# Accounting for Agreement Phenomena in Sentence Comprehension with Transformer Language Models: Effects of Similarity-based Interference on Surprisal and Attention

**Soo Hyun Ryu**
Department of Psychology
University of Michigan
soohyunr@umich.edu

**Richard L. Lewis**
Department of Psychology
University of Michigan
rickl@umich.edu

## Abstract

We advance a novel explanation of similarity-based interference effects in subject-verb and reflexive pronoun agreement processing, grounded in surprisal values computed from a pretrained large-scale Transformer model, GPT-2. Specifically, we show that surprisal of the verb or reflexive pronoun predicts *facilitatory interference effects* in ungrammatical sentences, where a distractor noun that matches in number with the verb or pronoun leads to faster reading times, despite the distractor not participating in the agreement relation. We review the human empirical evidence for such effects, including recent meta-analyses and large-scale studies. We also show that *attention patterns* (indexed by entropy and other measures) in the Transformer show patterns of diffuse attention in the presence of similar distractors, consistent with cue-based retrieval models of parsing. But in contrast to these models, the attentional cues and memory representations are learned entirely from the simple self-supervised task of predicting the next word.

## 1 Introduction

Deep Neural Network (DNN) language models (LeCun et al., 2015; Sundermeyer et al., 2012; Vaswani et al., 2017) have recently attracted the attention of researchers interested in assessing their linguistic competence (Chaves, 2020; Da Costa and Chaves, 2020; Ettinger, 2020; Wilcox et al., 2018, 2019) and potential to provide accounts of psycholinguistic phenomena in sentence processing (Futrell et al., 2018; Linzen and Baroni, 2021; Van Schijndel and Linzen, 2018; Wilcox et al., 2020). In this paper we show how attention-based transformer models (we use a pre-trained version of GPT-2) provide the basis for a new theoretical account of facilitatory interference effects in subject-verb and reflexive agreement processing. These effects, which we review in detail below, have played an important role in psycholinguistic theory because they show that properties of noun phrases that are not the grammatical targets of agreement relations may nonetheless exert an influence on processing time at points where those agreement relations are computed.

The explanation we propose here is a novel one grounded in surprisal (Hale, 2001; Levy, 2008), but with origins in graded attention and similarity-based interference (Van Dyke and Lewis, 2003; Lewis et al., 2006; Jäger et al., 2017). We use surprisal as the key predictor of reading time (Levy, 2013), and through targeted analyses of patterns of attention in the transformer, show that the model behaves in ways consistent with cue-based retrieval theories of sentence processing. The account thus provides a new integration of surprisal and similarity-based interference theories of sentence processing, adding to a growing literature of work integrating noisy memory and surprisal (Futrell et al., 2020). In this case, the noisy representations arise from training the transformer, and interference must exert its influence on reading times through a *surprisal bottleneck* (Levy, 2008).

The remainder of this paper is organized as follows. We first provide an overview of some of key empirical work in human sentence processing concerning subject-verb and reflexive pronoun agreement. We then provide a brief overview of the GPT-2 architecture, its interesting psycholinguistic properties, and the method and metrics that we will use to examine the agreement effects. We then apply GPT-2 to the materials used in several different human reading time studies. We conclude with some theoretical reflections, identification of weaknesses, and suggestions for future work.

## 2 Agreement Interference Effects in Human Sentence Processing

One long-standing focus of work in sentence comprehension is understanding how the structure of human short-term memory might support and con-

*Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 61–71
Online Event, June 10, 2021. ©2021 Association for Computational Linguistics

strain the incremental formation of linguistic dependencies among phrases and words (Gibson, 1998; Lewis, 1996; Lewis et al., 2006; Miller and Chomsky, 1963; Nicenboim et al., 2015). A key property of human memory thought to shape sentence processing is *similarity-based interference* (Miller and Chomsky, 1963; Lewis, 1993, 1996). Figure 1 shows a simple example of how such interference arises in cue-based retrieval models of sentence processing, as a function of the compatibility of *retrieval targets* and *distractors* with retrieval *cues* (Lewis and Vasishth, 2005; Lewis et al., 2006; Van Dyke and Lewis, 2003) (Corresponding sentences are from Wagers et al. (2009)'s Exp 4–6 shown in Table 1). *Inhibitory interference effects* occur when features of the target perfectly match the retrieval cue and features of a distractor partially matches, while *facilitatory interference effects* occur when the features of both target and distractor partially match the features of retrieval cue.

In this study, we focus on interference effects in subject-verb number agreement and reflexive pronoun-antecedent agreement, specifically in languages where the agreement features include *syntactic number* which is morphologically marked on the verb or pronoun. In such cases, number is plausibly a useful retrieval cue, and it is easy to manipulate the number of distractor noun phrases to allow for carefully controlled empirical contrasts.

