A Transparent Model of Syntactic and Semantic Cue-based Retrieval

Shisen Yue
Johns Hopkins University
syue11@jh.edu

John Hale
Johns Hopkins University
jthale@jhu.edu

Abstract

Human comprehenders have greater difficulty forming pairwise grammatical dependencies in cases where the earlier word competes with a "distractor" to which it is similar. Cue-based retrieval theories (see e.g., Lewis et al., 2006) address this "interference" phenomenon with explicit quantifications of memory retrieval difficulty. We propose a computational model, consistent with Cue-based retrieval, that separately quantifies two different kinds of similarity. A linear combination of the two reproduces the graded interference pattern reported in Van Dyke (2007). This simple account offers a more straightforward mechanistic interpretation than Attention-based predictors from opaque Transformer based models.

1 Introduction

A longstanding question in human language processing is how comprehenders retrieve previously encountered words from memory to establish dependencies with the current word. Cue-based retrieval theory proposes that memory access is content-addressable: morphological, syntactic, and semantic cues from the current word probe memory for matching candidates, while partially matching distractors create similarity-based interference during retrieval (for an overview, see Lewis et al., 2006).

Recent work has drawn an analogy between this retrieval process and the attention mechanism in Transformer language models. Under this view, a word's features act as queries that retrieve matching keys from the preceding context, and attention patterns provide candidate predictors of retrieval success and competition (Ryu and Lewis, 2021; Oh and Schuler, 2022; Ryu and Lewis, 2025). While such predictors often correlate with human processing difficulty, their interpretability remains limited: attention signals are distributed across many layers and heads, and there is no foolproof way to identify which components correspond to specific linguistic operations (Clark et al., 2019; Voita et al., 2019; Jain and Wallace, 2019).

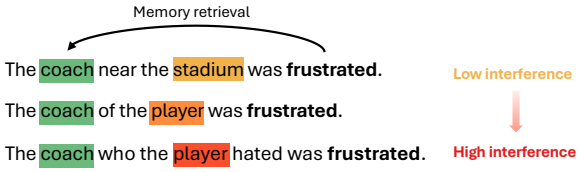

Figure 1: Retrieval of the subject *coach* becomes harder as the distractor increasingly matches the target in relevant linguistic features, producing a gradient from low to high interference. The full model computation for this example is shown in Appendix Figure 9.

One response has been to build smaller models that parameterize retrieval computations more explicitly. For example, Timkey and Linzen (2023) implement query–key–value–style retrieval operations while constraining model capacity to approximate human working-memory limits. However, because retrieval signals are still implemented through distributed neural dynamics, the mapping between model parameters and specific linguistic properties remains difficult to characterize mechanistically.

In this spirit, we ask whether the interaction between syntactic structure and semantic plausibility can be explained using a simple, explicitly parameterized model rather than opaque neural architectures. To investigate this possibility, we develop a lightweight linear model that factors syntactic and semantic contributions to interference while remaining close to the cue-based retrieval theory. Our approach treats each dependency arc as a retrieval trace and models retrieval as an asymmetric linear transformation from a cue representation into the target's representational space. The syntactic component (Grammar-bilinear model) uses supertags from Combinatory Categorial Grammar (CCG), a lexicalized syntactic theory whose

Proceedings of the Society for Computation in Linguistics (SCiL) 2026, pages 411–422

word-level supertags provide fine-grained syntactic descriptions that has been argued to align well with human sentence comprehension (Steedman, 2001; Steedman and Baldridge, 2011; Stanojević et al., 2023). The semantic component (Asymw2v) adapts dependency-based embeddings (Levy and Goldberg, 2014) to learn distinct input (cue) and output (context-bound target) spaces. By separating these two sources of competition, the model provides a transparent parameterization of interference that can be analyzed directly in terms of syntactic structure and lexical compatibility. We evaluate the model on the interference materials of Van Dyke (2007), which manipulate syntactic similarity and semantic plausibility in a 2×2 design. Our results show that: (1) the syntactic and semantic submodels produce independent interference signals corresponding to structural similarity and thematic plausibility; (2) a simple linear combination of these signals yields graded retrieval predictions across the four experimental conditions; and (3) the resulting mixed model outperforms attention heads within GPT-2 that are putatively specialized for grammatical subjecthood, both in retrieval accuracy and in alignment with human reading difficulty.

2 Background

Cue-based memory retrieval theory Cue-based retrieval theory proposes that sentence comprehension involves retrieving information that was introduced by earlier words using features of the current words as memory cues. Morphological, syntactic, and semantic properties of the cue probe a content-addressable memory store to retrieve the most compatible candidate (Lewis et al., 2006; Parker et al., 2017; Vasishth and Engelmann, 2021). Interference arises when multiple items partially match these cues, leading to slower processing and increased comprehension difficulty. Computational implementations such as Lewis and Vasishth (2005) characterize retrieval latency and accuracy as a function of cue–target similarity and competition. In syntactic dependency processing, each dependency arc can be viewed as a retrieval event: when encountering a verb such as *frustrated*, the processor must retrieve its subject (e.g., *coach*) from preceding context. Interference occurs when distractors (e.g., *player*) share syntactic or semantic features with the intended target, producing measurable effects on reading times and comprehension accuracy (Van Dyke, 2007; Oberauer and Lewandowsky, 2008).

