Highlights

Surprisal does not explain syntactic disambiguation difficulty: evidence from a large-scale benchmark

Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, Tal Linzen

- We collect a large reading time dataset for English syntactically complex sentences.
- We evaluate to what extent surprisal can explain processing difficulty.
- Surprisal was estimated using neural network language models.
- Surprisal underestimated processing difficulty for most constructions.
- It failed to predict relative difficulty among different garden path constructions.
- It also failed to predict across-item variation.

Surprisal does not explain syntactic disambiguation difficulty: evidence from a large-scale benchmark

Kuan-Jung Huang[a,1], Suhas Arehalli[b,2], Mari Kugemoto[c,2], Christian Muxica[d,2], Grusha Prasad[e,2], Brian Dillon[c,2], Tal Linzen[f,2]

[a]*Department of Psychological and Brain Sciences, University of Massachusetts, 135 Hicks Way, Amherst, MA, 01003, USA*
[b]*Department of Cognitive Science, Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD, 21218, USA*
[c]*Department of Linguistics, University of Massachusetts, 650 N Pleasant St, Amherst, MA, 01003, USA*
[d]*Department of Linguistics, University of California, Los Angeles, 3125 Campbell Hall, Los Angeles, CA, 90095, USA*
[e]*Department of Computer Science, Colgate University, 13 Oak Drive, Hamilton, NY, 13346, USA*
[f]*Department of Linguistics and Center for Data Science, New York University, 60 5th Avenue, New York, NY, 10012, USA*

Abstract

Prediction has been proposed as an overarching principle that explains human information processing in language and beyond. To what degree can processing difficulty in syntactically complex sentences—one of the major concerns of psycholinguistics—be explained by predictability, as estimated using computational language models? A precise, quantitative test of this question requires a much larger scale data collection effort than has been done in the past. We present the Syntactic Ambiguity Processing Benchmark, a dataset of self-paced reading times from 2000 participants, who read a diverse set of complex English sentences. This dataset makes it possible to measure processing difficulty associated with individual syntactic constructions, and even individual sentences, precisely enough to rigorously test the predictions of computational models of language comprehension. We find that the predictions of language models with two different architectures sharply diverge

[1]To whom correspondence should be addressed. E-mail: kuanjunghuan@umass.edu
[2]S.A., M.K., C.M., G.P. are listed in alphabetical order, and so are the senior authors B.D. and T.L.

from the reading time data, dramatically underpredicting processing difficulty, failing to predict relative difficulty among different syntactic ambiguous constructions, and only partially explaining item-wise variability. These findings suggest that prediction is most likely insufficient on its own to explain human syntactic processing.

Keywords: Sentence processing, Prediction, Surprisal, Language models

---

1. Introduction

Language comprehension proceeds quickly and efficiently. A central factor invoked to explain this fact is *prediction*: by anticipating upcoming words, readers are enabled to rapidly integrate them into their interpretation of the sentence (Kutas et al., 2011). This explanation fits with the growing evidence for prediction as an organizing principle of linguistic cognition in particular (Dell et al., 2021; Pickering and Garrod, 2013) and the brain more generally (Bar, 2007). In parallel, much recent work has shown that language models—computational systems trained to predict the next word in a sentence—serve as a powerful foundation for language understanding by computers (Peters et al., 2018; Brown et al., 2020). The conjunction of these two trends has given rise to the hypothesis that there is a close correspondence between the predictive mechanisms used by language models and humans (Goldstein et al., 2022; Schrimpf et al., 2021). In this paper we ask, using predictability estimates derived from language models, to what extent human language comprehension at the sentence level can be explained by prediction.

The hypothesis that prediction plays a central role in human language comprehension is supported by comprehenders' pervasive sensitivity to word-level predictability, which is reflected by measures such as word-by-word processing difficulty (Ehrlich and Rayner, 1981; Staub, 2015) and the N400 electrophysiological response (Kutas et al., 2011). Traditionally, word predictability was estimated using the cloze task, in which participants were asked to provide the next word in a sentence (Taylor, 1953). As the quality of computational language models has improved, these models have been increasingly used as a proxy for human predictability (Smith and Levy, 2013; Goodkind and Bicknell, 2018; Goldstein et al., 2022). There is growing evidence that the processing difficulty on a word that can be attributed to its predictability, as estimated by a language model, is proportional to the word's surprisal (Hale, 2001; Levy, 2008), that is, the negative log probability assigned by the

language model to that word in context (Smith and Levy 2013; Wilcox et al. 2020; Shain et al. 2022; though see Brothers and Kuperberg 2021).

1.1. To what extent can predictability explain sentence processing difficulty?

While there is compelling evidence that word predictability affects human language comprehension, just how much of language comprehension difficulty can be explained using word predictability remains an open question. On what is perhaps the strongest view on this matter, word surprisal is a "causal bottleneck" that explains much of word-level processing difficulty (Levy 2008; Smith and Levy 2013). This strong view is appealing on parsimony grounds: Since prediction is independently necessary to explain findings from language comprehension and other cognitive domains, it is worthwhile to explore the extent to which it can account for findings that have traditionally been explained using other factors. This methodological approach has been invoked to qualitatively explain a number of phenomena in sentence processing. These phenomena, most of which are described in more detail below, include antilocality effects (Konieczny, 2000; Levy, 2008), garden path effects (Bever, 1970; Hale, 2001; Levy, 2013), the relative difficulty of object-extracted compared to subject-extracted relative clauses (Vani et al., 2021), and the so-called "ambiguity advantage effect" (Traxler et al., 1998).

These qualitative accounts of processing difficulty in specific syntactic phenomena join quantitative studies based on measurements taken while participants were reading natural texts; these studies, which have found that up to 80% of the explainable variance in word reading times and nearly 100% of the explainable variance in neural responses to sentences can be predicted by the internal vector representations of next-word-prediction language models (Schrimpf et al., 2021), were taken to further suggest that prediction can explain much of sentence comprehension (though for a note of caution about the interpretation of such studies, see Section 5.2 and Antonello and Huth 2023).

There are limits to the conclusions we can draw from studies that use materials from naturalistic sources such as newspaper articles, however. Such materials may contain predominantly simple, unchallenging structures, and at most a small number of low-frequency syntactic constructions (Futrell et al., 2021). Crucially, the predictions of cognitive theories often diverge most sharply in less frequent constructions (Levy, 2008; Levy et al., 2012); even if the corpus does occasionally contain such examples, they are likely to be vastly outnumbered by syntactically simple sentences, and as such will

have a negligible impact on the model's fit to reading times (for a similar argument in the case of language model evaluation, see Marvin and Linzen 2018).

Adopting a more targeted approach to the quantitative assessment of predictability as an explanatory account of syntactic processing difficulty, van Schijndel and Linzen (2021) tested the predictions made by surprisal for three types of *garden path sentences*. Such sentences involve incremental syntactic ambiguity that is ultimately disambiguated towards a less preferred (and typically less likely) structure; such sentences are said to "lead the reader down the garden path." For example, in (1a) below, the word *conducted* signals that the preceding material should be parsed as a reduced relative clause, a low probability syntactic analysis; compare this sentence to (1b), which is a minimally different sentence that does not display such ambiguity. Following prior work, we use the term *garden path effect* to refer to the amount of excess reading time triggered by the disambiguating word in (1a) relative to the baseline condition (1b), where the syntax of the sentence is instead disambiguated prior to the critical word.

(1) a. The experienced soldiers warned about the dangers **conducted** the midnight raid.

b. The experienced soldiers who were warned about the dangers **conducted** the midnight raid.

Under the surprisal hypothesis, the processing difficulty on the boldfaced words in (1a) can be *fully explained* by the fact that these words constitute a highly improbable continuation compared to the same words in (1b). In other words, for surprisal theory to serve as an adequate theory of processing difficulty in garden path sentences, it needs to predict not only the existence of garden path effects, but also their full magnitude. Van Schijndel and Linzen tested this hypothesis using surprisal estimates derived from long short-term memory (LSTM) recurrent neural network language models. They showed that while surprisal correctly predicted that reading times on the boldfaced words in (1a) are longer than the reading times on the same words in (1b), it predicted a much smaller excess processing difficulty on (1a) than empirically observed (for similar results for other linguistic constructions, obtained using the maze task, see Wilcox et al. 2021). Such substantial underestimation of processing difficulty by surprisal may indicate that additional processes, such

as syntactic reanalysis (Fodor and Ferreira, 1998; Paape and Vasishth, 2022), are recruited during the comprehension of syntactically complex sentences.

1.2. High-sensitivity model evaluation: The Syntactic Ambiguity Processing Benchmark

While van Schijndel and Linzen (2021) provide a blueprint for testing whether processing difficulty in complex sentences can be reduced to surprisal, the empirical scope of their work is limited, in a number of ways. First, they only examined three garden path constructions, out of the range of syntactically complex English constructions documented in the psycholinguistics literature. Second, they were unable to determine conclusively whether surprisal predicts the relative processing difficulty across different constructions: The two evaluation sets used by van Schijndel and Linzen, collected from 73 and 224 participants respectively, did not permit drawing statistically significant conclusions regarding the relative difficulty among the three garden path types. Third, again due to limited power, they only report results at the construction level, and did not examine whether surprisal can explain item-wise variability; this is despite the fact that, as we show below, language models' predictability estimates vary widely not only from construction to construction, but also from item to item in the same construction (cf. Garnsey et al. 1997; Frank and Hoeks 2019). Finally, their ability to compare processing difficulty across constructions was limited by the fact that each of the constructions was read by a different set of participants, precluding within-subjects comparisons.

This is a typical situation in psycholinguistics: Datasets from existing experiments with classic factorial designs, which enable researchers to carefully control irrelevant factors and isolate the comparisons of interest, typically involve a relatively small number of participants. Such datasets sometimes do not even afford enough power to test coarse, directional predictions at the construction level (Vasishth et al., 2018), let alone the precise quantitative predictions at the construction and item level that can be derived from language models. For all these reasons, a thorough empirical test of the surprisal hypotheses requires a new data collection effort.

Motivated by these issues, we present the Syntactic Ambiguity Processing (SAP) Benchmark, a large-scale dataset that consists of self-paced reading times from a range of constructions that have motivated psycholinguistic theories. This benchmark seeks to strike a balance between classic factorial

designs and broad-coverage model evaluation that prioritizes explaining item-level variability. Our goal is to create a dataset that will yield effect size estimates precise enough to evaluate the predictions of language models at the level not only of constructions but also individual items. Unlike most prior work, we have the same participants read all types of constructions included in the experiment; this makes it possible to carry out within-participant comparisons of the magnitude of effects across constructions. Overall, by including various syntactic phenomena in the same study, and analyzing reading times in the same way, we can address more directly the question of whether prediction can serve as a *unified* account for language comprehension. Beyond the specific theoretical question that we set out to address as to the scope of the explanatory power of predictability, we see this dataset as a standard yardstick against which any quantitative theories of human sentence processing can be evaluated.

1.3. Summary of the research questions addressed by this paper

In summary, we aim to address three central questions regarding prediction in language comprehension. First, we ask if surprisal can explain the full magnitude of processing difficulty in the constructions that have driven psycholinguistic theory development. Our dataset includes the three garden-path constructions examined by van Schijndel and Linzen (2021); this subset of the SAP Benchmark can be seen as a high-power replication of their work, with material that are more tightly matched across constructions (see Section 3.4). In addition to these three constructions, we also evaluate whether predictability can explain the relative difficulty of object-extracted relative clauses compared to subject-extracted ones, the ambiguity advantage in prepositional phrase attachment, and the ungrammaticality penalty in subject-verb agreement dependencies.

The second question we ask is whether surprisal can correctly predict the relative difficulty among the three garden path constructions. In van Schijndel and Linzen's study language models made predictions that appeared not to match the order of human processing difficulty across constructions, their analyses had limited statistical power to detect differences between constructions. This issue is addressed in the current large-scale study, which has 8000 observations per condition. Furthermore, in addition to the LSTM language model used by van Schijndel and Linzen, we also evaluate a more powerful language model based on the Transformer architecture. This makes it possible to examine whether our conclusions with regards to surprisal theory

are sensitive to the technical aspects of the model used to derive surprisal estimates (see Section 3.8.1).

Finally, we ask whether surprisal can explain itemwise variation in processing difficulty within the same syntactic construction. While broad-coverage modeling has demonstrated seemingly impressive predictivity (Schrimpf et al., 2021; Smith and Levy, 2013; Wilcox et al., 2020), evaluations of item-level predictivity on more targeted linguistic contrasts have only been made with relatively small sample sizes (Frank and Hoeks, 2019). In this study, we collect between 220 and 440 observations per item. As we show below, this results in effect sizes for individual items that are much more precise than has been possible before, and enables robust item-wise analyses.

2. Data Availability

The materials, reading time data, and analysis scripts are available at the following website: `https://github.com/caplabnyu/sapbenchmark`.

3. Methods

3.1. The Syntactic Ambiguity Processing Benchmark: Dataset construction

As we described in the Introduction, the SAP Benchmark is a large-scale dataset that serves two purposes. First, we use it to empirically evaluate the ability of surprisal to explain human comprehension difficulty in syntactically complex sentences; and second, we intend it serve as a resource for quantitatively evaluating other theories of sentence processing. In this section, we describe how the benchmark was constructed.

To ensure that we had sufficient statistical power to obtain tight estimates of construction-level effects as well as item-level effects, we collected data from 2000 participants. Participants were recruited using the crowdsourcing service Prolific. Participants read a range of critical stimuli using the self-paced reading paradigm (Just et al., 1982). The materials included seven distinct English constructions, grouped into four subsets. We also included filler sentences from a naturalistic corpus that did not include syntactically complex structures (Luke and Christianson, 2018). Importantly, there are arguments in the literature attributing processing difficulty in all of our target constructions to word-level predictability under surprisal theory (Hale, 2001; Levy, 2008; Vani et al., 2021; Wilcox et al., 2021). The constructions are

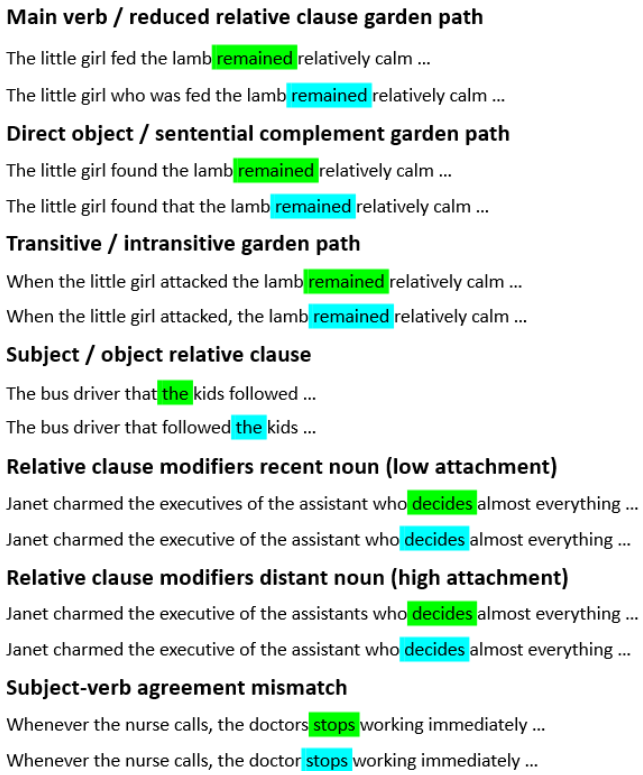

Figure 1: Effects of interest in the Syntactic Ambiguity Processing benchmark. Each sentence pair illustrates a construction tested in our dataset. An effect of interest is defined as the difference in reading times associated with a disambiguating or ungrammatical word, marked in green, minus the reading time associated with that same word in a context where it is grammatical and does not disambiguate the structure of the sentence, marked in turquoise.

exemplified in Figure 1. We motivate and describe each of the four subsets in turn.