**Interference in subject-verb agreement.** Previous studies (Pearlmutter et al., 1999; Wagers et al., 2009; Dillon et al., 2013; Lago et al., 2015; Jäger et al., 2020) attest to both *inhibitory* interference (slower processing in the presence of an interfering distractor) and *facilitatory* interference (faster processing in the presence of an interfering distractor), but the existing empirical support for inhibitory interference is weak, and many studies fail to find any evidence for it (Dillon et al., 2013; Lago et al., 2015; Wagers et al., 2009). There is stronger evidence for facilitatory effects, which arise in ungrammatical structures where the verb or pronoun fails to agree in number with the structurally correct target noun phrase, but where either an intervening or preceding distractor noun phrase does match in number. Example A. below illustrates, taken from Wagers et al. (2009), where the subject and verb are boldfaced and the distractor noun is underlined:

A. The **slogan** on the <u>posters</u> **were** designed to get attention.

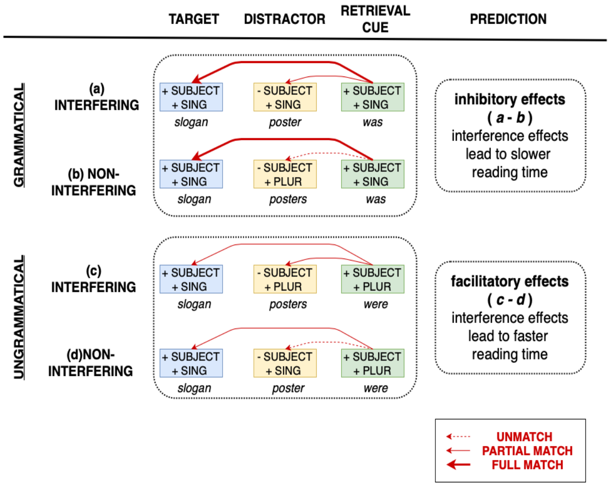
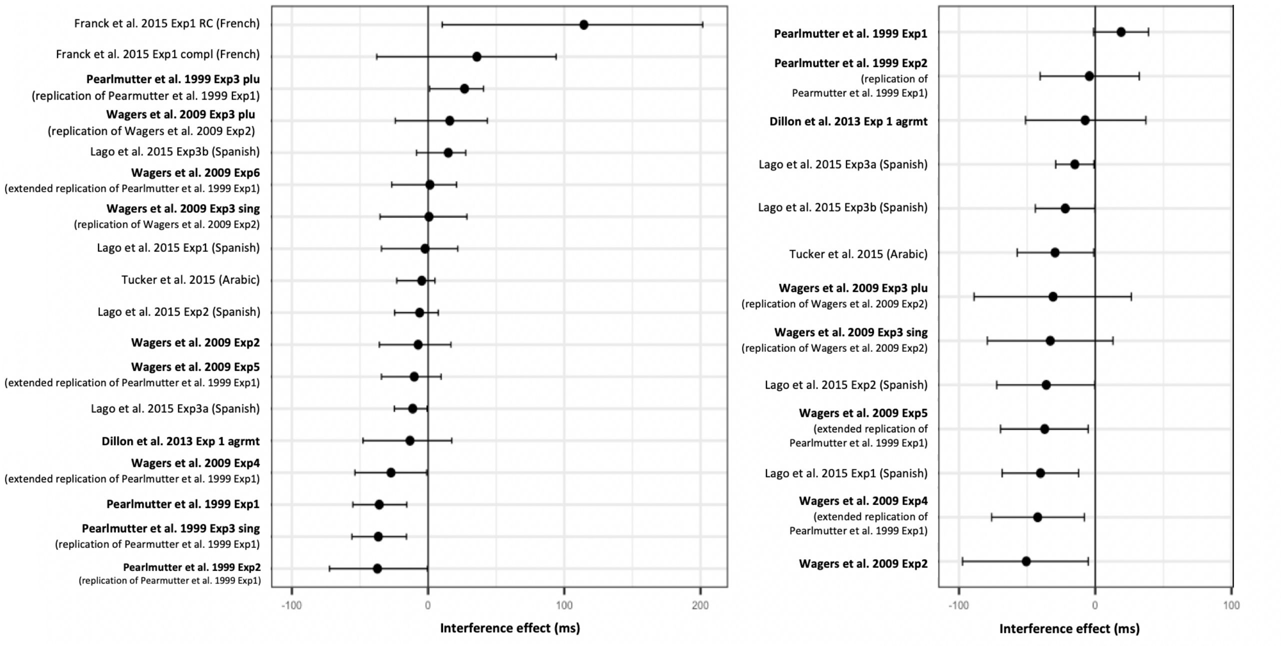


Figure 1: How facilitatory and inhibitory interference effects arise in subject-verb dependency creation in cue-based retrieval parsing. The critical manipulation concerns the overlap of number feature between the distractor, target, and retrieval cue.

A Bayesian meta-analysis of agreement phenomena was recently conducted with an extensive set of studies (Jäger et al., 2017; Vasishth and Engelmann, 2021). Their analysis of first-pass reading times from eye-tracking experiments on subject-verb number agreement is shown in Figure 1. The evidence from the meta-analysis is consistent with a very small or nonexistent inhibitory interference effect in in the grammatical conditions, with a small but robust facilitatory interference effects in the ungrammatical conditions. Concerned that the existing experiments did not have sufficient power to detect the inhibitory effects, Nicenboim et al. (2018) ran a large scale eye-tracking study (185 participants) with materials designed to increase the inhibition effect, and did detect a 9ms effect (95% credible posterior interval 0–18ms). This represents the strongest evidence to date for inhibitory effects in grammatical agreement structures, but even this evidence indicates the effect may be near zero.