Distributional representations Cue-based retrieval theory distinguishes between lexical information that serves as a memory cue and working-memory elements that are retrieved. In this work, we operationalize both cues and candidate memory elements as word embeddings, making distributional representations a natural way to model lexical retrieval. Distributional embedding models learn vector representations in which words that occur in similar contexts occupy nearby regions of a shared space. Saussure (1916) distinguishes between syntagmatic relations, in which words are associated with the words that surround them, and paradigmatic relations, in which words are related because they can substitute for one another. Building on this distinction, Kelly et al. (2020) show that higher-order distributional associations can induce syntactic lexical categories, such as part-of-speech and CCG syntactic types. Our approach instead focuses on the syntagmatic side of lexical knowledge: retrieval is modeled as an asymmetric relation between a cue and a role-bound target. Dependency-based word embeddings (Levy and Goldberg, 2014) provide a natural implementation of this idea by defining contexts not as linear windows but as syntactic dependency relations extracted from parsed text. Each training instance pairs a word with a relation-labeled context (e.g., *coach/nsubj* represents *coach* when appearing as a candidate word to be retrieved in a `nsubj` relation), allowing the model to capture functional relations (e.g., dog $\leftrightarrow$ bark) between words rather than topical co-occurrence (dog $\leftrightarrow$ paw). As in the Skip-Gram framework (Mikolov et al., 2013), the model learns two parameter matrices: an input matrix representing words when they serve as cues (queries) and an output matrix representing relation-bound contexts.

LLMs as psycholinguistic predictors Recent work has used autoregressive language models to derive predictors of human sentence processing difficulty, typically operationalized as next-word surprisal. Transformer-based models such as GPT-2 reliably capture well-known effects in reading behavior and have been shown to predict reading times and neural responses during language comprehension (Shain, 2024; Caucheteux et al., 2022). However, several studies report an inverse scaling effect: smaller models often correlate better with

human reading-time data than larger ones, possibly because larger models form overly confident expectations for rare or unpredictable words (Arehalli et al., 2022; Oh and Schuler, 2023b,a; Steuer et al., 2023). This observation has motivated approaches that derive psycholinguistic predictors from internal model representations rather than output probabilities alone. For example, attention-based entropy and distance metrics (Oh and Schuler, 2022) and analyses of syntactically specialized attention heads (Ryu and Lewis, 2021, 2025) attempt to link model components more directly to retrieval-like processes in human sentence comprehension. More broadly, recent work emphasizes cognitively constrained or lightweight architectures as potentially better aligned with human incremental processing mechanisms (Clark et al., 2025; Timkey and Linzen, 2023).

3 Model

We model interference during cue-based retrieval in sentence processing. Retrieval is treated as the operation that links the current word (the cue) to a previously encountered word (the target) in memory, corresponding to dependency relations in the sentence (Van Dyke and McElree, 2006; Hofmeister and Vasishth, 2014). We formalize retrieval as an asymmetric transformation that projects a cue representation into the representational space of candidate targets and produces a probability distribution over preceding words, reflecting their relative retrieval scores.

The model separates syntactic and semantic sources of interference using two components. Section 3.1 introduces the grammar-bilinear model, which approximates syntactic compatibility using CCG-based features. Section 3.2 presents the Asymw2v lexical model, which models semantic selectional preferences using dependency-based embeddings. Section 3.3 then combines these signals into a mixed model. While the present study focuses on the `nsubj` dependency as a case study, the framework naturally extends to other dependency relations.

3.1 Grammar-bilinear model

The grammar-bilinear model captures syntactic interference by scoring the compatibility between a cue word and candidate targets in its left context. Retrieval is formulated as a bilinear similarity between cue and candidate representations, producing a probability distribution over preceding words.

the	coach	of	the	player	was	frustrated
NP/N	N	$(NP\backslash NP)/NP$	NP/N	N	$(S[dcl]\backslash NP)/(S[adj]\backslash NP)$	$S[adj]\backslash NP$

Figure 2: Example of CCG supertags assigned to a sentence. Each word is annotated with a lexical category encoding its combinatory behavior. For example, the verb *was* receives the category $(S[dcl]\backslash NP)/(S[adj]\backslash NP)$, indicating that it combines with an adjectival predicate on the right to form a verb phrase that in turn combines with a subject noun phrase on the left.

Representation The model's syntactic information is limited to CCG supertags. These tags encode combinatory potential including subcategorization. Because these categories are highly informative about syntactic composition, assigning supertags is often described as "almost parsing" (Bangalore and Joshi, 1999). An example of supertag assignments is shown in Figure 2, where each word in the sentence is annotated with a lexical category describing how it combines with neighboring constituents. Supertag therefore provide a rich feature space for modeling fine-grained combinatory constraints in machine-learning models (Kasai et al., 2019).

Motivated by these properties, we approximate the syntactic environment of a word k as the concatenation of the supertags of the current word and its two preceding words:

$$h_k = [G_{k-2}; G_{k-1}; G_k]$$

where G_i is a one-hot vector over the supertag vocabulary. Training data are drawn from the English Web Treebank, using CCG supertags from *ccgtools*[1] and dependency relations from the Universal Dependencies annotations (Silveira et al., 2014).