The first subset includes three **classic garden path constructions** that generate reliable garden path effects: the Direct Object/Sentential Complement (NP/S) ambiguity (Frazier, 1979), the Transitive/Intransitive ambiguity (sometimes referred to as the NP/Z ambiguity; Frazier 1979), and the Main Verb/Reduced Relative ambiguity (Bever, 1970). These constructions have long been reported to differ in the severity of the garden path effect that occurs in each; this observation has been corroborated using reading time data for the NP/S and NP/Z constructions by Sturt et al. (1999) and Grodner et al. (2003). Sturt and colleagues created lexically matched item sets for the NP/S and NP/Z constructions; we extend this methodology to the MV/RR construction and create 24 lexically matched item sets for each of these three garden path constructions.

The second subset of items within the SAP Benchmark contained **relative clauses**. We constructed lexically matched subject-extracted relative clauses (SRCs) and object-extracted relative clauses (ORCs). In English, as in many other languages, ORCs are generally more difficult to process than SRCs (Lau and Tanaka, 2021). This difficulty is thought in part to reflect the relative unpredictability of ORCs (Chen and Hale, 2021; Hale, 2001; Staub, 2010; Vani et al., 2021). However, unlike the garden path constructions, the overall comprehension difficulty associated with ORCs has occasionally been argued to involve memory-related difficulties above and beyond the effects of predictability even by proponents of surprisal theory (Levy, 2013).

The third subset contained relative clause **attachment ambiguities**. In this construction, a relative clause (RC) can modify either of two noun phrases, a closer or more distant one. Including this subset in the benchmark allows us to contrast the processing of globally ambiguous relative clause attachment configurations and unambiguous relative clause attachment. Previous work has found a processing advantage associated with globally ambiguous RC attachment (the **ambiguity advantage effect**; Traxler et al. 1998, 2002; Van Gompel et al. 2005). Like garden path effects, this effect has been argued to arise from predictability: Unambiguous continuations are generally less predictable than ambiguous continuations in this context, and hence should be associated with greater processing difficulty (Levy, 2008).

The last subset examines the processing of **subject-verb agreement**. These sentences contain agreement errors that are caused by a mismatch between the inflectional features on a verb and those of its subject. Like

garden path sentences, agreement mismatches are triggered by material that is highly syntactically unlikely given the left context, and correspondingly cause processing difficulty (Wagers et al., 2009). Unlike garden paths, however, it is not possible to reanalyze these items to yield an acceptable structure: Under no reading or parse is the sentence well-formed.

The first three groups of items — classic English garden path constructions, relative clauses, and attachment ambiguities — can be seen as different instances of garden path effects: in each, the sentence is initially ambiguous between two syntactic analyses, and is later disambiguated at a critical point in the sentence. That critical point is highlighted in orange in Figure 1.

For all constructions, we estimated a specific *effect of interest* (EOI). In the three types of constructions that involve syntactic ambiguity, we defined the EOI as the processing time associated with the critical word that disambiguates the sentence, relative to that same word in a sentence where it does not. In the subject-verb agreement subset, the EOI was defined as the processing time on the verb when it mismatches the number features of the verb minus the processing time when it matches. A summary of the effects of interest for the seven constructions we tested is presented in Figure 1. Across all constructions, the EOI represents a key index of processing difficulty in this constructions, either the magnitude of the garden path effect in a given sentence pair (an *item*), or the slowdown in reading times associated with recognizing ungrammatical subject-verb agreement.

These EOIs are the target of our modeling efforts: They isolate the unique processing difficulty associated with a pair of sentences, controlling for lexical factors such as unigram log-frequency and length. We will consider a model successful to the extent that it can successfully predict the magnitude of our EOIs across constructions and across individual sentences.

3.2. Participants

We aimed to recruit a total of 2000 participants who spoke English as their first language. Participants were recruited on Prolific. Of the 2000 recruited, 1867 were speakers of North American English either from Canada or the United States. Due to an error in recruitment, 133 participants were recruited from the United Kingdom and 16 from other regions. After observing no evidence of difference in the results between participants from the UK and North America, we decided to include them in the final analysis. We excluded the 16 remaining participants. The age of all participants was between 18 and 45.

Our exclusion criteria are detailed in our preregistration document, available at `https://osf.io/9865s`. We excluded from analysis all participants whose accuracy on the comprehension questions for the fillers was below 80%. This resulted in the exclusion of an additional 179 participants.

3.3. Procedure

Participants in the experiment read our sentences in a self-paced reading paradigm. In this paradigm, the words of the sentence are first replaced by dashes. The participant presses a key to reveal the words of the sentence one at a time, with each word replaced by dashes once the participant moves on from it. The time taken to proceed to the next word is commonly seen as an indicator of the difficulty of processing the current word. In an experimental session, a participant read 92 sentences. This included 52 sentences of interest and 40 fillers, with four practice trials in the beginning. Each sentence was followed by a comprehension question. An experimental session lasted approximately 25 minutes on average.

To avoid syntactic adaptation, only four items from each condition were presented to each participant (counterbalanced using a Latin square). Items were generally presented in a random order, subject to the constraint that no two consecutive trials come from the same condition.[3]

3.4. Materials

We created 24 items for each subset, except for the subject-verb agreement subset, which had 18 items. For the classic garden path subset and the agreement subset, we created new materials. The materials for the ambiguity advantage subset were based on Dillon et al. (2019), and the materials for the relative clause subset were based on Staub (2010); we made small modifications to both subsets to ensure that all vocabulary items were included in the vocabulary of the LSTM language model we used (Gulordava et al., 2018). Filler items were drawn from the Provo Corpus (Luke and Christianson, 2018). See Appendix D for a full list of the items.

The strength of garden path effects is influenced by various factors such as plausibility and verb subcategorization frequencies. To control for these

[3]Because of an error in implementing this pseudorandomization scheme, this constraint was not enforced for a small number of participants. To account for this, we excluded all trials that immediately followed another trial from the same condition (1670 out of 104000 trials). This decision was not preregistered.

and other factors, we took a number of precautions during stimuli creation. We first searched the Corpus of Contemporary American English (COCA; Davies 2019) for verbs with at least one attested use which mirrored one of our garden-path constructions. Specifically, we searched for verbs where the less frequent parse was attested in a locally ambiguous form. This process helped to ensure that garden path items in the ambiguous condition were in fact ambiguous.

For example, the garden path effect in a Transitive/Intransitive construction such as 'After the woman moved the mail disappeared mysteriously from the delivery system' depends on the ambiguous transitivity of 'move' and the lack of a comma between 'moved' and 'the mail'. As such, a verb was only eligible for use in the Transitive/Intransitive frame if the less frequent intransitive use was attested without such a comma. We enforced a similar constraint on Direct Object/Sentential Complement verbs, where the less frequent parse includes a sentential complement and local ambiguity arises from the absence of a complementizer. Lastly, we only included Main Verb/Reduced Relative verbs that were attested in a reduced relative clause lacking the complementizer and copula ('who/that was').

We queried COCA for all sentences matching the pattern **DP VERB DP** for each verb considered (e.g., DP *moved* DP). All of these sentences were then parsed using the spaCy natural language processing library (Honnibal and Montani, 2017) and the disambiguating verb was labelled as a matrix verb or not. The output of spaCy was then manually verified and corrected. In the final set of 24 garden path items, we used 12 unique verbs for the Transitive/Intransitive and Direct Object/Sentential Complement conditions and 9 for the Main Verb/Reduced Relative condition. Consequently, each Transitive/Intransitive and Direct Object/Sentential Complement verb occurred twice in our stimuli, while six Main Verb/Reduced Relative verbs occurred twice and the remaining three occurred four times. Crucially, any repetition of a verb occurred in an entirely different frame. Our Latin square counterbalancing scheme ensured that every ambiguous verb and contextual frame seen by a given participant was unique within an experimental session. Each specific item was seen by between 220 and 440 participants.

The items in the agreement subset of the benchmark were partially derived from the Transitive/Intransitive conditions of the garden path subset, as follows. For each agreement item we repeated the ambiguous verb, disambiguating verb, and first word of the spillover region from a garden path item in the Transitive/Intransitive condition. For instance, the ungrammatical

agreement item 'When the magician moves, the cards disappears mysteriously from his assistant's hand' corresponds to the previous Transitive/Intransitive example 'After the woman moved the mail disappeared mysteriously from the delivery system'. This constraint was enforced to allow for a closer comparison of reading times across these subsets of the benchmark. Due to the difficulty of satisfying all of these constraints, the agreement subset was limited to 18 items.

A number of additional standard controls for reading experiments were imposed across both subsets. The disambiguating region was always of six or more characters to prevent the skipping of target words in a future eyetracking-during-reading version of the benchmark. We also included a spillover region of two words to capture spillover effects during reading. The total length in characters of each sentence was limited such that the sentences fit in a single line. Finally, we checked at all stages that the vocabulary of our stimuli was a subset of the vocabularies of both the Penn Treebank Corpus and the multilingual Wikipedia training data from (Gulordava et al., 2018). This was done to ensure that a wide range of models could be tested using the SAP Benchmark, including those trained on supervised parses from the Penn Treebank.

3.4.1. Norming

Multiple rounds of norming were conducted to ensure high levels of plausibility for the unambiguous garden path and grammatical agreement items (e.g., 'After the woman moved, the mail disappeared mysteriously from the delivery system.'). This ensured that the ultimate parse for each item was acceptable despite the difficulty associated with parsing a garden path construction. We included a number of implausible fillers in the norming experiments. These fillers provided a highly implausible baseline so that participants would see a wide range of acceptability. All judgments were provided using a 7-point Likert scale. Adjustments were made to items in between multiple rounds of norming until the mean plausibility rating of the whole items for each condition exceeded 5 points on the 7-point scale. The final round of norming satisfying these restrictions included judgements from 68 participants. Norming studies were run on *PCIbex* and advertised on *Prolific* at a rate of $11 per hour of participation.

In addition to norming plausibility of each full sentence, we also normed the plausibility of parts of each sentence, such as the plausibility of the temporary garden path interpreation. The results of this supplementary

norming were used for a separate analysis. The details are explained in Appendix B.

3.5. Additional data exclusion criteria

In addition to subject-level exclusions described above, we also excluded from analysis all observations at the critical positions with RTs greater than 7000 milliseconds. We reasoned that such long latency between key presses is unlikely to reflect normal reading processes. We determined the precise value of this cutoff based on the RT distributions from the first 150 participants we collected. This pre-processing step resulted in the exclusion of less than 0.03% of the critical data points.

3.6. Estimating the Effects of Interest

For each construction in the SAP Benchmark, we used Bayesian mixed-effects regression models to estimate both the empirical human comprehension difficulty and the comprehension difficulty predicted by surprisal. We used the BRMS package in R to fit these models (Bürkner, 2017). In this section we motivate our analysis decisions and describe the structure of the models.

3.6.1. Analyzing raw RTs

In all analyses below, we analyze raw, rather than log RTs. The choice to not log transform our dependent variable may seem unusual: RTs in general are not normally distributed, with a heavy right skew (e.g., Frank et al. 2013). And since regression models assume that residuals are normally distributed, RT distributions therefore violate this assumption. Log transformation can address this issue. While this practice does reduce rightward skew, transformation of our dependent variable has undesirable consequences in the context of predicting RTs from surprisal estimates. Surprisal theory assumes that the effect of surprisal is additive with the other factors affecting RTs, as illustrated by the formulae below, where $\mathcal{S}$ is the surprisal of a word given some context, and $\mathcal{F}_1, \ldots, \mathcal{F}_n$ are other factors that influence reading time of a word, such as length or log-frequency.

$$\begin{aligned} RT(word \mid ambiguous\ context) = k * \mathcal{S}(word \mid ambiguous\ context) \\ + x_1 * \mathcal{F}_1(word) + ... + x_n * \mathcal{F}_n(word) \end{aligned}$$

$$RT(word \mid unambiguous\ context) = k * \mathcal{S}(word \mid unambiguous\ context) + x_1 * \mathcal{F}_1(word) + ... + x_n * \mathcal{F}_n(word)$$

Since the terms $\mathcal{F}_1, \ldots, \mathcal{F}_n$ are identical in both ambiguous and unambiguous contexts, we can express surprisal's influence on comprehension difficulty on a linear scale as follows (here, $\mathcal{D}$ denotes comprehension difficulty):

$$\begin{aligned}\mathcal{D}(word) &= RT(word \mid ambiguous\ context) - RT(word \mid unambiguous\ context) \\ &= k * \mathcal{S}(word \mid ambiguous\ context) - k * \mathcal{S}(word \mid unambiguous\ context) \\ &= k * \Big(\mathcal{S}(word \mid ambiguous\ context) - \mathcal{S}(word \mid unambiguous\ context)\Big)\end{aligned}$$

By contrast, when RTs are log-transformed, surprisal's effect on difficulty is given by the equation below, where $\mathcal{F} = x_1 * \mathcal{F}_1(word) + ... + x_n * \mathcal{F}_n(word)$:

$$\begin{aligned}\mathcal{D}(word) &= \log(k * \mathcal{S}(word \mid ambiguous\ context) + \mathcal{F}) \\ &\quad - \log(k * \mathcal{S}(word \mid unambiguous\ context) + \mathcal{F}) \\ &= \Big(\log(k * \mathcal{S}(word \mid ambiguous\ context)) * \log(\mathcal{F})\Big) \\ &\quad - \Big(\log(k * \mathcal{S}(word \mid unambiguous\ context)) * \log(\mathcal{F})\Big)\end{aligned}$$

Crucially, as this equation shows, when we log-transform RTs, the influence of surprisal is *dependent* on the other factors that influence comprehension difficulty, thus violating the additive assumption made by surprisal theory. This observation strongly motivates our decision to analyze raw RTs instead of log-transformed RTs, even at the cost of violating the normality assumption for our statistical regression models: Our key regression coefficients with a log-transformed dependent variable would not be readily interpretable.