**Interference in reflexive pronoun agreement.** Example B. below shows a pair of sentences from Dillon et al. (2013) used to probe facilitatory effects in reflexive pronoun agreement (again, the target antecedent and pronoun are boldfaced and the distractor is underlined):

B. (1) *interfering* The basketball **coach** who trained the star <u>players</u> usually blamed **themselves** for the ...



(a) Interference effects in grammatical sentences

(b) Interference effects in ungrammatical sentences

Figure 2: Results of the meta-analysis on subject-verb number agreement from Vasishth and Engelmann (2021). The materials from boldfaced studies are those that we used in our GPT-2 experiments.

(2) *non-interfering* The basketball **coach** who trained the star player usually blamed **themselves** for the ...

The empirical record concerning facilitatory effects in reflexive agreement is mixed. Some have claimed that such effects do not arise (Sturt, 2003; Xiang et al., 2009; Dillon et al., 2013), and that this is expected under a model in which the structural constraints from binding theory (Chomsky et al., 1982) serve to effectively filter candidates for retrieval—in short, the parser does not consider or make contact with the ungrammatical distractor noun phrases (Sturt, 2003; Dillon et al., 2013).

However, a recent Bayesian meta-analysis of key experiments by Dillon et al. (2013) indicates substantially overlapping posterior estimates of facilitatory effects for subject-verb agreement and reflexive agreement (Vasishth and Engelmann, 2021). Concerned again about under-powered studies, Jäger et al. (2020) undertook a large scale (181 participants) eye-tracking replication and did find evidence for nearly equivalent facilitatory speedups for reflexive and subject-verb agreement (Figure 3). This result is not inconsistent with the meta-analysis, but provides stronger evidence that the facilitation effects in reflexives are real.

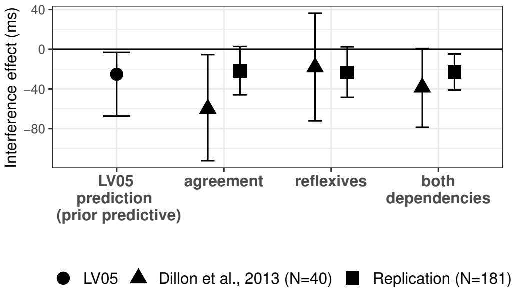


Figure 3: From Jäger et al. (2020). Posterior estimates of facilitatory interference effects in subject-verb and reflexive agreement processing in a large scale replication of Dillon et al. (2013), the original effects, and predictions from the Lewis and Vasishth (2005) model.

We take advantage of the very broad coverage of GPT-2 by having GPT-2 process the same set of sentence materials as human subjects in four different agreement experiments. To anticipate our key results, we find GPT-2 yields lower surprisal, i.e. facilitatory effects, in both subject-verb and reflexive pronoun conditions. Furthermore, we show that attention at the verb or pronoun is distributed to both target and distractor in just those conditions where the distractor matches the hypothesized number retrieval cue (Lin et al., 2019). Finally, we show that the surprisal contrasts between matching and non-matching distractors in the grammatical (inhibitory) interference conditions are essentially zero.

## 3 GPT-2 for Psycholinguistic Analysis

**The psycholinguistic relevance of GPT-2 and its training method.** GPT-2 (Generative Pre-trained Transformer-2), introduced by OpenAI in Radford et al. (2019), is a language model with a decoder-only Transformer architecture (Vaswani et al., 2017), and has achieved state-of-the-art performance in diverse downstream tasks. GPT-2 and other large-scaled language models based on transformer architectures were trained on billions of words of text, and engineered with performance in mind, not with concern for psycholinguistic plausibility. Why then should we then take them seriously as the basis of psycholinguistic models?

We believe that the new transformer-based models have three important properties that make them of psycholinguistic interest. (a) The models are among the first to serve as the basis of systems that achieve human-level performance on a range of linguistic tasks, and they directly generate a key quantity, *surprisal of the next word*, that we know is an important predictor of reading times in humans (Hale, 2001; Levy, 2008). (b) Although the data requirements are currently much greater than that for human language acquisition, the models are trained on a simple task—predict the next word—that may plausibly serve as the basis of a self-supervised learning signal in human language acquisition. The representations that arise from such learning are thus psycholinguistically interesting. (c) The learned soft-attention and parallel content-based retrieval of representations of prior input are architectural properties of the GPT models that align very closely with retrieval-based models of sentence comprehension (Lewis et al., 2006). And the structure of these psycholinguistic models was proposed as a response to the challenges of computing long-distance dependencies—the same challenge that motivated the transformer as a departure from standard recurrent architectures (Vaswani et al., 2017; Galassi et al., 2020).

**Identifying specialized heads in GPT-2.** Here we use the medium-sized GPT-2 which is constructed with 12 layers, each of which includes 12 attention heads. Previous studies have revealed that individual attention heads in Transformer models serve are at least partially specialized in function (Clark et al., 2019; Vig, 2019; Vig and Belinkov, 2019; Voita et al., 2019). Specifically, Voita et al. (2019) found that certain attention heads are specialized for different dependency relations.