Retrieval scoring Given cue representation h_k, a trainable matrix W projects the cue into the candidate space. If the supertag vocabulary has size V, then $h \in \mathbb{R}^{3V}$ and $W \in \mathbb{R}^{3V \times 3V}$. For each candidate c in the preceding context, the retrieval score is

$$s_c = h_c^\top W h_k$$

Scores are normalized over all preceding words using a softmax:

$$\hat{p}(w_c|w_k) = \frac{e^{h_c^\top W h_k}}{\sum_{i<k} e^{h_i^\top W h_k}}$$

[1] https://github.com/stanojevic/ccgtools

The model is trained to maximize the likelihood of retrieving the correct subject specified by the dependency parse. Because supertags encode detailed combinatory constraints, the model learns to assign higher scores to candidates occupying syntactically compatible positions, while structurally incompatible candidates receive lower retrieval probabilities.

3.2 Asymw2v

The Asymw2v model captures semantic interference through dependency-based distributional representations. It models the compatibility between a cue word and candidate targets using embeddings trained on dependency-labeled contexts, allowing the model to encode selectional preferences associated with particular grammatical relations.

Representation and parameters Following the dependency-based embedding framework introduced in Section 2, each training pair includes a word and a relation-labeled context derived from the dependency arcs (e.g., *frustrated*, *coach/nsubj*). The model learns two sets of embeddings: an input embedding $v_w \in \mathbb{R}^d$ for each word w, and an output embedding $u_c \in \mathbb{R}^d$ for each relation-bound context c. If the word vocabulary has size $|V|$ and the set of dependency contexts has size $|C|$, the model contains $d(|V| + |C|)$ parameters.

Embeddings are trained using the Skip-gram objective on a dependency-parsed Wikipedia corpus[2]. After training, the input embedding of the current word serves as the cue vector, while the output embeddings of relation-labeled candidates (e.g., *coach/nsubj*) serve as role-specific target representations.

Retrieval scoring Given cue embedding v_k, the compatibility between the cue and a candidate context c is computed using a dot product

$$s_c = v_k^\top u_c$$

and normalized across candidate targets using a softmax over all preceding words in the sentence. Because contexts are tied to specific dependency relations, the model learns graded preferences for plausible fillers of each relation. For example, verbs that frequently take animate subjects assign higher scores to animate `nsubj` candidates than to inanimate distractors, providing a distributional signal for semantic interference during retrieval.

[2] https://huggingface.co/datasets/wikimedia/wikipedia

3.3 Mixed model

The mixed model combines syntactic and lexical retrieval signals by linearly combining their scores before normalization. For each candidate i,

$$s_i^{mix} = \alpha s_i^{syn} + \beta s_i^{lex}$$

where s_i^{syn} and s_i^{lex} are the scores produced by the grammar-bilinear and Asymw2v models respectively. The resulting scores are converted into retrieval probabilities using a softmax over all preceding candidates:

$$p_i^{mix} = \text{softmax}(s^{mix})$$

This late-fusion formulation reflects cue-based retrieval accounts in which syntactic constraints provide the primary retrieval cues while semantic compatibility contributes graded competition among candidates (Van Dyke and Lewis, 2003; Van Dyke and McElree, 2011). By combining the two signals at the scoring stage, the model preserves strong syntactic filtering while allowing lexical-semantic similarity to modulate retrieval interference.

4 Experiments

We evaluate whether the proposed mixed model (as well as its two subcomponents) derive the inference patterns that we would expect under cue-based retrieval theory. Specifically, we ask whether the grammar-bilinear component captures syntactic interference, whether Asymw2v captures semantic interference, and whether their combination recovers the full human-like interaction between the two.

4.1 Experimental Materials

We evaluate the model using the sentence materials from Tan et al. (2017), adapted from Van Dyke (2007). These stimuli implement a 2×2 factorial design crossing syntactic interference (Low vs. High) and semantic interference (Low vs. High). Each item set contains four sentences corresponding to the conditions LoSynLoSem, LoSynHiSem, HiSynLoSem, and HiSynHiSem.

In these materials, the retrieval cue occurs at the main verb (e.g., *will forget*), which retrieves its grammatical subject (*the hostess*). Syntactic interference is manipulated by varying whether the intervening noun phrase occupies subject or object position in the embedded clause. Semantic interference is manipulated by varying the plausibility of the distractor noun as the subject of the main verb

(e.g., *toddler* vs. *room*). High interference arises when the distractor overlaps with the retrieval cues in syntactic or semantic features.

Table 1 illustrates the structure of the stimuli.

Sentence Region	Example Stimulus
Introduction	The hostess
Intervening region	
LoSyn/LoSem	who had yelled about the dirty **room** loudly
LoSyn/HiSem	who had yelled about the dirty **toddler** loudly
HiSyn/LoSem	who yelled that the **room** was dirty loudly
HiSyn/HiSem	who yelled that the **toddler** was dirty loudly
Critical region	will forget
Spillover region	about the
Last word	mess

Table 1: Experimental materials from Tan et al. (2017). Interference words are bold and differ in their syntactic/semantic compatibility to the verb.