While acknowledging the violation of the normality assumption, we believe it implausible that this will materially impact the conclusions we draw in this paper given our sample size: Recent simulations have found that violations of the normality assumption led to very little bias in regression coefficient estimates and no change in power for $n >= 1000$ (Knief and Forstmeier, 2021).

3.6.2. Priors for the Bayesian models

Table 1 lists the weakly informative priors we use in our Bayesian regression models. Given the large number of participants in our dataset, the choice of the prior did not substantially influence our estimates: All of the coefficients estimated using the Bayesian models were nearly identical to the coefficients estimated using frequentist linear mixed-effects models. The table also provides an intuition for the set of values on which most of the prior probability mass is concentrated.

Class	**Distribution**	**Intuition**
Intercept	Normal(300, 1000)	Under treatment coding, the intercept is the mean RT per word in the baseline condition. This is unlikely to be greater than 2000 ms.
Coefficients	Normal(0, 150)	The mean difference between any two conditions is unlikely to be greater than 250 ms or less than −250 ms.
Standard deviation (random effects)	Normal(0, 200)	The standard deviations of the random slopes and intercepts are unlikely to be greater than 350 ms.
Standard deviation (residuals)	Normal(0,500)	The standard deviation of the residuals is unlikely to be greater than 800 ms.

Table 1: The priors we used for our Bayesian mixed-effects models.

3.7. Estimating Empirical EOIs

We fit the following four sets of Bayesian mixed-effects models, one for each subset of the SAP Benchmark, and used the models to estimate the 95% posterior credible interval over the effect size for each construction, and for

each item. Construction-level EOIs were derived from the posterior estimates of the model's fixed effects. Item-specific estimates were computed from the random effects corresponding to each item. For each subset, we fit three models: One at the critical disambiguating word, one at the immediately following word (the first spillover word), and one at the word following that word (the second spillover word).

We describe the models using R formula notation. The models were fit with four chains. Each chain was run for 6000 iterations by default, with half of the iterations discarded as warm-up samples. The number of iterations was increased when necessary. The between-chains variability, as indexed by $\hat{R}$, was lower than 1.05 for all our models, indicating convergence (Nalborczyk et al., 2019).

Classic English garden path constructions. For this subset, we used the following formula:

$$\begin{aligned} RT \sim\ & ambiguity * sentence\_type + \\ & (1 + ambiguity * sentence\_type \,||\, item) + \\ & (1 + ambiguity * sentence\_type \,||\, participant\_id) \end{aligned}$$

Here, *ambiguity* was a predictor with two levels, ambiguous and unambiguous, with unambiguous coded as the baseline (unambiguous coded as 0, ambiguous coded as 1). The predictor *sentence_type* had three levels: Transitive/Intransitive, Direct Object/Sentential Complement, and Main Verb/Reduced Relative. We used treatment coding with Main Verb/Reduced Relative coded as the baseline with two contrasts, (0, 1, 0) and (0, 0, 1), where 1 indexes Direct Object/Sentential Complement and Transitive/Intransitive respectively. There were three EOIs associated with this subset, one for each of the garden path effects. Given this coding scheme, the Main Verb/Reduced Relative garden path effect can be directly obtained from the posterior of the ambiguity contrast coefficient. The Direct Object/Sentential Complement garden path effect can be recovered by adding the ambiguity contrast coefficient and the first interaction coefficient; this was done for each posterior sample first, and then averaged across all the samples from the four chains. The standard error is the standard deviation of this aggregated posterior sample distribution. Finally, the Transitive/Intransitive garden path effect can be recovered by adding the ambiguity contrast coefficient and the second interaction coefficient. As mentioned above, we fit three separate models, one for the disambiguating verb and one for each of the following two words.

Relative clauses. In this subset, the critical word of interest was the determiner, which occurred at different linear positions across conditions.[4] To account for any independent effect that word position may have on RTs, we first corrected for the effect of position by fitting the following linear mixed-effects model to the filler sentences. We then used this model to regress out word position for the critical sentences (Van Dyke and Lewis, 2003):

$$\begin{aligned} RT \sim\ & scale(position) + \\ & (1 + scale(position) \mid participant) \end{aligned}$$

After residualizing RTs in this fashion, we fit three models, one each for the determiner, noun and verb in the relative clause, using the following model:

$$\begin{aligned} RT\_corrected \sim\ & RC\_type + \\ & (1 + RC\_type \mid\mid item) + \\ & (0 + RC\_type \mid\mid participant) \end{aligned}$$

RC_type was a predictor with two levels—subject RC and object RC—with subject RC coded as the baseline (subject RC coded as 0, object RC coded as 1). There is one effect of interest associated with this subset: the difference in reading times on the critical word between subject and object RCs.

Attachment ambiguities. We used the following formula:

$$\begin{aligned} RT \sim\ & ambiguity * height + \\ & (1 + ambiguity + height \mid\mid item) + \\ & (1 + ambiguity + height \mid\mid participant) \end{aligned}$$

This subset had three types of sentences: ambiguous sentences, unambiguous sentences with high attachment, and unambiguous sentences with low attachment. The *ambiguity* predictor had two levels: ambiguous (coded with $2/3$) and unambiguous (coded with $-1/3$ for the two unambiguous conditions).

[4]This is a departure from our preregistration document, which incorrectly identifies the verb as the critical region. We follow Hale (2001) and Levy (2008) in expecting the excess processing cost of object RC to arise at the word disambiguating object RC from subject RC, which in our experimental sentences is the determiner.

The factor *height* had two levels: high (coded as 1/2) and low (coded as −1/2). There are two EOIs associated with this subset: high attachment garden path and low attachment garden path. Given this coding scheme, the high attachment effect can be recovered by adding half of the height contrast coefficient to the negative ambiguity contrast coefficient. The low attachment effect can be recovered by subtracting half of the height contrast coefficient from the negative ambiguity contrast coefficient.

Subject-verb agreement. We used the following formula:[5]

$$\begin{aligned} RT \sim\ & \mathit{grammaticality} + \\ & (1 + \mathit{grammaticality} \,||\, \mathit{item}) + \\ & (1 + \mathit{grammaticality} \,||\, \mathit{participant}) \end{aligned}$$

In this subset, we considered two kinds of sentences: grammatical, unambiguous sentences from the Transitive/Intransitive subset, and ungrammatical versions of those sentences containing an agreement error. The predictor *grammaticality* had two levels: grammatical and ungrammatical. The baseline was grammatical (grammatical coded as 0, ungrammatical coded as 1).

3.8. Estimating predicted EOIs

We generated the predicted EOIs in two steps. First, we derived surprisal values for the critical regions in all of our experimental items from our language models. Second, we estimated "conversion factors", coefficients that link surprisal estimates to reading times.

3.8.1. Computing language model surprisal

We derived surprisal values from two publicly available neural-network language models that differed in both architecture and training data. The first model we used, released by Gulordava et al. (2018), was based on the Long Short-Term Memory (LSTM) recurrent neural network architecture (Elman, 1991; Hochreiter and Schmidhuber, 1997). It was trained on approximately 80 million words of Wikipedia text. The second model we used was the 117-million parameter variant of GPT-2 (GPT-2 small; Radford et al. 2019); this

[5]Due to model convergence issues, we had to simplify the original preregistered formula and fitted a separate model separately for each word position instead.

model is based on the Transformer architecture (Vaswani et al., 2017), and was trained on approximately 40 GB of data scraped from the Web. These models have been shown in previous work to display substantial awareness of the constraints of English grammar, such as subject-verb agreement, garden path constructions and filler-gap dependencies (Gulordava et al., 2018; Hu et al., 2020; Warstadt et al., 2020; van Schijndel and Linzen, 2021), and as such are promising candidates for modeling human syntactic expectations.

3.8.2. Linking surprisal to reading times

We followed the methodology that van Schijndel and Linzen (2021) used to predict human reading times from model-based surprisal. Specifically, we first fit a linear mixed-effects model **to our filler items**. The goal of this model is to estimate the linear relationship between surprisal and reading time; the coefficient (slope) of this linear relationship, according to surprisal theory, should be the same in syntactically simple and complex sentences. In addition to surprisal-based predictors, this model included as predictors word position, word length, unigram frequency, and the interaction between word length and unigram frequency. We also included random intercepts by participant and by item, as well as a random slope for surprisal by participant. To account for spillover effects in self-paced reading (Mitchell, 1984), we included these predictors not only for the current word but also for the three preceding ones. All predictors were centered and scaled across the full dataset. We fit two of these linear mixed-effects models to the fillers, one for each of the language models. We excluded any words for which any of our predictors were not defined; this was the case, in particular, for the first three words of a sentence, which are not preceded by a three-word spillover context. We also followed prior work (Smith and Levy, 2013, for example) in excluding the final word of each sentences, as these words display wrap-up effects that are beyond the scope of our modeling goals (Just et al., 1982).

The resulting models (Table 2) offer a set of **conversion factors** that estimate how reading time on the fillers co-vary with surprisal and the other predictors.

LSTM		**GPT-2**	
Predictor	$\hat{\beta}$	*Predictor*	$\hat{\beta}$
Word position	-1.49	Word position	-1.26
Surprisal_{w_n}	1.38	Surprisal_{w_n}	1.62
$\text{Surprisal}_{w_{n-1}}$	1.18	$\text{Surprisal}_{w_{n-1}}$	1.62
$\text{Surprisal}_{w_{n-2}}$	0.17	$\text{Surprisal}_{w_{n-2}}$	0.83
$\text{Surprisal}_{w_{n-3}}$	0.54	$\text{Surprisal}_{w_{n-3}}$	0.34
Log-Freq_{w_n}	1.02	Log-Freq_{w_n}	0.43
$\text{Log-Freq}_{w_{n-1}}$	0.57	$\text{Log-Freq}_{w_{n-1}}$	0.07
$\text{Log-Freq}_{w_{n-2}}$	-0.32	$\text{Log-Freq}_{w_{n-2}}$	-0.33
$\text{Log-Freq}_{w_{n-3}}$	1.30	$\text{Log-Freq}_{w_{n-3}}$	0.89
Length_{w_n}	11.3	Length_{w_n}	9.53
$\text{Length}_{w_{n-1}}$	12.6	$\text{Length}_{w_{n-1}}$	10.9
$\text{Length}_{w_{n-2}}$	3.46	$\text{Length}_{w_{n-2}}$	2.73
$\text{Length}_{w_{n-3}}$	1.90	$\text{Length}_{w_{n-3}}$	1.46
$\text{Freq}\times\text{Length}_{w_n}$	-0.69	$\text{Freq}\times\text{Length}_{w_n}$	-0.51
$\text{Freq}\times\text{Length}_{w_{n-1}}$	-0.87	$\text{Freq}\times\text{Length}_{w_{n-1}}$	-0.69
$\text{Freq}\times\text{Length}_{w_{n-2}}$	-0.31	$\text{Freq}\times\text{Length}_{w_{n-2}}$	-0.23
$\text{Freq}\times\text{Length}_{w_{n-3}}$	-0.20	$\text{Freq}\times\text{Length}_{w_{n-3}}$	-0.14

Table 2: Coefficient estimates for the models fit to the fillers; for example, $\text{Surprisal}_{w_{n-1}}$ indicates the effect in milliseconds of each additional bit (unit) of surprisal on reading times on word n. Note that the models reported in this table used uncentered and unscaled variables for ease of interpretation and comparability with previous studies. Shaded cells indicate an effect significant at the $p < 0.05$ level.

3.8.3. Generating predicted reading times

We then used these conversion factors to generate predicted reading times for each of the critical subsets. We additionally fit a **No-surprisal baseline**: a mixed-effects model that included only our non-surprisal factors (word position, word length, unigram frequency, and the interaction between length and frequency). This model was fit using the same process outlined above. When assessing how well surprisal predicts the magnitude of our effects of interest, the difference between this baseline and the models that include surprisal provides a conservative estimate of how much of the empirical garden path effect the addition of surprisal accounts for—in other words, it quantifies

how much better the surprisal-based model is at predicting the magnitude of the garden path effect compared to a model that includes only unigram statistics (and their spillover effects).

3.9. Comparing empirical and predicted EOIs

To evaluate whether our empirical estimates of processing difficulty *at the construction level* align with language- model-derived surprisal, we fit the Bayesian mixed-effect models described in Section 3.7 to both the empirical and predicted data. Then, we compared the resulting coefficients from these two sets of models.

To evaluate how well our surprisal estimates predict *item-wise variation*, that is, how well the surprisal on a given item predicts the EOI on that item, we estimated the uncertainty of this correlation coefficient for each construction using a Monte Carlo-based approach that leveraged the itemwise posterior EOI estimates, as follows. We independently sampled one observation from the posterior distribution of each item's EOI, as well as one observation from the corresponding model-based prediction for the EOI. This resulted in two numbers for each item. We then computed the correlation between the two quantities—empirical and predicted—across all items within a construction to yield that construction's correlation coefficient. We repeated this procedure 1000 times. We did this separately for each of the language models and for the No-surprisal baseline model.

Any correlation coefficient should be interpreted in the context of the intrinsic noise in our reading time measures, which limits the highest possible correlations that could be observed (Schrimpf et al., 2021). We estimated the explainable variance of our empirical dataset for each of the four experimental subsets by running 15 split-half reliability analyses. In each of the 15 iterations of this procedure, the 2000 participants were randomly split into two halves. Each half of the dataset was then entered into a frequentist linear mixed-effect model with the same structure as that used in the main analyses, yielding point estimates of item-level processing difficulty for each effect of interest (two point estimates for each item, each based on 1000 different participants). We then computed the correlation between the two sets of item-level estimates within each effect of interest. The average of these 15 correlation coefficients was then entered into the Spearman-Brown prophecy formula (Brown, 1910) to calculate the corrected reliability coefficient. We used this predicted reliability effect as an estimate of the highest possible correlation (Vul et al., 2009). The item-level correlation between the empirical and predicted effects for

each EOI was eventually divided by this ceiling to compute the proportion of explainable variance that was in fact explained.

We also calculated a comparable noise ceiling for our filler items. This was done similarly to the split-half analysis, with one exception: The model only contained a fixed intercept, a random participant intercept, and a random word intercept. That is, instead of treating each filler sentence as one item, each *word* in a sentence is treated as a unique item, as now this is a broad-coverage approach. For each iteration, 24 words out of 498 words were randomly selected. The split-half correlation thus is the correlation between the 24 word RTs estimated from a subset of 1000 participants and those estimated from the remaining 1000 participants.