Following Voita et al. (2019)'s method, we identified heads that are specialized for subject-verb relations and reflexive anaphora resolution. Voita et al. (2019)'s method works as follows. First, sentences are parsed using CoreNLP dependency parser (Manning et al., 2014). Then, relative string positions (e.g., one token back, two tokens back) of all instances in each syntactic dependency were counted. Considering the proportion of the most frequent relative position as the baseline, attention heads are selected as specialized for a particular dependency relation if attention is paid for the corresponding dependent at least 10% more often than the baseline. In other words, there must be some evidence that the attention head is sensitive to the dependency and not merely string position.

To find attention heads responsible for the relation between subjects and verbs, we used the CoreNLP parser on 148,376 sentences from the Brown corpus and Gutenberg corpus provided via Natural Language Toolkit (NLTK) (Bird et al., 2009), extracting 49,145 *nsubj* relations, which associate nominal subjects and their governors which are mostly verbs. The most frequent relative position for *nsubj* dependency relation is -1, which means that the nominal subjects usually come right before their governor, taking up 42% of the cases.

After analyzing the attention distribution pattern using GPT-2, we obtained four syntactic heads that were found to be partly specialized for *nsubj* dependency relations: *head4_3* (59%); *head3_6* (51%); *head6_0* (49%); *head2_9* (49%)[1]. Although we expect that the four syntactic heads responsible for *nsubj* dependency relation may play distinct roles, in our analyses here we simply use the best performing head (*head4_3*).

The same method was implemented to find attention heads responsible for reflexive anaphora resolution. The only difference was that we used NeuralCoref (Wolf et al., 2018) to count relative position of antecedents to reflexive anaphora since the dependency parser does not associate antecedents and anaphora. Out of 2,660 sentences that includes reflexive anaphora, we extracted 510 sentences where NeuralCoref identified a single unique antecedent for the reflexive pronoun. The most fre-

[1] head*n_m* refers to the *m*-th attention head in the *n*-th layer. Numbers in parentheses indicate accuracies of heads in paying the highest attention to the subject/antecedent by the verb/pronoun.

Table 1: A set of data included for the experiment on subject-verb agreement. (Wagers et al. (2009)'s Exp3 also included sets with plural subjects in the ungrammatical conditions.)

| | Interference | Grammaticality | Example sentences |
|---|---|---|---|
| **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 ... |

quent relative position for reflexive anaphora and their antecedents was -2, meaning that antecedents appear before reflexive anaphora having one word in between. The proportion of the highest relative position was 22%, requiring 24.2 % of accuracy for attention heads to be considered responsible for reflexive anaphora resolution. We found four heads whose accuracies are higher than the threshold: *head1_5* (44%); *head3_5* (39%); *head4_3* (27%); *head6_0* (25%), and we again take the best performing head (*head1_5*) for further analysis.

**Metrics.** We define here three metrics for our analyses: *surprisal*, *attention entropy from syntactic heads*, and *attention to target*. We use surprisal for making reading time predictions, but use the attention metrics to provide insight into the processing at the critical region and therefore the representations computed in the prefix before the critical region. Surprisal is thus based on the final prediction of the entire model, but the attention metrics are associated with the attention heads most specialized for our dependencies of interest.

**Surprisal** (Hale, 2001; Levy, 2008) is defined as the negative log probability of the word given left context.

$$\mathrm{Surprisal}(w) = -\log_2 P(w|context) \quad (1)$$

Any use of surprisal requires adoption of some kind of language model; e.g. some past work has used probabilistic CFGs (Levy, 2008). Here we use GPT-2, which computes after each word a probability distribution over its large lexicon that is conditioned on its internal representation of the left context.

**Attention to target** is simply the value of the soft attention vector element that corresponds to the target word position, which we denote $\mathrm{Attn}(w_{cue}, w_{target})$, and indicates how much attention is allocated to the target by one of the specialized attention heads (*head4_3* for subject-verb and *head1_5* for reflexives.)

**Attention entropy** is a variant of Shannon (1948)'s information entropy that we use as a measure of how sharply focused (low entropy) or diffuse (high entropy) the attention pattern is. (It may be thought of as a measure of the uncertainty about the attentional target, but because the attention values are not probabilities from which targets are sampled, this interpretation is not strictly warranted).

$$\mathrm{Entropy}(w_i) = \sum_{j=1}^{i-1} \mathrm{Attn}(w_i, w_j) \times \log_2 \mathrm{Attn}(w_i, w_j) \quad (2)$$

where *i* refers to the location of the critical word, *j* are locations of prior words, and $\mathrm{Attn}(w_i, w_j)$ is attention allocated to $w_j$ from $w_i$.