4.2 Experimental Setup

The mixed model introduced in Section 3.3 combines the syntactic and lexical components with fixed weights $\alpha = 0.4$ and $\beta = 0.6$, which are selected without an extensive hyperparameter search for a reason to provide a simple, fixed combination of the two components.

To quantify retrieval competition, we compute the relative retrieval score $Relret$ of the correct subject. Following Timkey and Linzen (2023), this measure represents the proportion of retrieval probability assigned to the correct target relative to the distractor:

$$RelRet(v, s) = \frac{Ret(v, s)}{Ret(v, s) + Ret(v, d)}$$

where $Ret(v, s)$ denotes the model's retrieval score from cue v to the correct subject s, and d denotes the distractor candidate. Given the balanced 2×2 design, a two-way repeated-measures ANOVA is also performed to assess main effects of syntactic and semantic interference and their interaction.

4.3 Experiment 1: Component Validation

We first test whether the two model components independently capture their intended interference dimensions. The grammar-bilinear model should exhibit sensitivity to syntactic interference, while the Asymw2v model should respond primarily to semantic interference.

Figure 3 shows the relative retrieval scores predicted by the two models across the experimental conditions. As expected, the grammar-bilinear model shows a strong effect of syntactic interference, with significantly lower retrieval scores in HiSyn conditions than LoSyn conditions, remaining insensitive to semantic interference. In contrast, the Asymw2v model shows that retrieval scores decrease under high semantic interference but remain unaffected by syntactic interference. The paired t-test statistics for these contrasts demonstrate that all predicted effects are statistically significant ($p < 0.05$), confirming that the two components capture distinct sources of retrieval interference.

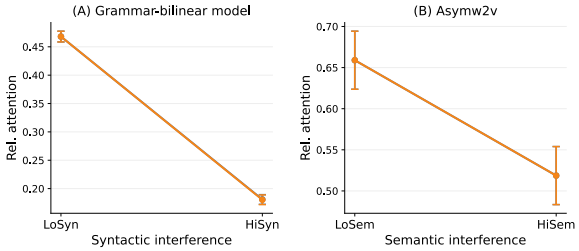

Figure 3: Separate validation of the two components. Mean relative retrieval scores predicted by the grammar-bilinear model and Asymw2v. The grammar-bilinear model (A) is sensitive to syntactic interference, whereas the Asymw2v model (B) is sensitive to semantic interference.

4.4 Experiment 2: Mixed Model Evaluation

Table 2 compares retrieval accuracy for the mixed model, two GPT-2-small attention heads, and human data. A prediction is correct when the highest retrieval or attention score is assigned to the intended subject. The GPT-2 heads are identified by Ryu and Lewis (2021) as `nsubj`-sensitive using the head-specialization diagnostic of Voita et al. (2019).

Our model captures the human-like accuracy hierarchy, with highest accuracy in LoSynLoSem and lowest accuracy in HiSynHiSem. The Transformer heads fail in complementary ways: $head4\_3$ robustly attends to the correct subject under low syntactic interference but shows little semantic sensitivity, whereas $head3\_6$ produces more graded condition-wise relative-attention scores but often assigns attention to the wrong item, yielding near-zero retrieval accuracy.

Figure 4 shows the same contrast in relative retrieval or attention scores, calculated following Timkey and Linzen (2023). The mixed model shows significant syntactic and semantic interference effects, with all pairwise comparisons signif-

Condition	Mixed model	Attention heads (Ryu and Lewis, 2021)		Human (Van Dyke, 2007)
		head4_3	head3_6	
LoSynLoSem	0.5625	0.4250	0.0250	0.90
LoSynHiSem	0.4500	0.4625	0.0375	0.82
HiSynLoSem	0.4875	0.2375	0.0000	0.86
HiSynHiSem	0.2750	0.2250	0.0000	0.73

Table 2: Accuracy across conditions for the mixed model GPT-2-small attention heads, and human data from Van Dyke (2007).

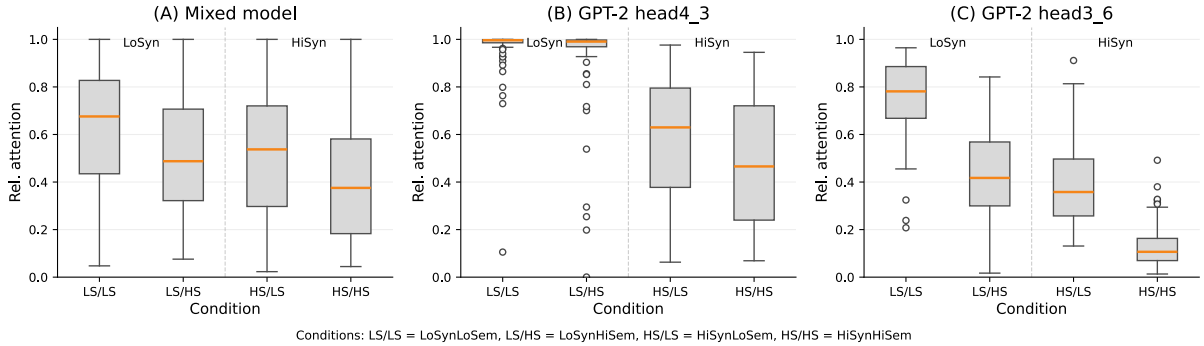

Figure 4: Comparison of relative retrieval score patterns across models and human data. Boxplots display model-predicted relative attention to the correct subject across the four interference conditions from Tan et al. (2017).

icant ($p < 0.05$), consistent with the qualitative human pattern (Van Dyke, 2007).