4. Results

4.1. Comprehension question accuracy

Accuracy on the comprehension questions for the fillers was high (mean across subjects = 91.4%, min = 80%), indicating participants were paying attention to the reading task. For our critical items, we designed some comprehension questions to specifically target successful resolution of the garden path. For example, for *The little girl fed the lamb remained relatively calm despite having asked for beef*, the comprehension question targeting ambiguity resolution was *Did the girl feed the lamb?*. The remaining comprehension questions targeted other aspects of the sentence.

Accuracy on these questions varied across constructions, with Transitive/Intransitive, Low Attachment and Main Verb/Reduced Relative displaying the lowest accuracy. The low accuracy associated with these three constructions is consistent with earlier findings (Christianson et al., 2001; Dillon et al., 2019; Prasad and Linzen, 2021). We present the full accuracy data in Appendix A.

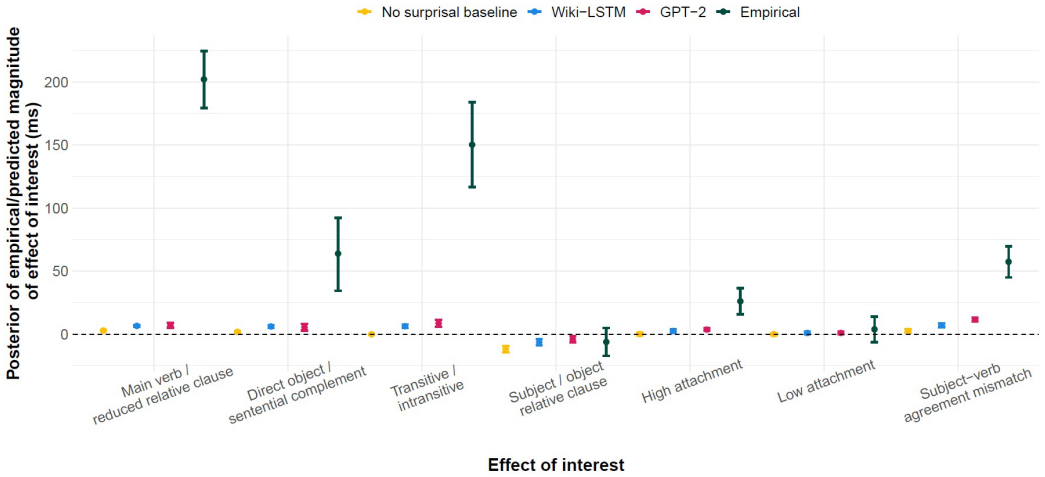

Figure 3: Empirical and predicted effects of interest at the first spillover region for all seven constructions in the SAP Benchmark . Empirical effects were estimated from a Bayesian mixed-effects regression model fit to raw RTs on the word that indexed the effect of interest. Error bars represent the 95% posterior credible interval on the construction-level size of this effect. Predicted effects were estimated from another Bayesian mixed-effects regression model with the same structure, whose coefficients were estimated from the filler items.

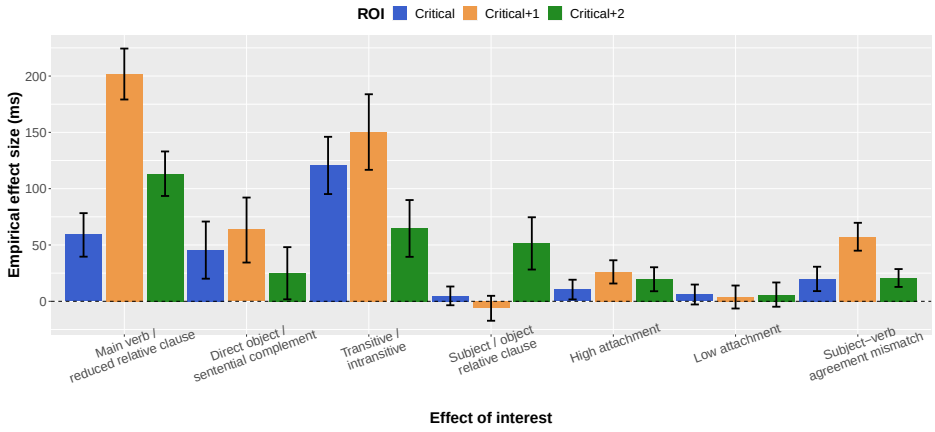

Figure 2: Posterior estimates of effect sizes at the three regions of interest for each effect of interest. Error bars represent 95% credible intervals. Note that for the Object vs. Subject Relative Clause construction, the critical words in the subject and object relative clauses are reversed. In this construction, we treat the determiner as the critical ROI, the noun as the first spillover region, and the verb as the second.

4.2. Construction-level reading times

Figure 2 presents the average effect of interest at the critical disambiguating word and the two immediately following spillover words, for each construction in our dataset. Here, we focus on the effects of interest on the first spillover word, rather than at the critical word. We focus on this region as for most effects of interest it showed the largest effect. Note that the two constructions where the EOIs were not the largest on this position are Object vs. Subject Relative Clause and Low Attachment. However, in both cases, we note that the effect was not present even at the critical disambiguating word.

We found robust garden path effects in four out of our six garden path constructions, with the largest effect in Main Verb/Reduced Relative (202.1 ms [179.2–224.4]; brackets indicate 95% credible intervals) and the second largest in Transitive/Intransitive (150.2 ms [116.7–183.8]). The garden path effects for Direct Object/Sentential Complement and the high attachment RC ambiguities were of smaller magnitudes (Direct Object/Sentential Complement: 63.9 ms [34.4–92.1]; High Attachment: 26.9 ms [15.8–36.4]). Finally, the ungrammaticality effect for Agreement Violations was somewhat smaller than the largest garden path effects but highly robust (57.4 ms [44.9–69.7]). The credible intervals for Object vs. Subject Relative Clause and Low Attachment overlapped with zero.

This pattern of results is consistent with four previously observed patterns. First, disambiguation is harder in Transitive/Intransitive than Direct Object/Sentential Complement (Sturt et al., 1999). Second, relative clauses with high attachment, but not low attachment, result in processing difficulty (Swets et al., 2008). Third, outright subject-verb agreement mismatch reliably slows down reading times (Wagers et al., 2009). Fourth, there was no reliable object relative clause difficulty at the determiner or noun position, as in prior self-paced reading studies (Grodner and Gibson, 2005). In addition to these previously established patterns, we demonstrated that disambiguation is harder in sentences with the Main Verb/Reduced Relative ambiguity, compared to sentences with the Transitive/Intransitive or Direct Object/Sentential Complement ambiguities. This establishes a difficulty ranking across these three widely studied garden paths for the first time in a within-items design.

The time course of the Object vs. Subject Relative Clause contrast is more complex. We followed Staub (2010) in comparing similar words across SRCs and ORCs, despite the fact that these words occur in a different linear order across the two conditions. In this analysis, we found no contrast between

SRCs and ORCs at the determiner or noun position. Instead, we saw slower RTs for ORCs compared to SRCs the verb position (see Figure 2). The timecourse of this effect appears to be inconsistent with the predictions of surprisal theory, which predict that the effect should localize to the subject noun phrase in an ORC construction (Hale, 2001). Instead, it appears to be more consistent with theories that attribute the difficulty with ORCs to difficulty integrating a distant argument at the verb (Gibson, 1998).

It is difficult to interpret this apparent time course effect too strongly, however. The effect we see at the verb position could reflect processing difficulty associated with the subject noun phrase showing up at the following region. Such spillover effects are common in self-paced reading. It is also possible that the lack of an effect at this position reflects spillover effects from the preceding context, which differed across the two conditions. This conjecture is supported by the observation that our no-surprisal baselines (whose effects derive solely from the spillover of unigram lexical effects from previous words) predicted a *negative* effect at the determiner and the noun regions. Since we did not see a negative effect in the empirical reading times in these regions, it is possible that the slowdown attributable to surprisal in the ORC conditions cancels out the speedup attributable to these other spillover factors. These results may therefore be consistent with other studies using reading paradigms that are less subject to spillover effects, which have observed processing costs at both the determiner (Vani et al., 2021) and verb position (Staub, 2010) in ORCs.

4.3. Variability across items

In most of the constructions, there was substantial variability across items in the size of the EOI. We used a split-half analysis to determine how reliable this item-level variability was (for details, see Section 3.9). The results varied by construction (see Table 3).

Effect of Interest	*Noise ceiling*
Main Verb/Reduced Relative	0.81
Direct Object/Sentential Complement	0.84
Transitive/Intransitive	0.82
Object vs. Subject Relative Clause	0.56
High Attachment	0.44
Low Attachment	0.18
Agreement Violations	0.45
Fillers	0.99

Table 3: Noise ceiling estimates based on Spearman-Brown-corrected split-half reliability for each effect of interest.

For the classic garden path constructions, Spearman-Brown-corrected split-half reliability (Brown, 1910) was quite high, all above 0.81. This indicates that there are highly stable item-wise EOIs for many of our target constructions. Reliability estimates were lower for the agreement errors and for the high and low relative clause attachment subsets, ranging from 0.18 to 0.45. We take the lower reliability by items here to reflect the fact that in the high and low attachment conditions the effects of interest were much smaller overall or nonexistent to begin with.

The extent of item-level variability differed across the constructions. We quantified this difference by computing *coefficients of variation* (CoVs), which capture the ratio of the standard deviation of the items' effect sizes to the mean effect size of the construction (see Figure 4).

The two constructions with the highest CoVs were Object vs. Subject Relative Clause and Low Attachment; this is due to the denominator (the mean effect of the construction) being close to zero. In the Direct Object/Sentential Complement and High Attachment constructions, only a subset of the items resulted in garden path effects that were distinguishable from 0 ms (about a third for Direct Object/Sentential Complement and about a half for High Attachment). In the items that did yield garden path effects, the magnitudes were generally large, with some items resulting in garden path effects as large as 100 ms. These constructions are the ones associated with higher CoVs.

For the Transitive/Intransitive construction, every item resulted in a garden path effect statistically greater than 0 ms. Yet even in this construction, there was a considerable item-level variability with effects ranging from 59.2

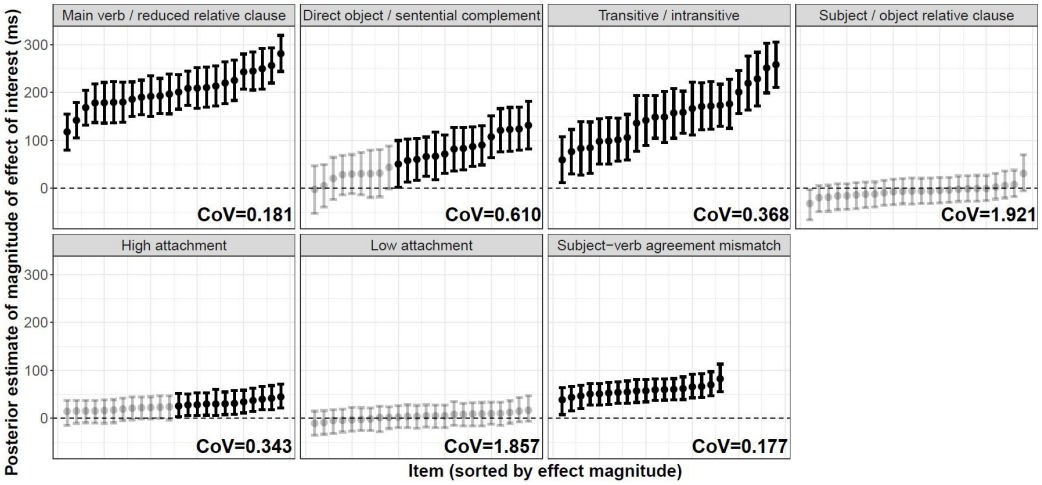

Figure 4: Empirical effects of interest for each individual item in all seven constructions in the SAP Benchmark. All effects were estimated from a Bayesian mixed-effects regression model fit to raw RTs on the word that indexed the effect of interest. Error bars represent the 95% posterior credible interval on the item-level size of this effect. The Coefficient of Variation (CoV) is the ratio between the standard deviation and the mean.

ms [12.1–107.5] (a 14.4% increase in reading time) to 258.3 ms [210.7–305.4] (a 58% increase in reading time). The Main Verb/Reduced Relative items likewise all showed a garden path effect statistically greater than 0 ms, though the variability across items was lower than for Transitive/Intransitive, with most effect sizes around 200 ms. Crucially, this item-level variability was not fully explained by easy-to-interpret variables like local-phrase plausibility or verb subcategorization bias (see Appendix B), making it an important target for modeling. Finally, in the Agreement Violations construction, the item-level effects were the least variable across individual sentences (hence the lowest CoV): While every item showed an effect of agreement violation, the effects only ranged from 38.6 ms [7.6–63.8] to 82.5 ms [55.6–113].

The differences in variability between constructions are not a simple by-product of differences in construction-wide effect sizes: Agreement Violations and Direct Object/Sentential Complement have similar mean effect sizes, but the former shows much smaller variability than the latter. Neither is it the case that the magnitude of item-level EOI is related to the absolute RT of the corresponding word in the unambiguous/grammatical sentence (see Appendix C).

4.4. Comparison to language model surprisal

As described above, we fit three linear mixed-effects models to the reading time data in our filler items. We then use these mixed-effects models to predict reading times at each word in our critical items, and from those predicted reading times we compute each model's prediction for the location, direction, and magnitude of each EOI (for additional details, see Section 3.8.1). The rest of this section reports the findings of this analysis.

Language model surprisal predicts the existence of human processing difficulty, but not its magnitude. Surprisal from both language models predicted the location and direction of most of the effects of interest tested, with the exception of the ORC/SRC condition: here both language models predicted a *negative* garden path effect, but such an effect was not seen in the human data (Figure 3). We suspect that this negative effect reflects differences in the unigram frequency of the pre-critical region, which was unmatched across the two conditions: No-surprisal models that only used lexical factors and their spillover predicted an even more dramatic negative difference in this EOI.

At the same time, the models failed to accurately predict the empirically observed rank order of the observed EOIs across constructions: the average garden path in Main Verb/Reduced Relative was greater than in Transitive/Intransitive, which was in turn greater than in Direct Object/Sentential Complement. However, the credible intervals for the Transitive/Intransitive, Direct Object/Sentential Complement, and Main Verb/Reduced Relative EOIs predicted by the two language models all overlapped.

In most constructions, we saw a clear quantitative misalignment between model predictions and the empirical data: Even when surprisal predicted an effect in the correct direction and at the correct position, the predicted effect size was orders of magnitude smaller than the empirically observed one. For example, in Main Verb/Reduced Relative, the observed effect of interest was 202.1 ms [179.2–224.4], whereas the predicted one was 6.6 ms [5.8–7.3] for Wiki-LSTM and 7.1 ms [5.1–9.1] for GPT-2. This quantitative misalignment held even for the relatively small empirical effects observed for Direct Object/Sentential Complement, Agreement Violations, or High Attachment.