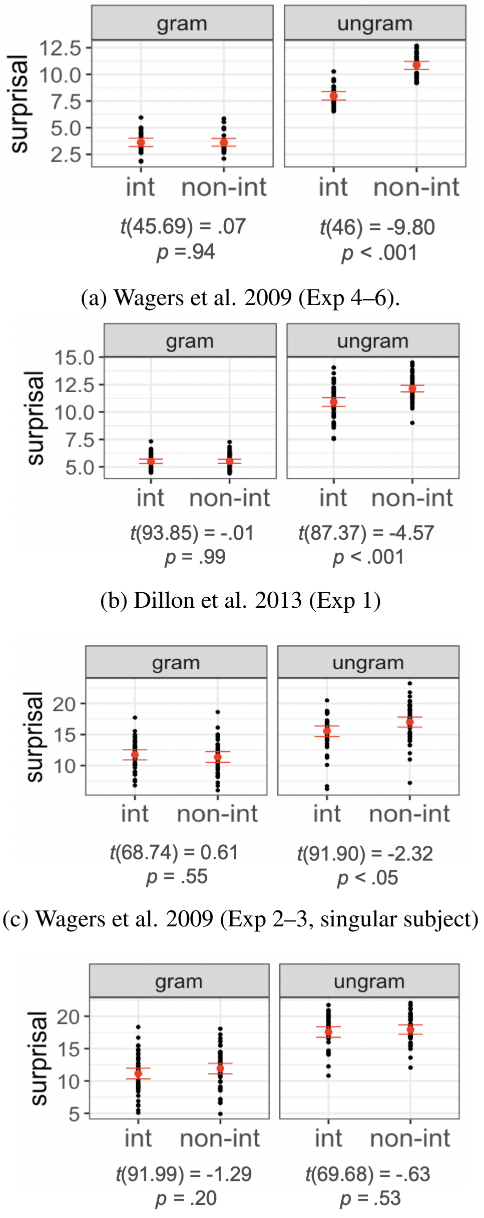
## 4 Subject-verb Agreement Experiments

To investigate whether GPT-2 may predict facilitatory interference effects in subject-verb agreement, we ran GPT-2 on materials from three studies (Dillon et al., 2013; Wagers et al., 2009): 48 sets of sentences from Experiments 2-3 in Wagers et al. (2009)[2]; 24 sets of sentences from Experiments 4-7 in Wagers et al. (2009); 48 sets of sentences from Dillon et al. (2013) (See Table 1).

These three sets of sentences have in common a 2 × 2 structure with the factors *grammaticality* (grammatical/ungrammatical) and *interference* (interfering/non-interfering), as described above. Additionally, Wagers et al. (2009)'s Exp 3 also includes an additional condition, *subject* (singular/plural) for investigating a possible singular-plural asymmetry, i.e., asking whether interference effects are equivalent for plural (for plural verbs) and singular (for singular verbs) distractors.

Note that sentences from Experiments 2–3 in Wagers et al. (2009) involve structures in which the distractor appears *before* the target, and so test effects of *proactive interference*. Thus the distractors are also more distant from verbs than in the other experimental materials.

**Results of surprisal analyses.** Figure 4 shows the surprisal computed at the critical verbs in each of the experiments and in each of the four conditions separately (red dots and intervals represent means and conventional 95% confidence intervals). Surprisal matches the important qualitative pattern found in the meta-analysis of first-pass reading times: lower surprisal—facilitatory effects—are found in the ungrammatical conditions when the distractor matches the verb's number, and no inhibitory effects are found in the grammatical conditions. Furthermore, the effects are largest for the case of *retroactive* interference, where the distractor follows the target and immediately precedes the verb (Figure 4a), compared to *proactive* inteference, where the distractor precedes the target (Figure 4c). The exception is that no facilitatory effects were found when the verb is singular and the target subject is plural (see Figure 4d). But the facilitatory effect in this condition was not reliably different from zero in the meta-analysis, and it mirrors a plural-singular asymmetry (or *markedness* effect) found in agreement attraction in production.



(a) Wagers et al. 2009 (Exp 4–6).

(b) Dillon et al. 2013 (Exp 1)

(c) Wagers et al. 2009 (Exp 2–3, singular subject)

(d) Wagers et al. 2009 (Exp 3, plural subject)

Figure 4: The surprisal of critical verbs computed by GPT-2 on the materials in four subject-verb number agreement experiments. Each small dot is a data point from one sentence; the red dots and intervals represent means and 95% confidence intervals.

**Results of attention analyses.** Our conjecture is that in the *interfering* conditions where the distractor matches the verb in number that the attention of the *nsubj*-specialized attention head *head4_3* will be distributed to both the target *and* the distractor. It is possible to visualize exactly this pattern using a tool developed by Vig (2019). Figure 5 shows an example visualization.

Analyses of the *attention entropy* and *attention to target* metrics provide quantitative evidence for

[2] Wagers et al. (2009)'s materials are an extended and slightly modified version of Pearlmutter et al. (1999)