In contrast, $head4\_3$ exhibits near-ceiling relative attention to the correct subject under low syntactic interference and show minimal semantic interference effects, which is consistent with results from Timkey and Linzen (2023). $head3\_6$ shows more balanced and separable retrieval-score distributions across all four conditions; however, its low retrieval accuracy indicates that these condition-level effects do not correspond to reliable retrieval of the intended subject.

4.5 Error Analysis

To diagnose this mismatch between condition-level separation and target-level accuracy, we compute each model's average output distribution over six linguistically interpretable regions preceding the verb. The regions include: (1) the determiner of the subject noun phrase, (2) the real subject that should agree with the main verb, (3) introductory elements preceding the subject, (4) the distractor noun within the embedded clause that creates interference, (5) the remainder of the embedded clause following the distractor, and (6) the token immediately preceding the verb (typically the auxiliary "was" in our stimuli). By aggregating attention weights within these syntactic regions, we can better assess whether attention heads are tracking linguistically relevant dependencies or responding to other structural patterns.

Figure 5 shows that $head3\_6$ allocates its attention mostly to the auxiliary that immediately precedes it, rather than on the words that are possible as the retrieval cue's intended target. This pattern contrasts sharply with $head4\_3$ and our mixed model, which shows stronger attention to the real subjects and distractors. Importantly, when the mixed model makes retrieval errors, these errors tend to favor the interference distractor manipulated in Van Dyke's human experiment rather than linguistically irrelevant tokens, suggesting that the model's failures are themselves aligned with the intended cue-based competition.

This observation raises a broader interpretive question: if a psycholinguistic effect appears in a model component not designed for that function, should we treat it as meaningful evidence of the underlying cognitive mechanism, or does it instead point to a different kind of specialization within the network? This is discussed in greater detail in Section 5.

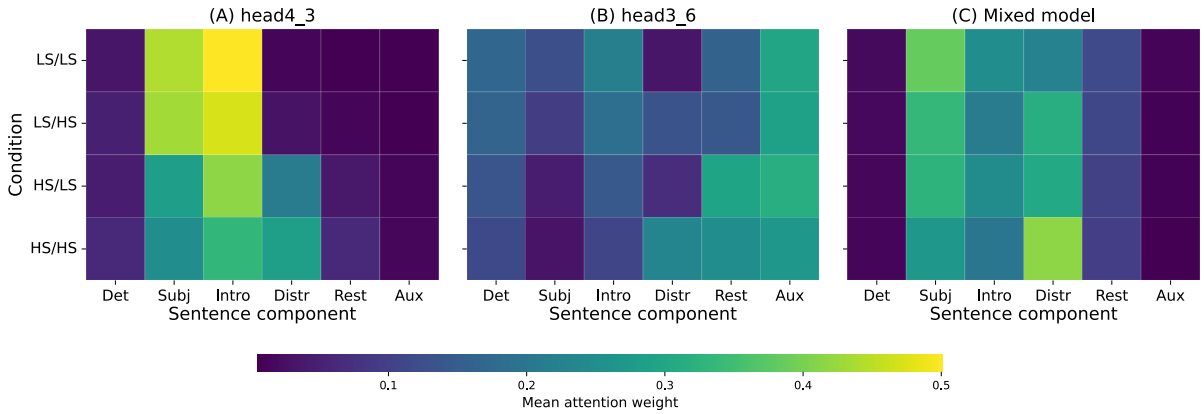

Figure 5: Error analysis from this paper (C) as well as two previously-proposed models of retrieval (A,B). Mean attention weights from the main verb to preceding tokens for mixed model, $head4\_3$, and $head3\_6$. Rows indicate the syntactic region before the verb cue, columns represent syntactic-semantic interference conditions. Color intensity corresponds to attention weight magnitude (see individual colorbars).

5 Discussion

5.1 Relation to attention-based accounts of retrieval

In our modeling with Tan et al. (2017)'s materials, the attention head proposed by Ryu and Lewis (2021) as syntactically specialized for `nsubj` retrieval ($head4\_3$) shows a strong effect of syntactic interference but little sensitivity to semantic interference. This pattern is consistent with the observations of Timkey and Linzen (2023). Rather than contradicting prior work, it suggests that such heads may track structural compatibility cues relevant to subject dependencies while failing to capture the broader combination of syntactic and semantic competition implicated in human interference effects.

At the same time, another head ($head3\_6$) exhibits interference-like patterns in its relative attention scores across both syntactic and semantic dimensions. However, its retrieval accuracy is close to zero (Table 2), and the attention-distribution analysis in Figure 5 shows that it allocates most attention to the auxiliary preceding the verb rather than to plausible subject candidates. As noted by Ryu and Lewis (2025), this head is primarily associated with relations such as `nmod:poss` and `nummod`, rather than `nsubj`. The interference pattern therefore arises even though the head's attention is not primarily directed toward the candidates that a subject-retrieval account would identify as relevant. Taken together, these observations highlight a broader challenge for attention-based explanations: neural architectures distribute linguistic computations across many interacting components, and there is currently no foolproof method for identifying which internal units correspond to particular linguistic operations.