Language model surprisal does not accurately predict the variation across items. We next evaluated whether surprisal can account for the item-wise variation in our EOIs. Here, we assessed whether the models can predict

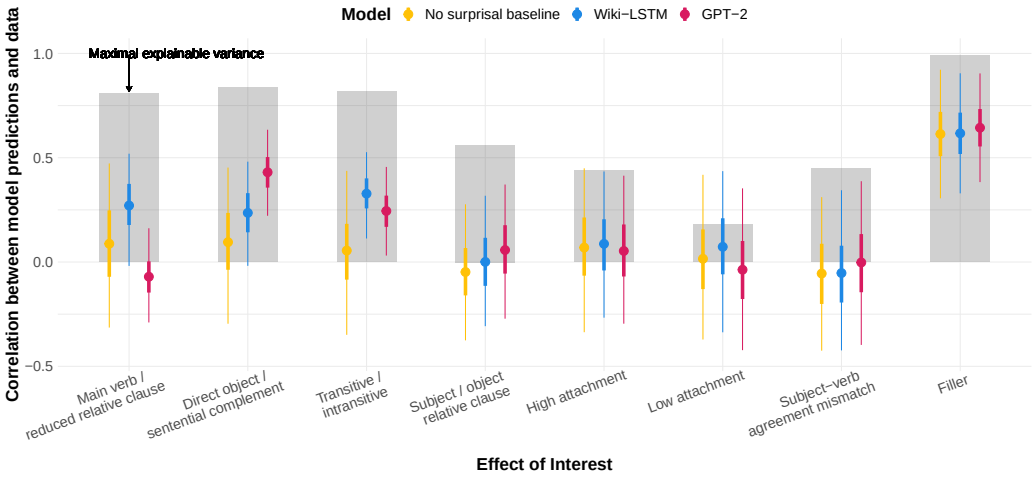

Figure 5: Correlation between the item-level predicted effects of interest and item-level empirical effects of interest. Error bars represent 95% credible intervals, and the grey bars indicate the explainable variance ceiling.

the correct rank ordering of items within each condition—in other words, whether they predicted higher processing difficulty for those items where humans showed longer reading times. For visual summaries of predicted vs. empirical EOIs by item, see Appendix C.

The results of this analysis are summarized in Figure 5, which plots the amount of item-wise variation in EOI captured by our models against the maximum amount of explainable variance (i.e. the Spearman-Brown-corrected split-half reliability for that construction). For filler items, the proportion of variance explained was relatively high, consistent with the reading time corpus findings reported by Schrimpf et al. (2021). By contrast, for the classic garden path subsets, the models accounted for less than half of the explainable variance, suggesting that much of the item-wise variation was not accounted for in these surprisal estimates. We note, however, that for these constructions, the surprisal-based models did generally explain more of the item-wise variance than the no-surprisal baseline.

For Object vs. Subject Relative Clause, High Attachment, and Agreement Violations, very little of the explainable variance was accounted for by the models. For Low Attachment, almost half of the explainable variance was predicted by GPT-2 model; we remind the reader, however, that this contrast had no reliable effect at the construction level to begin with.

5. Discussion

Prediction has been proposed as an organizing principle of human cognition in general and language in particular (Dell et al., 2021; Pickering and Garrod, 2013). In machine learning, deep-learning language models trained to predict upcoming words—or, more generally, some aspect of their input from another ("self-supervised learning")—have been immensely successful as a foundation for language technologies (Peters et al., 2018; Devlin et al., 2019), and have been shown to learn a surprising amount about language structure (Linzen and Baroni, 2021). This convergence between natural and artificial intelligence suggests the hypothesis that deep learning language models can be used as cognitive models of language processing (Goldstein et al., 2022; Schrimpf et al., 2021), with surprisal as linking function. We have evaluated this hypothesis with a large-scale self-paced reading dataset, the Syntactic Ambiguity Processing Benchmark; the scale of the dataset allowed us to evaluate the quantitative predictions of language model surprisal for individual sentences drawn from a set of targeted constructions of interest. If processing difficulty in these constructions arises from word-level unpredictability (Hale, 2001; Levy, 2008; Vani et al., 2021; Wilcox et al., 2021), we expect surprisal to track the magnitude of the effects observed in human reading (van Schijndel and Linzen, 2021).

Our results revealed three systematic misalignments between the predictions of the language models and human reading data. First, in a range of garden path constructions and ungrammatical sentences, language model underestimated the processing difficulty experienced by humans by orders of magnitude. Second, the models falsely predicted similar levels of processing difficulty among different garden path constructions even when they were very different empirically. Third, the models had only limited success in explaining item-wise variation in processing difficulty. For some of the constructions we tested, only slightly over half of the explainable variance across items was accounted for by surprisal; for others, language model surprisal did not capture any inter-item variation above and beyond a baseline model that did not include surprisal at all.

5.1. Implications for theories of sentence processing

At the broadest level, our results raise the question of how much of human sentence processing difficulty can ultimately be reduced to prediction. Our results cast doubt on the strong thesis that localized language processing

difficulty can be wholly reduced to word-by-word predictability (Levy, 2008), even for garden path constructions that are driven by syntactically unpredictable sentence completions, which would appear to be excellent candidates for a predictability-based account (Hale, 2001).

While we have established this conclusion only for the two specific language models we tested here, we believe that this conclusion would generalize to other large language models trained solely on a word prediction objective. First, both models failed in essentially similar ways on our contrasts, despite significant differences in architecture and training data. Second, both models directly optimize word-by-word perplexity over datasets that match or exceed the linguistic experience of a human's lifetime. Because they may be seen as directly optimizing the distribution over next-word predictions, they provide strong tests of hypothesis that human language processing rests on a fundamentally similar principle. Third, recent work has shown that even larger transformer models trained on even larger corpora—models that show excellent next-word prediction performance—nevertheless exhibit a worse fit to human reading times than less capable models such as the GPT-2 model we tested (Oh and Schuler, 2023; Shain et al., 2022), reversing an earlier trend observed with weaker models (Wilcox et al., 2020; Goodkind and Bicknell, 2018). This suggests that further improving the underlying language model's next-word-prediction accuracy is unlikely to improve its surprisal-based estimates of our effects of interest.

Our results do not license the stronger conclusion that prediction plays no role in language comprehension, of course: there is a wealth of converging evidence indicating that it does (Kutas et al., 2011). What they do suggest, instead, is either that the incremental predictions generated by humans diverge in substantial ways from the distributions encoded in models optimized to predict the next word, or that the role of predictability in moment-by-moment processing difficulty is more modest than often assumed, or both.

One approach, informed by the first hypothesis, involves creating language models whose predictions align more closely with those made by humans (Eisape et al., 2020). As a recent example, Arehalli et al. (2022) reweighted language models' predictability estimates to emphasize syntactic predictions more strongly than purely lexical ones, and found that doing so did bring model estimates of garden paths closer to the empirical effects. At the same time, the resulting estimates were still orders of magnitude smaller than observed effect sizes, suggesting that there are additional factors at play other than predictability.

An alternative approach, based on the second hypothesis, considers mechanisms that may operate in addition to or in tandem with prediction that may influence language processing difficulty. The language models we tested can, at least in principle, represent all possible analyses of the sentence in their hidden state (Aina and Linzen, 2021). This maps onto the fully parallel parsing assumption that tends to underlie "one-stage" models, such as standard formulation of surprisal theory (Hale, 2001; Levy, 2008) or the entropy reduction hypothesis (Hale, 2006). One interpretation of the massive cost of disambiguation we found is that the fully parallel parsing assumption is incorrect. Readers may not, in fact, consider most or all possible analyses of the sentence; instead, because of memory limitations on the number of concurrent interpretations of a sentence, when one of the grammatically possible interpretation is deemed unlikely, that interpretation drops out of consideration (Frazier, 1979; Gibson, 1991; Jurafsky, 1996). At the disambiguating region, when the favored interpretation is no longer consistent with the sentence, readers must construct the discarded interpretation based on their memory of the words they have read. Models like this are broadly referred to as "two-stage" models of sentence processing (Van Gompel and Pickering, 2007).

Two-stage models could make sense of the observation that garden paths take much more time to process than predicted by surprisal alone, and that at the level of individual sentence tokens, the model-derived surprisal correlates only modestly with the time it takes to resolve a garden path. The process of discarding unlikely parses may well be probabilistic, and the difference in reanalysis difficulty across constructions and items could be due to the different expectations generated by each construction and item (Jurafsky, 1996; Garnsey et al., 1997), or other structural or contextual factors not captured in our models (Frazier and Clifton, 1998; Sturt et al., 1999). It is our suspicion that integrating particle filters (Levy et al., 2008) or a limited beam width over symbolic parses (Hale et al., 2018) into neural language models will be important for rising to these challenges; we stress, however, that any such model would need to be supplemented with a mechanism for reconstructing a discarded parse.

Other aspects of our results are consistent with our conjecture that limited beam parsers are best suited to capture our results. We observed processing difficulty for 'high attachment' of relative clauses, but not low attachment. This pattern is the pattern predicted by 'two-stage' models of sentence processing when there is a bias to attach the relative clause to the most recent

noun (Frazier, 1979). If readers attach the relative clause to the most recent noun, they will not be garden pathed when that is the ultimately correct analysis. Hence there should be little measurable processing difficulty for low attachment, and that is what we observe. However, this observation contrasts with the results of a number of other studies, which have found that both low and high attachment of the relative cause processing difficulty when compared against their globally ambiguous baseline (Traxler et al., 1998; Van Gompel et al., 2005). This 'ambiguity advantage' pattern is a natural prediction of single-stage models (see Levy 2008 for details). Thus the findings from the relative clause attachment subset are broadly more consistent with two-stage models. But why should our results contrast with these previous reports? Previous self-paced reading work suggests that the ambiguity advantage pattern is modulated by the overall difficulty of the experimental context (Swets et al., 2008). In particular, it may be seen only when if the task context permits 'shallow' processing where comprehenders don't need to fully resolve the structure of the input (Swets et al., 2008; Logačev and Vasishth, 2016). If this is correct, then our finding of a high attachment penalty only may suggest that the participants in our experiment were engaged in 'deep' processing that pushed them to commit more strongly to a small number of analyses of the input. This raises the possibility that the misalignments we observe specifically characterize 'deeper' modes of language comprehension, a possibility that would need to be evaluated in further research.

Finally, some other surprisal-based theories that also explicitly take memory constraints into consideration have been proposed (Futrell et al., 2020; Levy, 2013). This framework assumes that contexts are often encoded or maintained imperfectly; therefore the conditional probability of an upcoming word might not always be based on the literal input presented so far. Similarly, a reader might rationally reconstruct what they think they have read, after seeing new evidence. Such an inference mechanism could potentially explain why the agreement mismatch costs in our study were much smaller than those induced by unlikely but grammatical continuation (i.e., garden path sentences), despite agreement mismatch being *outright ungrammatical*: The participants might have attributed the perceived ungrammaticality to their own memory error or to a production mistake. A full reanalysis of the literal ungrammatical sequence is therefore not always required. We note, however, that lossy-context models are unlikely to better explain our empirical data in the classic garden path subset, since, if anything, they should predict smaller garden path effects (e.g., seeing *remained* after *When the little girl*

attacked the lamb, the readers may sometimes infer/hallucinate a presence of a comma to rationalize the unpredicted *remained*, which will reduce processing difficulty). It remains to be seen whether our intuitive prediction is correct. Computationally implementable models of such (Hahn et al., 2022) can be evaluated against our benchmark dataset in the future.

5.2. Surprisal-based vs. embedding-based linking functions

The quantitative misalignments we have observed stand in contrast to recent studies in which measures derived from next-word-prediction models explained a substantial portion of the variance in human measurements, in particular neuroimaging data (Schrimpf et al., 2021; Goldstein et al., 2022; Caucheteux et al., 2023). The success of those analyses was taken to support a strong prediction-based account of language processing, of the sort that we have been arguing against. We see a number of overlapping explanations for this discrepancy; these explanations have to do with differences in materials and modeling approach between our study and the studies mentioned above.

The first difference between our study and the neuroimaging studies is in the linguistic materials: compared to the syntactically complex sentences included in the Syntactic Ambiguity Processing benchmark, other studies have tended to use simpler linguistic materials, perhaps more comparable to our fillers. As we have argued above, it is essential to evaluate models not only on sentences from a natural corpus, but also on theoretically critical constructions, whose frequency in a natural corpus may be low (Marvin and Linzen, 2018).

Second, our linking function was radically different. We used surprisal, a highly constrained, theoretically motivated linking function: each word is associated with a single scalar that represents that word's predictability. To fit the human data, we only needed to fit a handful of scalar "conversion factors", translating bits of surprisal to reading times. By contrast, in the neuroimaging studies mentioned above, an encoding model—typically, a dense linear layer—was trained to predict the human measurements from the language model's internal vector representations (embeddings). Such encoding models often have a vast number of parameters, and consequently may achieve a surprisingly good fit to human data even when trained to predict it from embeddings drawn from randomly initialized language models (Schrimpf et al., 2021) or systems trained to perform tasks that are not directly related to English next-word prediction, such as English to German translation (Antonello and Huth, 2023). The expressivity of these linking

functions makes it challenging to interpret the success of such analyses as providing support for prediction as the primary factor underlying human language processing, and motivates more theoretically constrained linking functions such as surprisal.

Third, our analysis was based on a generalization paradigm: if prediction is a unified mechanism that explains processing in both simple and complex sentences, we expect a linking function with parameters fit to simpler items to generalize to more complex ones. This is a higher bar for the models than the one used in previous studies, where the training and test set for the encoding model came from the same distribution: in those studies, the encoding model was in principle free to learn a separate processing mechanism for each construction, which leads to a much weaker support for prediction as a unified theory of sentence processing. Indeed, if our paradigm were flexible enough to fit a separate conversion factor for each construction, we would dramatically and trivially improve our model's fit to the human data (by construction, if not by item).

In summary, our approach differs along multiple dimensions from the approaches used in recent neuroimaging studies. The potential explanations we have discussed for the discrepancy between our results and the results of those studies can be disentangled in a neuroimaging study using our materials and following the generalization-based training/test split we have proposed.

5.3. The SAP Benchmark as a tool for theory evaluation

Stepping back from theoretical issues raised by the present data, the SAP Benchmark provides a framework that allows targeted testing of quantitatively explicit models of sentence processing. The dataset is large enough to provide relatively precise item-level estimates of effects for a range of widely studied processing effects in the sentence processing literature. These effects, such as garden path constructions or relative clause processing difficulty, have long been key phenomena that qualitative theories of sentence processing are expected to explain (Christianson et al., 2001; Traxler et al., 2002; Pearlmutter et al., 1999; Swets et al., 2008). The SAP Benchmark provides one way to leverage these important contrasts to quantitatively evaluate proposals about algorithmic-level claims (such as beam width of parser) or how to align theoretical models and psycholinguistic measures (relationship between neurophysiological measures and surprisal).