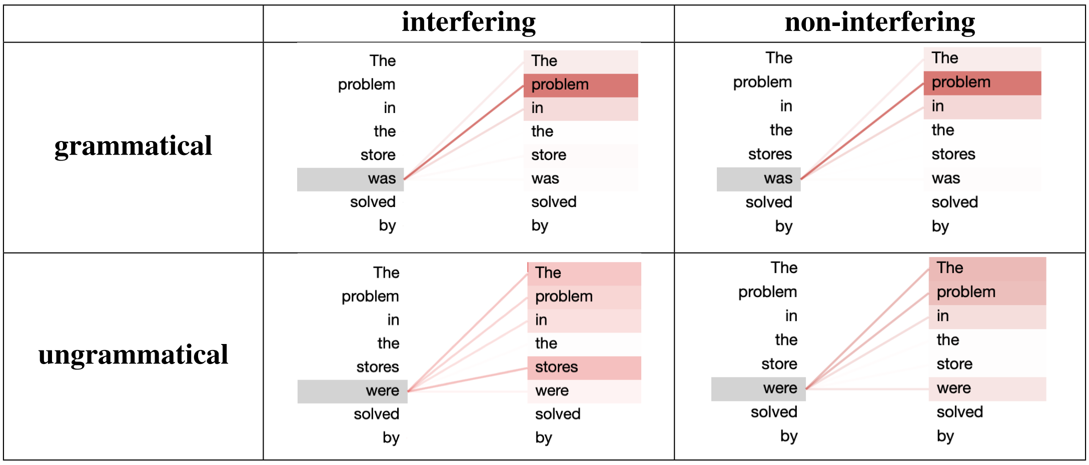

interfering
non-interfering
grammatical
ungrammatical
The problem in the store was solved by
The problem in the stores was solved by
The problem in the stores were solved by
The problem in the store were solved by

Figure 5: An example of the attention distribution of an attention head specialized for subject-verb dependencies in the four conditions of the subject-verb agreement experiments.

this conjecture: Figure 6 shows two metrics across the four datasets. The interfering conditions always show the highest value of *attention entropy* and the lowest value of *attention to target*, which means that the head most specialized for subject-verb relations distributes attention more diffusely and away from the target subject. There is evidence for the expected attention effects even in the grammatical conditions, but in these conditions there is no effect of surprisal. Thus, under a theory in which similarity-based interference exerts its effects on reading time through a *surprisal bottleneck* (Levy, 2008), no reading time differences are expected here—even though the underlying representations and attention patterns may reflect the interference.

**Preliminary corpus analysis of ungrammatical subject-verb agreement sentences.** One possible explanation for the observed facilitatory interference effects is that GPT-2 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, we 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 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 un-edited 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 Experiment 4–6, we 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.

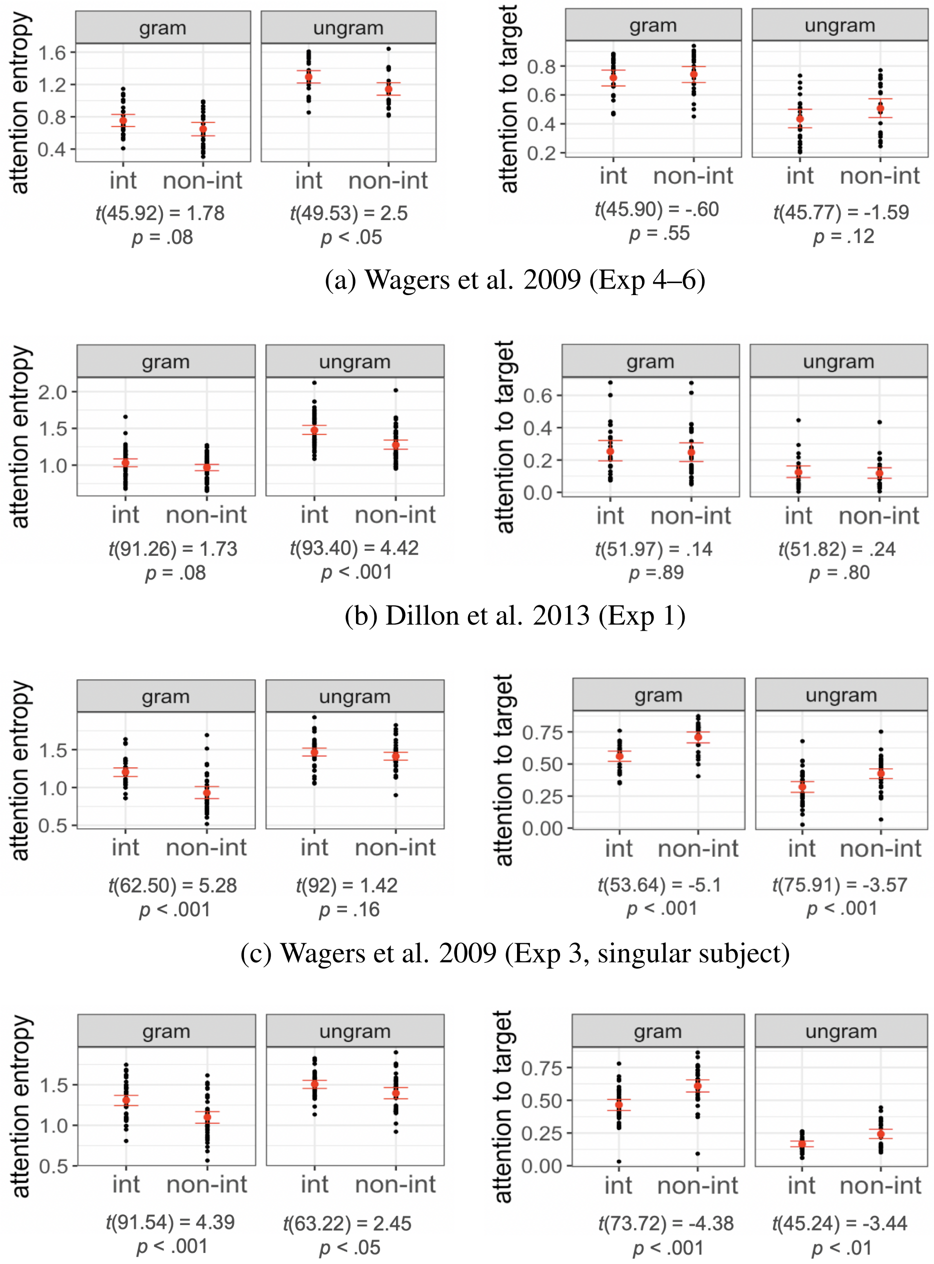
| | singular subj | plural subj |
|---|---|---|
| interfering | 80 | 71 |
| non-interfering | 39 | 51 |