5.2 Implications for computational psycholinguistic modeling

These results suggest that relatively simple computational mechanisms may be sufficient to capture key interference phenomena in sentence processing. Syntax and semantics may not be as hopelessly entangled as they appear from the perspective of opaque pretrained neural networks.

More broadly, the findings highlight a trade-off between fit to human data and mechanistic transparency in computational psycholinguistic modeling. In the search for scientific insight into human language comprehension, neural language models can provide strong predictors of human behavior, but the relationship between internal model components and cognitive mechanisms is often difficult to establish. By contrast, explicitly parameterized models such as the one proposed here can transparently identify the source of interference.

From this perspective, the goal of computational psycholinguistics is not only to identify predictors that correlate with human behavior, but also to develop models that clarify the mechanisms generating those effects (Hale, 2017). The present model contributes to this effort by providing a simple formalization of cue-based retrieval in which syntactic and semantic interference arise from independently interpretable computations. Such models can serve as a useful bridge between symbolic accounts of

memory retrieval and distributional representations learned from large corpora.

6 Conclusion

This study presents a computational model of cue-based retrieval interference. By combining these two components within a simple linear retrieval framework, the model captures the graded interference patterns observed in experimental materials manipulating structural similarity and semantic plausibility.

Our comparison with attention-based predictors from GPT-2-small highlights the difficulty of interpreting neural model components as direct implementations of linguistic retrieval mechanisms. While some attention heads exhibit partial interference patterns, their behavior does not consistently align with the retrieval targets predicted by cue-based theory.

More broadly, the results show that key interference effects in sentence processing can be modeled using a transparent and explicitly parameterized architecture. This approach provides a computational bridge between symbolic psycholinguistic theories of memory retrieval and distributional representations learned from large corpora, offering a clearer mechanistic account of how syntactic and semantic similarity interact during comprehension.

Acknowledgments

We thank the anonymous reviewers for their helpful comments and suggestions. We are also grateful to Miloš Stanojević for valuable discussions and feedback on earlier versions of this work.

References

Suhas Arehalli, Brian Dillon, and Tal Linzen. 2022. Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 301–313, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Srinivas Bangalore and Aravind K. Joshi. 1999. Supertagging: An approach to almost parsing. *Computational Linguistics*, 25(2):237–265.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. 2022. Deep language algorithms predict semantic comprehension from brain activity. *Scientific Reports*, 12(1):16327.

Christian Clark, Byung-Doh Oh, and William Schuler. 2025. Linear recency bias during training improves transformers' fit to reading times. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7735–7747, Abu Dhabi, UAE. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What Does BERT Look at? An Analysis of BERT's Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

John Hale. 2017. Models of human sentence comprehension in computational psycholinguistics. In Mark Aronoff, editor, *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.

Philip Hofmeister and Shravan Vasishth. 2014. Distinctiveness and encoding effects in online sentence comprehension. *Frontiers in Psychology*, 5.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Jungo Kasai, Dan Friedman, Robert Frank, Dragomir Radev, and Owen Rambow. 2019. Syntax-aware neural semantic role labeling with supertags. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 701–709, Minneapolis, Minnesota. Association for Computational Linguistics.

Mary Alexandria Kelly, Moojan Ghafurian, Robert L West, and David Reitter. 2020. Indirect associations in learning semantic and syntactic lexical relationships. *Journal of Memory and Language*, 115:104153.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, Maryland. Association for Computational Linguistics.

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

Richard L. Lewis, Shravan Vasishth, and Julie A. Van Dyke. 2006. Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, 10(10):447–454.

Bofang Li, Tao Liu, Zhe Zhao, Buzhou Tang, Aleksandr Drozd, Anna Rogers, and Xiaoyong Du. 2017. Investigating different syntactic context types and context representations for learning word embeddings. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages

2421–2431, Copenhagen, Denmark. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Klaus Oberauer and Stephan Lewandowsky. 2008. Forgetting in immediate serial recall: Decay, temporal distinctiveness, or interference? *Psychological review*, 115(3):544.

Byung-Doh Oh and William Schuler. 2022. Entropy- and distance-based predictors from GPT-2 attention patterns predict reading times over and above GPT-2 surprisal. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9324–9334, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Byung-Doh Oh and William Schuler. 2023a. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1915–1921, Singapore. Association for Computational Linguistics.

Byung-Doh Oh and William Schuler. 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.

Dan Parker, Michael Shvartsman, and Julie A. Van Dyke. 2017. The cue-based retrieval theory of sentence comprehension: New findings and new challenges. In Linda Escobar, Vicenç Torrens, and Teresa Parodi, editors, *Language Processing and Disorders*, pages 121–144. Cambridge Scholars Publishing, Newcastle.

Soo Hyun Ryu and Richard Lewis. 2021. Accounting for agreement phenomena in sentence comprehension with transformer language models: Effects of similarity-based interference on surprisal and attention. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 61–71, Online. Association for Computational Linguistics.

Soo Hyun Ryu and Richard L Lewis. 2025. Memory for prediction: A transformer-based theory of sentence processing. *Journal of Memory and Language*, 145:104670.