Moreover, having a single benchmark with multiple phenomena makes it possible to better evaluate the successes and failures of a range of different

theories (Oberauer et al., 2018). For example, surprisal fares quite well in some cases, but less well in others. The same likely could be said for other theories. But synthesizing these results to advance the debate is difficult given existing datasets. Advancing this state of the art requires the sort of higher precision, within-subject data provided by the SAP Benchmark.

6. Conclusion

This study tested a strong prediction-based linking hypothesis between deep learning word prediction models and human reading, namely that the surprisal of the word being read can be mapped linearly onto reading times. We found only modest support for this hypothesis, with two major misalignments between the predictions of the theory and human data. First, model-based surprisal systematically underpredicted the magnitude of garden path effects. Second, model-based surprisal showed only limited success at capturing variation across garden path effects in individual sentence tokens. Taken together, our results cast doubt on the strong hypothesis that word-by-word prediction difficulty predicted by deep learning models is sufficient to explain processing difficulty in syntactically complex contexts such as garden path constructions. Our work leaves open the possibility that these models could serve as one component of a cognitive model of syntactic processing, however, perhaps in conjunction with an additional syntactic reanalysis component (see Section 5.1).

More broadly than the specific theoretical questions we tested, our dataset clarifies the empirical picture in a range of syntactically complex English constructions. Against the backdrop of the so-called replication crisis in psychology (Open Science Collaboration, 2015), we were able to robustly replicate fundamental results from the psycholinguistic literature: English object-extracted relative clauses are harder to process than subject-extracted relative clauses (Grodner and Gibson, 2005); disambiguation in favor of an unexpected parse of a structurally ambiguous sentence causes processing difficulty (Frazier and Rayner, 1982); and subject-verb agreement errors are detected quickly and cause a slowdown in reading (Pearlmutter et al., 1999; Wagers et al., 2009).

We not only provided a high-powered replication of classic results, but also expanded the empirical picture by using an experimental design that allowed us to directly comparing reading times across constructions and items. We observed that Transitive/Intransitive garden path effect are about twice as

large as Direct Object/Sentential Complement ones, confirming the results of earlier studies (Sturt et al., 1999) with much more precise effect size estimates; we extended this observation by showing that Main Verb/Reduced Relative garden paths are the most difficult of all. We also saw that constructions differed in how much item-wise variation there was in the garden path effect. Quite aside of the debate around the limits of prediction from deep learning models as an explanatory factor for human language comprehension, then, it is our view that the empirical picture of the difficulty associated with the constructions and items in the SAP Benchmark should serve as a much more detailed, robust, replicable modeling target for any computational model of syntactic processing.

7. Acknowledgement

This work was supported by the National Science Foundation, grant nos. BCS-2020945 and BCS-2020914. We thank the audience of the 35th Conference of Human Sentence Processing, of the UPenn Linguistics Speaker Series, the UMass psycholinguistics community, Adrian Staub, and Shravan Vasishth for comments. This work was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise.

Appendix A. Comprehension accuracy by question types

Accuracy on the comprehension questions for the fillers was high (mean = 91.4%, min = 80%), indicating that participants were paying attention to the reading task. For our critical items, whenever possible, the comprehension questions were designed to specifically target the ambiguity resolution (e.g., given *The little girl fed the lamb remained relatively calm despite having asked for beef*, the comprehension question targeting ambiguity resolution was Did the girl feed the lamb?.) Table A.1 reports mean accuracy for each construction separately for questions that targeted ambiguity resolution and those that did not. As with the fillers, questions not targeting ambiguity resolution were answered with high accuracy across the board (82.2–96.4%). For questions targeting ambiguity resolution, accuracy varied across constructions. Accuracy was fairly high for Direct Object/Sentential Complement, Object vs. Subject Relative Clause, Agreement Violations, and High Attachment, ranging from 72.9% to 87.3%. For Transitive/Intransitive and Main Verb/Reduced Relative, by contrast, accuracy was extremely low when the sentences contained local

ambiguity (37% and 44.1% respectively). Even when the sentences were unambiguous, the accuracy for Transitive/Intransitive was still relatively low (62.7%). The very low accuracy associated with these two constructions was consistent with early findings (Christianson et al., 2001; Prasad and Linzen, 2021). Finally, accuracy for Low Attachment was only 55.2%; this was similar to the acceptability rate using grammatical judgement paradigm (Dillon et al., 2019).

	Question targeting ambiguity?	
Construction	No	Yes
MV/RR (ambiguous)	92.2% (1.3)	44.1% (0.7)
MV/RR (unambiguous)	96.4% (0.9)	77.8% (0.5)
NP/S (ambiguous)	94.5% (0.3)	78.7% (1.1)
NP/S (unambiguous)	92.9% (0.3)	87.3% (0.9)
NP/Z (ambiguous)	91.2% (0.4)	37.0% (1.0)
NP/Z (unambiguous)	92.2% (0.4)	62.7% (1.1)
Object RC	82.2% (0.6)	77.9% (0.7)
Subject RC	82.4% (0.7)	74.1% (0.7)
High Attachment	92.6% (0.4)	72.9% (0.7)
Low Attachment	92.4% (0.4)	55.2% (0.8)
Agreement (grammatical)	93.9% (0.3)	79.0% (1.2)
Agreement (ungrammatical)	93.5% (0.3)	77.1% (1.2)

Table A.1: Comprehension question accuracy for each experimental construction, for questions targeting ambiguity resolution and on questions not targeting ambiguity resolution. Standard errors (by-subject) in parentheses.

Appendix B. Can verb bias and plausibility explain item-wise variability in the garden-path subset?

Garnsey et al. (1997) showed that the Direct Object/Sentential Complement garden path effect is smaller for verbs that are more likely to take a sentential complement; when readers come across such verbs, they are more likely to predict the ultimately correct sentential complement parse than the direct object parse. They also showed that the strength of the garden path effect is affected by plausibility of the direct object reading of the ambiguous region, which again is hypothesized to affect the likelihood the readers adopt the direct object parse, which ends up being incorrect.

Inspired by the Garnsey et al. (1997) analysis, this appendix reports analyses that test to what degree item-wise variation in the garden path effect size can be explained using these two factors. We verb bias estimates from two sources—a Cloze task and the Corpus of Contemporary American English—as well as plausibility judgments we collected for the temporary but ultimately incorrect parse of each garden path sentence.

B.1. Predictors

Local phrase plausibility norms. In an online norming task ($N = 100$), we provided participants a fragment of each of the stimuli from the garden path subset of our self-paced reading experiment. The fragments continued through the second noun of the ambiguous sentences, forming a complete sentence. Examples of the fragment sentences corresponding to the three main garden path constructions are given in (2); cf. 1 for the complete sentences. Note that for Transitive/Intransitive we also removed the subordinating preposition (*after*, *when*, etc.).

(2) a. **Main Verb/Reduced Relative**: The little girl fed the lamb.

b. **Direct Object/Sentential Complement**: The little girl found the lamb.

c. **Transitive/Intransitive**: The little girl attacked the lamb.

These fragments correspond to the temporary syntactic analysis that later turns out to be incorrect in each of our target constructions. We refer to this as the **local phrase** for a given item.

Participants rated the plausibility of each sentence on a scale of 1 to 7, where 7 is plausible. We then defined a given item's local phrase plausibility as the arithmetic mean of each item's ratings.

Cloze-based verb bias estimates. We also conducted a Cloze-based norming task ($N = 332$). Participants were presented with the fragments of the experimental materials up until the disambiguating region, mixed with other materials irrelevant to the current project. Here the Transitive/Intransitive fragments did include the subordinating preposition.

(3) a. **Main Verb/Reduced Relative**: The little girl fed the lamb...

b. **Direct Object/Sentential Complement**: The little girl found the lamb...

c. **Transitive/Intransitive**: After the little girl attacked the lamb...

For each trial, participants were instructed to continue the fragment in whichever way they wished, with no time pressure. The responses were manually coded into either of the two possible parses or assigned a NA label. For instance, for the Transitive/Intransitive fragment in the example above, responses such as *After the little girl attacked the lamb* ***she*** were labeled as Transitive parses, while responses such as *After the little girl attacked the lamb* ***ran*** were labeled as Intransitive parses. The responses were labeled as NA when they did not contain enough information about the parse adopted by the participant (e.g., *After the little girl attacked the lamb* ***violently***).

We defined the verb bias as the proportion of responses that resulted in the target construction for a given item, out of all non-NA continuations. For example, an Direct Object/Sentential Complement item that resulted in a sentential complement continuation 70% of the time, after excluding the NA continuations, would receive a value of .70.

Corpus-based verb bias estimates. Finally, we performed a corpus analysis as described in *Material* section in the the main text, extracting from COCA sentences containing DP VERB DP for each of the verbs in question (e.g., DP *moved* DP). All of the results were parsed and labelled using the spaCy Python library. We then coded these parses into the three categories we used for coding the results of the Cloze task, and use those to compute verb bias, as before.

B.2. Hypotheses

Following Garnsey et al. (1997), we expected our two measures of verb bias to be inversely correlated with the size of a garden path effect for a given item. That is, the greater an item's verb bias towards the ultimately correct parse, the less surprising the critical disambiguating word should be. We also predicted that local phrase plausibility would positively correlate with garden path effects. The more plausible the local phrase, the more likely readers are to adopt or accept that parse, and hence the more processing difficulty we would expect at the disambiguating verb.

B.3. Analysis and results

The analyses we report in this Appendix are simpler linear regression models which did not consider spillover effects from the independent variables. We focused on the word immediately following the disambiguating word (the first spillover word), where garden path effects were largest. Table B.1 presents the correlation matrix of all the variables from simple regressions.

There were no significant correlations with local phrase plausibility, except for a non-significant trend in the expected direction for the Direct Object/Sentential Complement ambiguity.

Cloze-based verb bias strongly correlated with item-wise EOIs for Direct Object/Sentential Complement, replicating Garnsey et al. (1997). We did not find a similar correlation for the other two constructions. Corpus-based verb bias only showed a significant correlation with item-wise EOIs for the Main Verb/Reduced Relative ambiguity. This effect was in an unexpected direction: the more likely a reduced relative clause was for a given item, the *larger* its garden path effect was.

Table B.2 presents multiple regression results. The multiple regressions included all three predictors described in this section, as well as "surprisal difference", which we define as the surprisal of the critical verb in the unambiguous or grammatical sentence subtracted from the surprisal of the critical verb in the matching ambiguous or ungrammatical sentence. The results were consistent with those from the simple correlation tests: Only for the Direct Object/Sentential Complement ambiguity did cloze-based verb bias show a robust strong negative effect on the garden path effect size, and only for the Main Verb/Reduced Relative ambiguity effect did corpus-based verb bias show a robust strong positive effect. Overall, the Direct Object/Sentential Complement ambiguity is the only type of garden paths for which both word surprisal and syntactic surprisal adequately tracked item-level effects.

Due to the unexpected direction of the effect of corpus-based verb bias for the Main Verb/Reduced Relative subset, caution should be exercised in interpreting the regression results in this subset: When corpus-based verb bias was added to the regression model, it yielded few more spurious significant effects (e.g., an unexpected *negative* plausibility effect), but at the same time the adjusted R-squared, while significantly different from zero, is fairly low.

The significant effect of corpus-based verb bias in the unexpected direction appears to have been driven by two data points. For both items, the critical ambiguous verb was *fed*, which has a relatively high verb bias in the COCA

counts. A closer examination revealed that the reduced relative clause uses of *fed* were almost exclusively drawn from academic texts. As such, this bias might not be representative of the average reader's experience with this verb.

In conjunction with the results of the surprisal analysis from Section 4.4, we conclude item-wise variation in the magnitude of garden path effects, while substantial and reliable, cannot be readily explainable by word surprisal, syntactic surprisal (cloze-based and corpus-based), or local phrase plausibility, at least not with a simple linear linking function.

Main Verb/Reduced Relative

Variables	EOI size	LSTM	GPT-2	Plausibility	Cloze	COCA
EOI size	–	0.38	-0.08	-0.03	-0.10	0.47*
LSTM		–	0.32	0.29	-0.02	0.10
GPT-2			–	0.33	-0.14	-0.57**
Plausibility				–	0.05	0.24
RRC bias (Cloze)					–	0.05
RRC bias (COCA)						–

Direct Object/Sentential Complement

	EOI size	LSTM	GPT-2	Plausibility	Cloze	COCA
EOI size	–	0.60**	0.57**	0.24	-0.69***	-0.32
LSTM		–	0.40*	-0.15	-0.28	-0.43*
GPT-2			–	0.23	-0.55**	-0.14
Plausibility				–	-0.47*	-0.05
Sent bias (Cloze)					–	0.22
Sent bias (COCA)						–

Transitive/Intransitive

Variables	EOI size	LSTM	GPT-2	Plausibility	Cloze	COCA
EOI size	–	0.29	0.36	-0.01	-0.26	-0.18
LSTM		–	0.69***	0.02	-0.12	0.13
GPT-2			–	-0.24	-0.02	-0.23
Plausibility				–	-0.27	0.22
Intrans bias (Cloze)					–	0.43*
Intrans bias (COCA)						–

Table B.1: Correlation table for correlations between EOI size and the potentially relevant variables and correlations between the variables. LSTM and GPT-2: surprisal differences at the disambiguating verb, as estimated from the language models; Plaus: local phrase plausibility; Cloze: verb subcategorization bias as normed by the cloze task; COCA: verb subcategorization bias as estimated from COCA corpus; RRC: reduced relative clause; Sent: sentential complement; Intrans: intransitive. * $p<.05$; ** $p<.01$; *** $p<.001$

Main Verb/Reduced Relative

LSTM	GPT-2	Plausibility	Cloze	COCA	Adjusted R^2
1.94	–	-0.60	-0.39	–	0.05
–	-0.10	-0.17	-0.43	–	-0.15
2.05	–	-1.29	-0.56	2.51*	0.26
–	2.42*	-2.15*	-0.20	3.65**	0.32*

Direct Object/Sentential Complement

LSTM	GPT-2	Plausibility	Cloze	COCA	Adjusted R^2
2.68*	–	0.24	-3.05**	–	0.57***
–	1.25	-0.58	-2.81*	–	0.44**
2.28*	–	0.22	-2.96**	-0.11	0.54**
–	1.21	-0.50	-2.58*	-0.11	0.45**