Table 2: Results from a preliminary corpus analysis of patterns of ungrammatical subject-verb agreement. In the key case of a singular subject and a plural verb, the number of an intervening distractor is about twice as likely to be plural (interfering) rather than singular (non-interfering). See text for a discussion.

## 5 Reflexive Agreement Experiments

To examine whether the prediction of GPT-2 are consistent with the null interference effects argued for by Dillon et al. (2013), or show facilitatory interference effects as in the large scale Jäger et al. (2020) replication, we conducted an experiment using the same methodology as described above for the subject-verb experiments, but using the reflexive materials in Dillon et al. (2013), and focusing the attention analyses on the head most specialized for reflexive anaphor resolution. Examples of the materials are shown in Table 3.

**Results of the surprisal analyses.** Summaries of the surprisal (and attention metrics) measured at



(a) Wagers et al. 2009 (Exp 4–6)

(b) Dillon et al. 2013 (Exp 1)

(c) Wagers et al. 2009 (Exp 3, singular subject)

(d) Wagers et al. 2009 (Exp 3, plural subject)

Figure 6: Metrics quantifying attention patterns of the attention head most specialized for subject-verb relations, computed at the verb in the subject-verb agreement experiments.

reflexive anaphora are provided in Figure 7. Consistent with the large scale replication of Dillon et al. (2013) conducted by Jäger et al. (2020) (but inconsistent with the null results reported by Dillon et al), we found lower *surprisal* values in the ungrammatical interfering conditions, consistent with a facilitatory interference effect.

**Results of the attention analyses.** We found little or no differences between interfering and non-interfering cases in the two attention metrics *attention entropy* and *attention to target*. It is possible that this is because the attention head *head1_5* that we found to be partly specialized for reflexive anaphora resolution is actually not as specialized in reflexive anaphora resolution as *head4_3* specialized in *nsubj* dependency resolution. We cannot conclude yet whether there exist heads that serve this function better (that are not detected by the method of Voita et al. (2019)), whether GPT-2 is not reliably resolving the reflexive anaphora, or whether GPT-2 is doing so in a way that is dis-

|  | Interference | Grammaticality | Example sentences |
|---|---|---|---|
| **Dillon 2013 Exp 1 reflexive** | int | gram | The basketball **coach** who trained the star player usually blamed **himself** for the ... |
|  | non-int | gram | The basketball **coach** who trained the star players usually blamed **himself** for the ... |
|  | int | ungram | *The basketball **coach** who trained the star players usually blamed **themselves** for the ... |
|  | non-int | ungram | *The basketball **coach** who trained the star player usually blamed **themselves** for the ... |

Table 3: Examples from Dillon et al. (2013), used in the GPT-2 experiment on reflexive pronoun agreement.

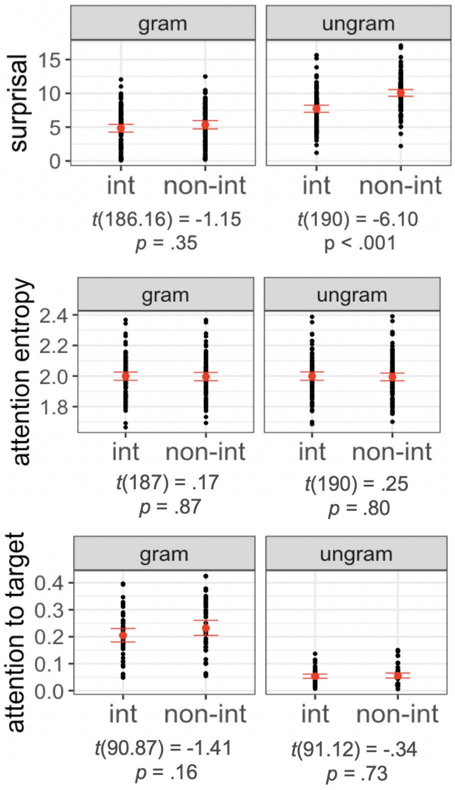


Figure 7: Results of the GPT-2 reflexive agreement experiment using materials from Dillon et al. (2013).

tributed across many attention heads.

