Modern language models refute chomsky's approach to language. *From fieldwork to linguistic theory: A tribute to Dan Everett*, pages 353–414.

Pritchett, B. L. (1988). Garden path phenomena and the grammatical basis of language processing. *Language*, pages 539–576.

Qian, P. and Levy, R. P. (2019). Neural language models as psycholinguistic subjects: representations of syntactic state. Association for Computational Linguistics.

Rayner, K. and Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review*, 3(4):504–509.

Reichle, E. D., Rayner, K., and Pollatsek, A. (2003). The ez reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and brain sciences*, 26(4):445–476.

Ryu, S. H. and Lewis, R. L. (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.

Schustack, M. W., Ehrlich, S. F., and Rayner, K. (1987). Local and global sources of contextual facilitation in reading. *Journal of Memory and language*, 26(3):322–340.

Shadlen, M. N. and Shohamy, D. (2016). Decision making and sequential sampling from memory. *Neuron*, 90(5):927–939.

Shain, C. (2021). Cdrnn: Discovering complex dynamics in human language processing. 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 3718–3734.

Shain, C. and Schuler, W. (2018). Deconvolutional time series regression: A technique for modeling temporally diffuse effects. In *Proceedings of the 2018 conference on empirical methods in natural language processing*.

Shain, C., Van Schijndel, M., Futrell, R., Gibson, E., and Schuler, W. (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.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Shiffrin, R. M. (2003). Modeling memory and perception. *Cognitive science*, 27(3):341–378.

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

Staub, A. (2010). Eye movements and processing difficulty in object relative clauses. *Cognition*, 116(1):71–86.

Stolz, W. S. (1967). A study of the ability to decode grammatically novel sentences. *Journal of Verbal Learning and Verbal Behavior*, 6(6):867–873.

Sturt, P., Pickering, M. J., and Crocker, M. W. (1999). Structural change and reanalysis difficulty in language comprehension. *Journal of Memory and Language*, 40(1):136–150.

Taylor, W. L. (1953). "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

Timkey, W. and Linzen, T. (2023). A language model with limited memory capacity captures interference in human sentence processing. *arXiv preprint arXiv:2310.16142*.

Trott, S., Jones, C., Chang, T., Michaelov, J., and Bergen, B. (2023). Do large language models know what humans know? *Cognitive Science*, 47(7):e13309.

Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(2):407.

Van Dyke, J. A. and Lewis, R. L. (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.

Van Schijndel, M. and Linzen, T. (2021). Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive science*, 45(6):e12988.

Vani, P., Wilcox, E. G., and Levy, R. (2021). Using the interpolated maze task to assess incremental processing in english relative clauses. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

Vasishth, S. and Engelmann, F. (2021). *Sentence Comprehension as a Cognitive Process: A computational approach*. Cambridge University Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Vig, J. (2019). A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42.

Vig, J. and Belinkov, Y. (2019). Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76.

Voita, E., Talbot, D., Moiseev, F., Sennrich, R., and Titov, I. (2019). Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808.

Wagers, M. W., Lau, E. F., and Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, 61(2):206–237.

Wald, A. and Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, pages 326–339.

Wilcox, E., Levy, R., Morita, T., and Futrell, R. (2018). What do rnn language models learn about filler-gap dependencies? *arXiv preprint arXiv:1809.00042*.

Wilcox, E. G., Futrell, R., and Levy, R. (2024a). Using computational models to test syntactic learnability. *Linguistic Inquiry*, 55(4):805–848.

Wilcox, E. G., Hu, M., Mueller, A., Linzen, T., Warstadt, A., Choshen, L., Zhuang, C., Cotterell, R., and Williams, A. (2024b). Bigger is not always better: The importance of human-scale language modeling for psycholinguistics.

Yadav, H., Paape, D., Smith, G., Dillon, B. W., and Vasishth, S. (2022). Individual differences in cue weighting in sentence comprehension: An evaluation using approximate bayesian computation. *Open Mind*, 6:1–24.

Yngve, V. H. (1960). A model and an hypothesis for language structure. *Proceedings of the American philosophical society*, 104(5):444–466.

Zini, J. E. and Awad, M. (2022). On the explainability of natural language processing deep models. *ACM Computing Surveys*, 55(5):1–31.