Ferdinand de Saussure. 1916. Rapports syntagmatiques et rapports associatifs. *Cours de linguistique générale. Payot, Paris, France*, pages 170–175.

Cory Shain. 2024. Word frequency and predictability dissociate in naturalistic reading. *Open Mind*, 8:177–201.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Miloš Stanojević, Jonathan R Brennan, Donald Dunagan, Mark Steedman, and John T Hale. 2023. Modeling structure-building in the brain with ccg parsing and large language models. *Cognitive science*, 47(7):e13312.

Mark Steedman. 2001. *The syntactic process*. MIT press.

Mark Steedman and Jason Baldridge. 2011. Combinatory categorial grammar. *Non-transformational syntax: Formal and explicit models of grammar*, pages 181–224.

Julius Steuer, Marius Mosbach, and Dietrich Klakow. 2023. Large GPT-like models are bad babies: A closer look at the relationship between linguistic competence and psycholinguistic measures. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 142–157, Singapore. Association for Computational Linguistics.

Yingying Tan, Randi C Martin, and Julie A Van Dyke. 2017. Semantic and syntactic interference in sentence comprehension: A comparison of working memory models. *Frontiers in psychology*, 8:198.

William Timkey and Tal Linzen. 2023. A language model with limited memory capacity captures interference in human sentence processing. *arXiv preprint arXiv:2310.16142*.

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

Julie A Van Dyke and Richard L Lewis. 2003. Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language*, 49(3):285–316.

Julie A. Van Dyke and Brian McElree. 2006. Retrieval interference in sentence comprehension. *Journal of Memory and Language*, 55(2):157–166.

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

Shravan Vasishth and Felix Engelmann. 2021. *Sentence comprehension as a cognitive process: A computational approach*. Cambridge University Press.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 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, Florence, Italy. Association for Computational Linguistics.

A Appendix

A.1 Training details

The grammar-bilinear model is trained for 20 epochs using a learning rate of 0.01 without L2 regularization.

For the asymmetric word2vec (Asymw2v) component, we follow the general training configuration of Li et al. (2017). Negative sampling uses 5 negative samples and a distribution smoothing parameter of 0.75. The model is trained for 2 epochs with a learning rate of 0.005 and a batch size of 2048. No dynamic context sampling or subsampling is applied.

The word and context vocabularies include tokens and dependency-labeled contexts with a minimum frequency threshold of 100.

A.2 First word bias

We observe the first word bias in attention heads of GPT-2 Small layer 4, head 3 indicated in (Ryu and Lewis, 2025), as illustrated in Figure 6. In this phenomenon, the attention mechanism disproportionately assigns the highest weight to the initial token of the sequence, regardless of its syntactic or semantic relevance to the current token.

Figure 6: First word bias shown in GPT-2-small *head*4_3

A.3 Ablation of the Grammar-Bilinear Model with CCG Context Windows

We further investigate the impact of the CCG supertag convolution window size k on the model's capacity to identify the correct subject among words preceding a verb. As illustrated in Figure 7, expanding the context window k results in a monotonic increase in prediction accuracy and a concomitant decrease in entropy. The performance gains begin to saturate beyond $k = 4$. This trend confirms that enriching representations with local syntactic history (via concatenated previous supertags) significantly reduces uncertainty and aids in resolving subject-verb dependencies relative to the context-agnostic baseline ($k = 0$). Furthermore, the observed stabilization suggests a theoretical upper bound on the model's performance when using purely local syntactic context.

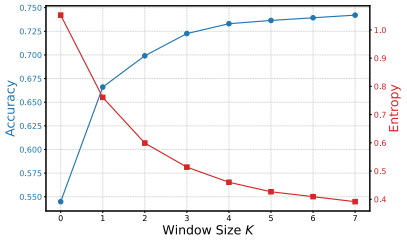

Figure 7: Accuracy and entropy of grammar-bilinear models on the English Web Treebank test set across varying convolution window sizes k. Here, k represents the number of previous tokens' supertags concatenated to the current token's supertag; $k = 0$ denotes a baseline where each token is represented solely by its own supertag without contextual information.

A.4 Probing for Semantic Animacy

To demonstrate that the learned Asymw2v vectors reflect animacy through vector distance in the high-dimensional space, we conduct a linear probe study. We use WordNet to construct a robust animacy lexicon, resulting in a dataset of 9,336 items ($N_{animate} = 4,336$, $N_{inanimate} = 5,000$) after removing OOV. Using these labels, we train a logistic regression classifier as a linear probe on the 300-dimensional frozen embeddings to determine if animacy is linearly separable in the vector space.

As shown in Table 3, the results indicate a strong alignment between the learned vector geometry and semantic animacy. The linear probe achieves an overall accuracy of 91.33% and a ROC-AUC score of 0.9674. The high precision and recall across both classes suggest that the embedding space effectively clusters entities based on agency without explicit semantic supervision.