## 6 Discussion and Future Directions

Effects of similarity-based interference have been the province of models of noisy memory rather than models of probabilistic expectations, because in standard probabilistic grammars the expectation for the agreement features of a licensor such as a verb or pronoun should not be conditioned upon the agreement features of constituents other than the target licensee. But we show here that a large-scale Transformer language model, GPT-2, trained only to predict the next word, nevertheless yields surprisal values that are consistent with facilitatory interference effects due to distractor noun phrases that do not participate in the agreement relations. We also confirmed that two metrics that are easily computed from the Transformers' attention mechanism, *attention entropy* and *attention to target*, show patterns in the subject-verb experiments that are consistent with cue-based retrieval models.

Our results are suggestive of a possible interesting link between surprisal and noisy memory representations. The attention patterns that we have discovered must reflect similarity between the representations of the target and distractor noun phrases. This representational similarity is the source of great generalization power, but this generalization can lead to linguistic expectations that are not derived by conventional grammatical analyses.

One limitation of our analyses of attention is that they depend on methods for identifying specialized heads for specific dependency types. It is not clear that we understand enough about Transformer models to do this reliably. But our results suggest that for at least some dependencies, these simple attention metrics and head selection methods can yield interesting insights.

The approach outlined may provide an important way to combine surprisal and noisy memory accounts, maintaining a surprisal bottleneck. Using trained Transformers has the significant theoretical advantage that the memory representations, the attention/retrieval cues, and thus the predicted similarity effects are *learned* via a self-supervised prediction task. And so such models naturally yield experience-driven sources of noisy representations that are independent of the process noise assumed in existing memory-based models. Combining the process- and experience-based noise in a single model is an important goal for psycholinguistic theory.

## References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Jonathan P Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. Convokit: A toolkit for the analysis of conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 57–60.

Rui P Chaves. 2020. What don't rnn language models learn about filler-gap dependencies? *Proceedings of the Society for Computation in Linguistics*, 3(1):20–30.

Noam Chomsky et al. 1982. *Some concepts and consequences of the theory of government and binding*. MIT press.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

Jillian K Da Costa and Rui P Chaves. 2020. Assessing the ability of transformer-based neural models to represent structurally unbounded dependencies. *Proceedings of the Society for Computation in Linguistics*, 3(1):189–198.

Brian Dillon, Alan Mishler, Shayne Sloggett, and Colin Phillips. 2013. Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, 69(2):85–103.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

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

Richard Futrell, Ethan Wilcox, Takashi Morita, and Roger Levy. 2018. Rnns as psycholinguistic subjects: Syntactic state and grammatical dependency. *arXiv preprint arXiv:1809.01329*.

Andrea Galassi, Marco Lippi, and Paolo Torroni. 2020. Attention in natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*.

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

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

Lena A Jäger, Felix Engelmann, and Shravan Vasishth. 2017. Similarity-based interference in sentence comprehension: Literature review and bayesian meta-analysis. *Journal of Memory and Language*, 94:316–339.

Lena A Jäger, Daniela Mertzen, Julie A Van Dyke, and Shravan Vasishth. 2020. Interference patterns in subject-verb agreement and reflexives revisited: A large-sample study. *Journal of Memory and Language*, 111:104063.

Sol Lago, Diego E Shalom, Mariano Sigman, Ellen F Lau, and Colin Phillips. 2015. Agreement attraction in spanish comprehension. *Journal of Memory and Language*, 82:133–149.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature*, 521(7553):436–444.

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

Roger Levy. 2013. Memory and surprisal in human sentence comprehension.

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

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

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.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside bert's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 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.

George A Miller and Noam Chomsky. 1963. Finitary models of language users.

Bruno Nicenboim, Shravan Vasishth, Felix Engelmann, and Katja Suckow. 2018. Exploratory and confirmatory analyses in sentence processing: A case study of number interference in german. *Cognitive science*, 42:1075–1100.

Bruno Nicenboim, Shravan Vasishth, Carolina Gattei, Mariano Sigman, and Reinhold Kliegl. 2015. Working memory differences in long-distance dependency resolution. *Frontiers in Psychology*, 6:312.

Neal J Pearlmutter, Susan M Garnsey, and Kathryn Bock. 1999. Agreement processes in sentence comprehension. *Journal of Memory and language*, 41(3):427–456.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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

Patrick Sturt. 2003. The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language*, 48(3):542–562.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2012. Lstm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*.

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.

Marten Van Schijndel and Tal Linzen. 2018. Modeling garden path effects without explicit hierarchical syntax. In *CogSci*.

Shravan Vasishth and Felix Engelmann. 2021. *Sentence Comprehension as a Cognitive Process: A computational approach*. Cambridge University Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Jesse Vig. 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.

Jesse Vig and Yonatan Belinkov. 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.

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.

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

Ethan Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. *arXiv preprint arXiv:2006.01912*.

Ethan Wilcox, Roger Levy, and Richard Futrell. 2019. What syntactic structures block dependencies in rnn language models? *arXiv preprint arXiv:1905.10431*.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do rnn language models learn about filler-gap dependencies? *arXiv preprint arXiv:1809.00042*.

Thomas Wolf, James Ravenscroft, Julien Chaumond, and Maxwell Rebo. 2018. Neuralcoref: Coreference resolution in spacy with neural networks.

Ming Xiang, Brian Dillon, and Colin Phillips. 2009. Illusory licensing effects across dependency types: Erp evidence. *Brain and Language*, 108(1):40–55.