Transitive/Intransitive

LSTM	GPT-2	Plausibility	Cloze	COCA	Adjusted R^2
0.13	–	0.07	-1.07	–	-0.09
–	0.90	0.25	-1.04	–	-0.04
0.21	–	0.38	-0.48	-0.75	-0.11
–	0.73	0.45	-0.56	-0.55	-0.08

Table B.2: T-values from multiple regressions predicting EOI sizes with different potentially relevant variables. Each row corresponds to a separate regression analysis, in which the variables marked with a dash were left out. All predictors were centered. LSTM, GPT-2: surprisal difference at the disambiguating verb, as estimated from each of the language models; Plausibility: local phrase plausibility; Cloze: verb subcategorization bias as normed by the cloze task; COCA: verb subcategorization bias as estimated from the COCA corpus. * $p<.05$; ** $p<.01$; *** $p<.001$

Appendix C. Item-level correlation plots

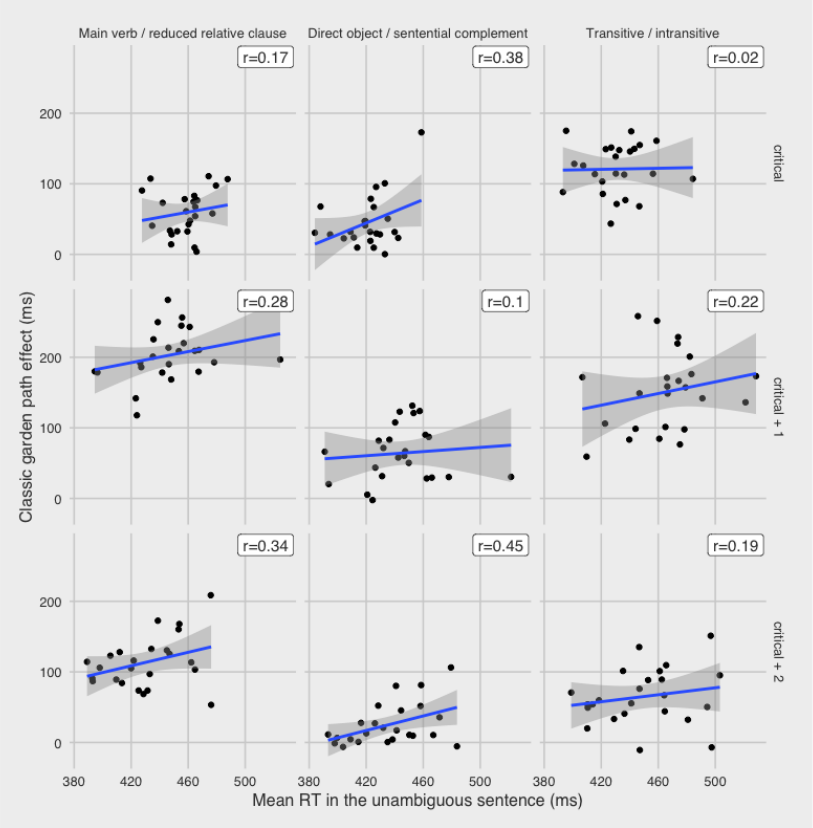

Figure C.1: Scatterplot of empirical garden path effect sizes and the raw RTs on the corresponding target word in the unambiguous sentences. Here, the EOI reflects the excess processing cost on the critical word in the temporarily ambiguous sentence compared to the unambiguous one.

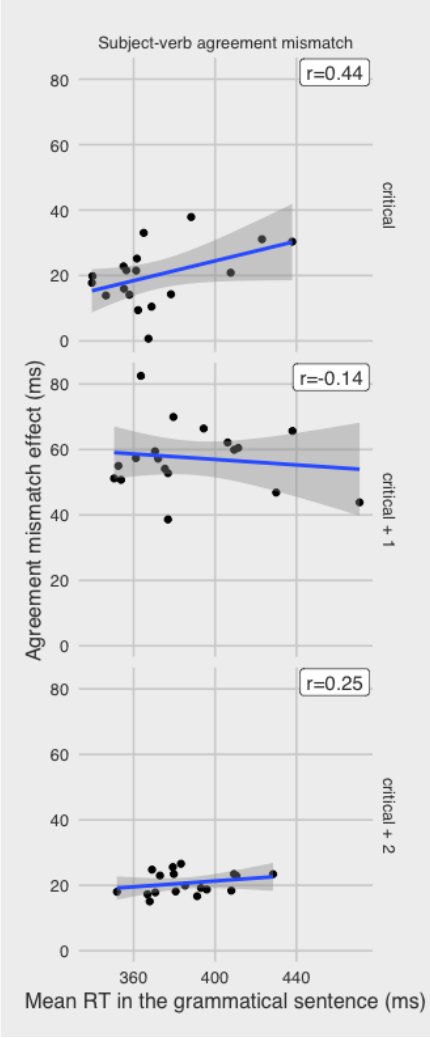

Figure C.2: Scatterplot of empirical object relative clause effect sizes and the raw RTs on the corresponding target word in the subject relative clause sentences.

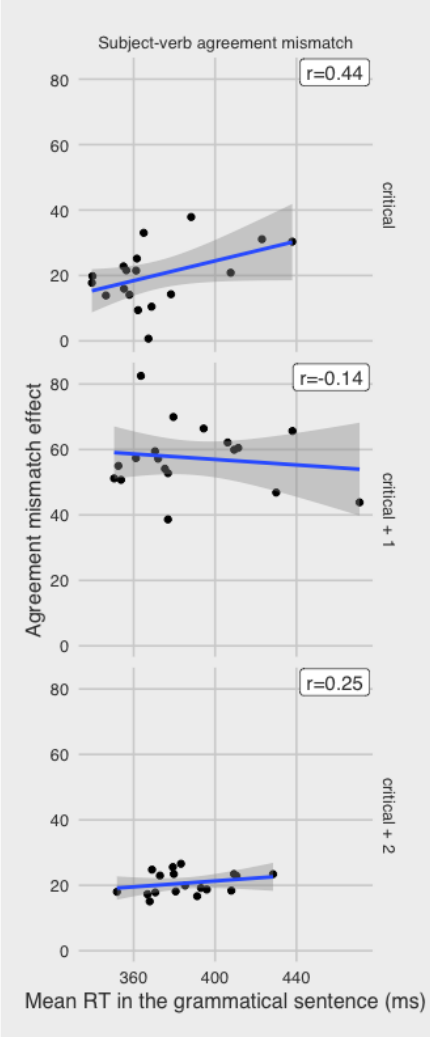

Figure C.3: Scatterplot of empirical agreement mismatch effect sizes and the raw RTs on the corresponding target word in the grammatical sentences.

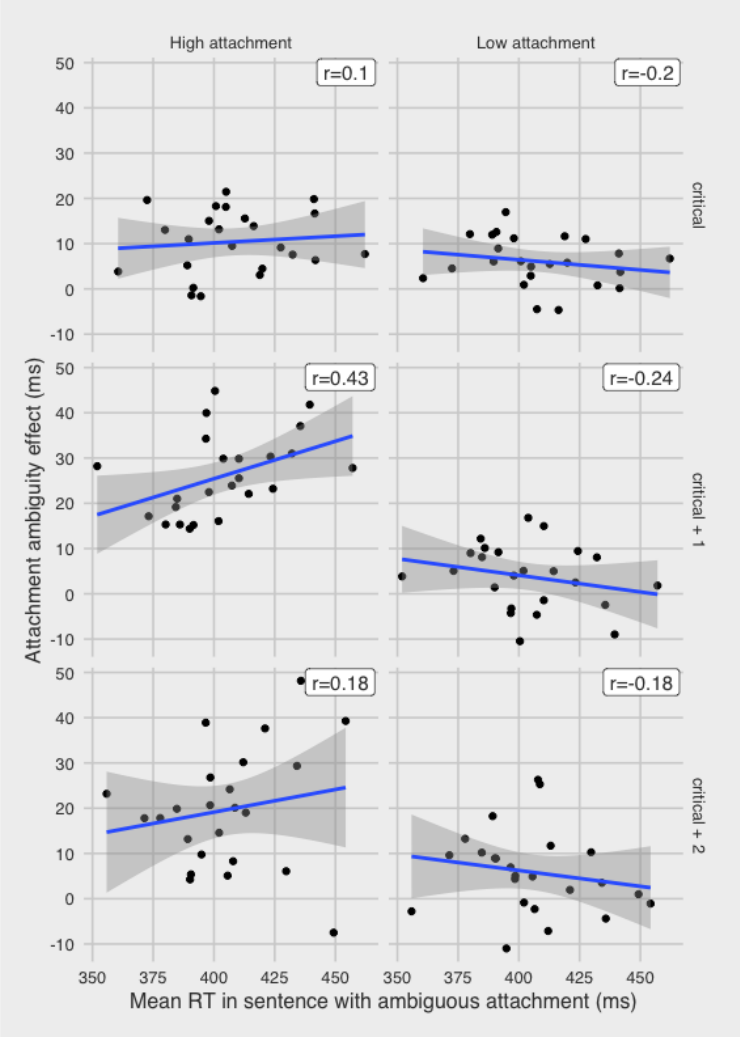

Figure C.4: Scatterplot of empirical attachment effect sizes and the raw RTs on the corresponding target word in the multi-attachment sentences.

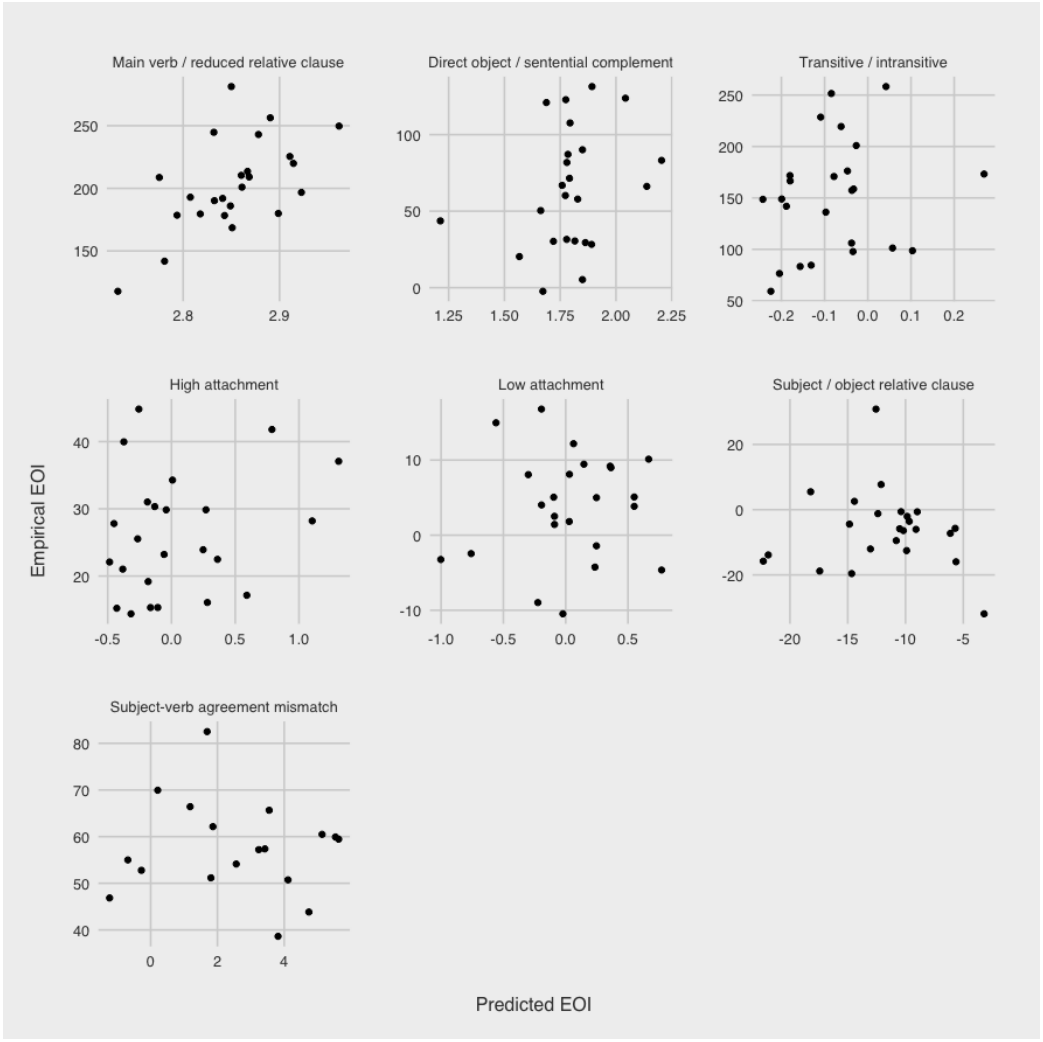

Figure C.5: Scatterplot of item-level empirical EOI sizes and predicted EOI sizes (No surprisal baseline model).

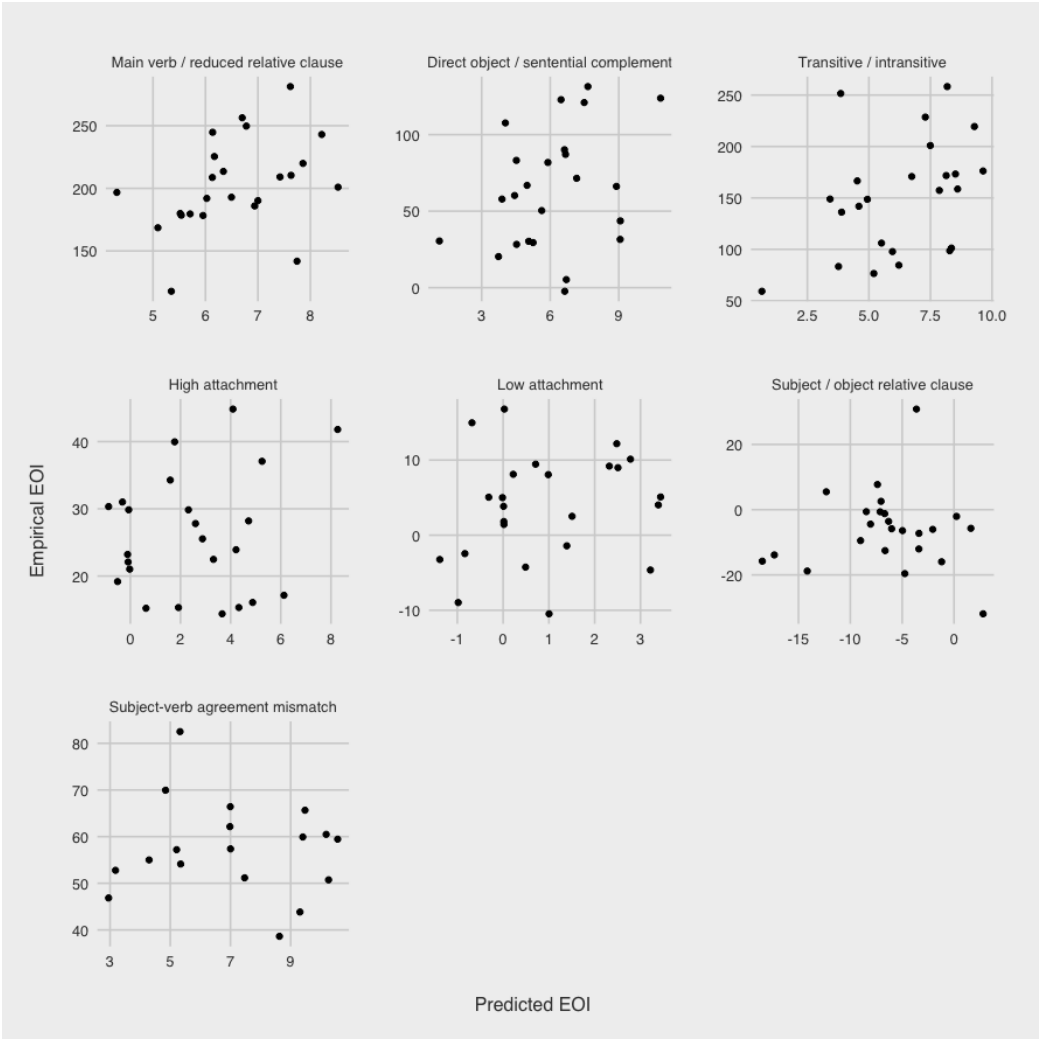

Figure C.6: Scatterplot of item-level empirical EOI sizes and predicted EOI sizes (Wikitext LSTM+).