Figure 8 visualize the separation by projecting test-set embeddings onto the normal vector of the learned decision boundary, yielding a one-dimensional representation along the discrimina-

Metric/Class	Precision	Recall	F1-Score	Support
Inanimate	0.94	0.90	0.92	1000
Animate	0.89	0.93	0.91	868
Overall Accuracy		0.9133		
ROC-AUC		0.9674		

Table 3: Linear probe results for animacy classification using logistic regression on 300-dimensional embeddings. The model demonstrates high separability between animate and inanimate classes.

tive axis. The resulting histogram shows two distinct, largely non-overlapping distributions. The inanimate terms (blue) cluster around a negative mean projection value, while animate terms (orange) cluster around a positive mean. The minimal overlap between the two densities corroborates the high ROC-AUC score reported in Table 3, demonstrating that the semantic distinction between animate and inanimate tokens is geometrically realized linearly in the embedding space.

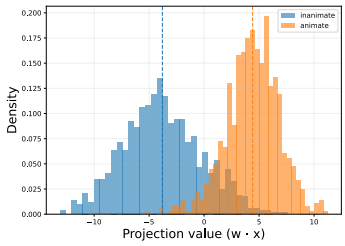

Figure 8: Histogram of projection values ($w \cdot x$) for the test set vocabulary on the learned animacy axis. Dashed lines indicate the mean projection value for each class.

A.5 Model demonstration

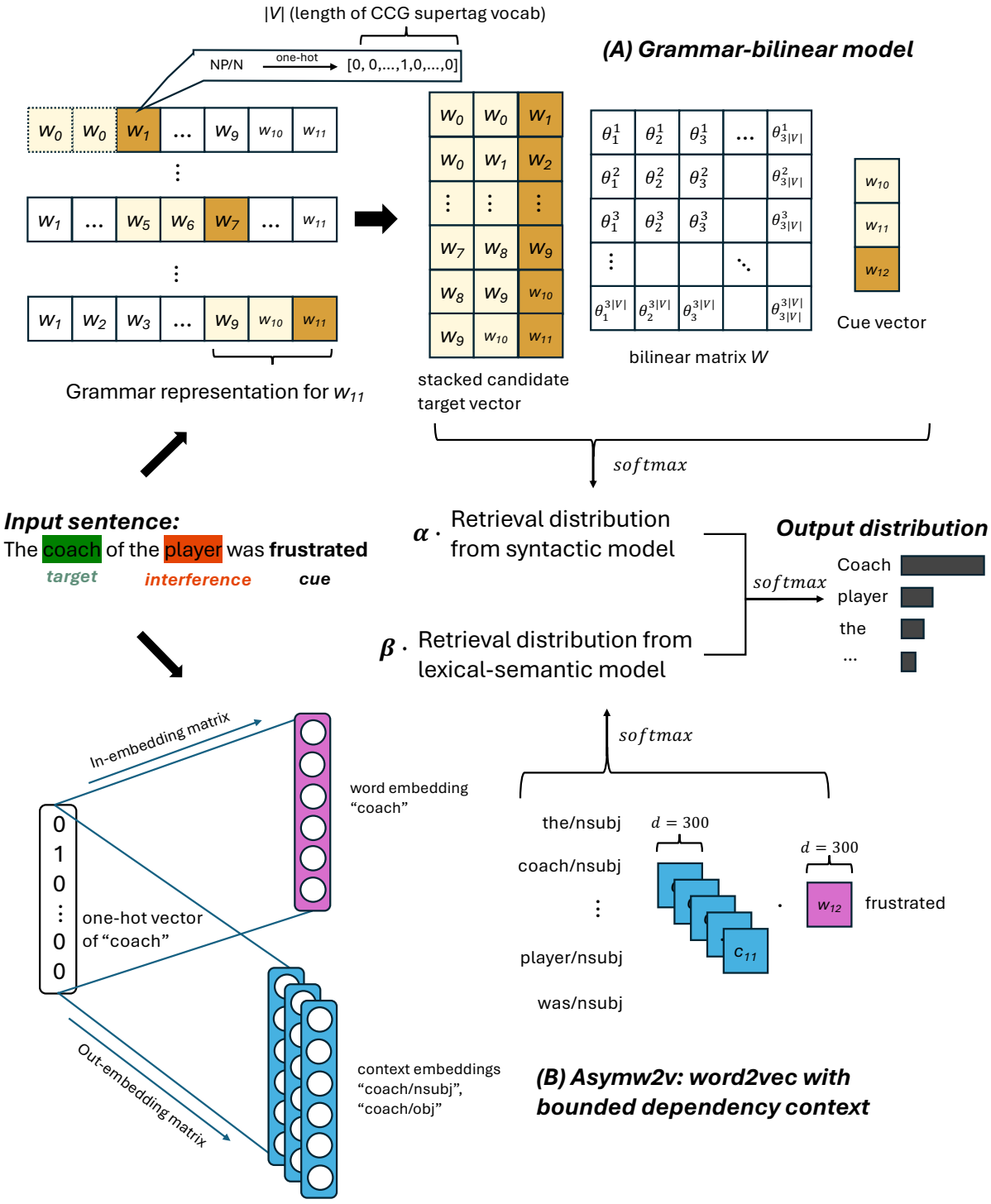

Figure 9: Architecture of the proposed retrieval model. The grammar-bilinear component (A) represents syntactic context using CCG supertags and computes cue–target compatibility through a bilinear transformation to produce a syntactic retrieval distribution. The lexical-semantic component (B) uses dependency-based word embeddings to model semantic compatibility between the cue and candidate targets. The two distributions are combined linearly and normalized to produce the final retrieval probability over candidate subjects.