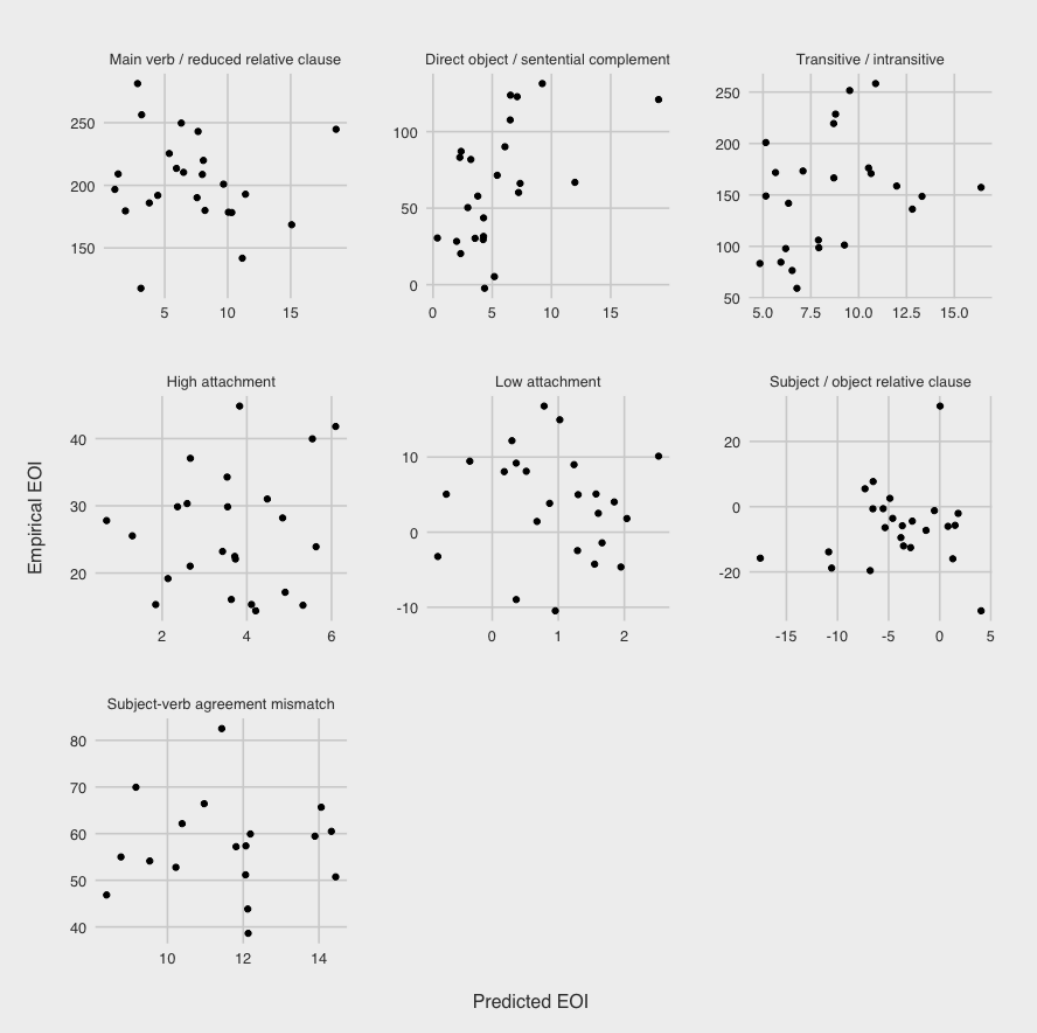

Figure C.7: Scatterplot of item-level empirical EOI sizes and predicted EOI sizes (GPT-2+).

Appendix D. Materials

This section lists the materials for all four subsets.

D.1. Classic garden path

We spell out all six versions of the first item, and use a more concise notation for the remaining items.

(1) a. Direct Object/Sentential Complement:
   i. The suspect showed that the file **deserved further investigation** during the murder trial.
   ii. The suspect showed the file **deserved further investigation** during the murder trial.

b. Transitive/Intransitive:
   i. Because the suspect changed, the file **deserved further investigation** during the jury discussions.
   ii. Because the suspect changed the file **deserved further investigation** during the jury discussions.

c. Main Verb/Reduced Relative:
   i. The suspect who was sent the file **deserved further investigation** given the new evidence.
   ii. The suspect sent the file **deserved further investigation** given the new evidence.

(2) a. The woman maintained (that) the mail disappeared mysteriously from her front porch.

b. After the woman moved(,) the mail disappeared mysteriously from the delivery system.

c. The woman (who was) brought the mail disappeared mysteriously after reading the bad news in it.

(3) a. The boy found (that) the chicken stayed surprisingly happy in the new barn.

b. Although the boy attacked(,) the chicken stayed surprisingly happy as if nothing happened.

c. The boy (who was) fed the chicken stayed surprisingly happy despite having a mild allergic reaction.

(4) a. The new doctor demonstrated (that) the operation appeared increasingly likely to succeed.

b. After the new doctor left(,) the operation appeared increasingly likely to succeed.

c. The new doctor (who was) offered the operation appeared increasingly likely to succeed in her career.

(5) a. The professor noticed (that) the grant gained more attention from marine biologists.

b. After the professor read(,) the grant gained more attention due to her excellent description.

c. The professor (who was) awarded the grant gained more attention from marine biologists.

(6) a. The technician reported (that) the service stopped working almost immediately after the storm started.

b. After the technician called(,) the service stopped working almost immediately to his surprise.

c. The technician (who was) refused the service stopped working almost immediately after the argument.

(7) a. The mechanic observed (that) the truck needed several more hours to be repaired.

b. Because the mechanic stopped(,) the truck needed several more hours before it could be fully repaired.

c. The mechanic (who was) brought the truck needed several more hours to fully repair it.

(8) a. The guitarist knew (that) the song failed dramatically because of the tensions within the band.

b. After the guitarist began(,) the song failed dramatically because he skipped the sound check.

c. The guitarist (who was) assigned the song failed dramatically because he never practiced enough.

(9) a. The player revealed (that) the bonus remained essentially the same as in the original contract.

b. Although the player lost(,) the bonus remained essentially the same as in the original contract.

c. The player (who was) paid the bonus remained essentially the same despite his sudden fame and wealth.

(10) a. The recent hire claimed (that) the job prepared many students for careers in media.

b. Once the recent hire started(,) the job prepared many students for careers in media.

c. The recent hire (who was) offered the job prepared many students for careers in media.

(11) a. The assistant manager discovered (that) the training seemed unnecessarily demanding for new staff.

b. While the assistant manager worked(,) the training seemed unnecessarily demanding to him.

c. The assistant manager (who was) assigned the training seemed unnecessarily demanding to new staff.

(12) a. The mayor showed (that) the document provided sufficient evidence to prove her innocence.

b. Although the mayor changed(,) the document provided sufficient evidence for what he had promised.

c. The mayor (who was) sent the document provided sufficient evidence that it was simply blackmail.

(13) a. The basketball player mentioned (that) the contract created another controversy in the NBA.

b. After the basketball player signed(,) the contract created another political controversy in the NBA.

c. The basketball player (who was) handed the contract created another controversy in the NBA.

(14) a. The engineer maintained (that) the equipment required constant supervision from senior technicians.

b. After the engineer moved(,) the equipment required constant supervision from senior technicians.

c. The engineer (who was) brought the equipment required constant supervision from senior technicians.

(15) a. The little girl found (that) the lamb remained relatively calm despite the absence of its mother.

b. When the little girl attacked(,) the lamb remained relatively calm despite the sudden assault.

c. The little girl (who was) fed the lamb remained relatively calm despite having asked for beef.

(16) a. The yoga instructor demonstrated (that) the position demanded immense physical effort from everyone.

b. Before the yoga instructor left(,) the position demanded immense physical effort from everyone.

c. The yoga instructor (who was) offered the position demanded immense physical effort from everyone.

(17) a. The governor noticed (that) the contract received sweeping support across the entire state.

b. While the governor read(,) the contract received sweeping support from the audience at the rally.

c. The governor (who was) awarded the contract received sweeping support across the entire state.

(18) a. The patient reported (that) the treatment continued causing uncomfortable side effects like nausea.

b. Before the patient called(,) the treatment continued causing uncomfortable side effects like nausea.

c. The patient (who was) refused the treatment continued causing uncomfortable scenes in the ER.

(19) a. The operator observed (that) the machine started working efficiently all of a sudden.

b. Once the operator stopped(,) the machine started working efficiently without any supervision.

c. The operator (who was) brought the machine started working efficiently with the added automation.

(20) a. The dancer knew (that) the ballet achieved incredible success for a small local production.
b. Once the dancer began(,) the ballet achieved incredible success for a show with a new performer.
c. The dancer (who was) assigned the ballet achieved incredible success for a new performer.

(21) a. The contestant revealed (that) the money became unavailable to him when the show's budget shrank.
b. After the contestant lost(,) the money became unavailable despite his previous three wins in a row.
c. The contestant (who was) paid the money became unavailable and suddenly terminated his contract.

(22) a. The new chef claimed (that) the restaurant separated mediocre cooks from gifted ones.
b. Once the new chef started(,) the restaurant separated mediocre cooks from gifted ones.
c. The new chef (who was) offered the restaurant separated mediocre cooks from gifted ones.

(23) a. The apprentice baker discovered (that) the oven produced smaller cakes because it heated too fast.
b. When the apprentice baker worked(,) the oven produced smaller cakes because he lacked experience.
c. The apprentice baker (who was) assigned the oven produced smaller cakes because he lacked experience.

D.2. Agreement violations

(24) a. If the supervisor changes, the schedules deserves further inspection by the rest of the staff.
b. If the supervisor changes, the schedule deserves further inspection by the rest of the staff.

(25) a. When the magician moves, the cards disappears mysteriously from his assistantś hand.
b. When the magician moves, the card disappears mysteriously from his assistantś hand.

(26) a. Whenever the lawyer leaves, his clients appears increasingly uncomfortable in the courtroom.

b. Whenever the lawyer leaves, his client appears increasingly uncomfortable in the courtroom.

(27) a. After the esteemed reviewer reads, the books gains more attention due to his glowing praise.

b. After the esteemed reviewer reads, the book gains more attention due to his glowing praise.

(28) a. Whenever the nurse calls, the doctors stops working immediately to check on the patient.

b. Whenever the nurse calls, the doctor stops working immediately to check on the patient.

(29) a. When the lecturer stops, her audiences needs several minutes to reflect on the content.

b. When the lecturer stops, her audience needs several minutes to reflect on the content.

(30) a. When the actress begins, the scenes fails dramatically despite the months she spent rehearsing.

b. When the actress begins, the scene fails dramatically despite the months she spent rehearsing.

(31) a. After the worst team loses, the tournaments remains essentially the same for the rest of the year.

b. After the worst team loses, the tournament remains essentially the same for the rest of the year.

(32) a. When the supervisor works, the shifts seems unnecessarily stressful on a Friday night.

b. When the supervisor works, the shift seems unnecessarily stressful on a Friday night.

(33) a. After the diplomat signs, the agreements creates another border conflict as a side effect.

b. After the diplomat signs, the agreement creates another border conflict as a side effect.

(34) a. Whenever the reporter moves, the cameras requires constant adjustment from the director.

b. Whenever the reporter moves, the camera requires constant adjustment from the director.

(35) a. Unless the dog attacks, the cats remains relatively tranquil throughout the day.

b. Unless the dog attacks, the cat remains relatively tranquil throughout the day.

(36) a. Until the lead architect leaves, the projects demands immense patience from the engineers.

b. Until the lead architect leaves, the project demands immense patience from the engineers.

(37) a. Even if the mother calls, her boys continues causing problems with the other kids on the playground.

b. Even if the mother calls, her boy continues causing problems with the other kids on the playground.

(38) a. After the tutor stops, the students starts working independently on the questions.

b. After the tutor stops, the student starts working independently on the questions.

(39) a. Once the head surgeon begins, the operations achieves incredible results given the risks involved.

b. Once the head surgeon begins, the operation achieves incredible results given the risks involved.

(40) a. After the producer starts, the auditions separates mediocre actors from talented ones.

b. After the producer starts, the audition separates mediocre actors from talented ones.

(41) a. However hard the scientist works, his experiments produces smaller amounts of alcohol than expected.

b. However hard the scientist works, his experiment produces smaller amounts of alcohol than expected.

D.3. Relative clauses

(42) a. The bus driver who followed the kids wondered about the location of a hotel.

b. The bus driver who the kids followed wondered about the location of a hotel.

(43) a. The chef who distracted the cameraman poured the flour onto the counter.

b. The chef who the cameraman distracted poured the flour onto the counter.

(44) a. The children who woke the father bothered him about the trip to the beach.

b. The children who the father woke bothered him about the trip to the beach.

(45) a. The class that disliked the teacher skimmed the reading for the week.

b. The class that the teacher disliked skimmed the reading for the week.

(46) a. The dancer that loved the audience ignored some basic principles.

b. The dancer that the audience loved ignored some basic principles.

(47) a. The employees that noticed the fireman hurried across the open field.

b. The employees that the fireman noticed hurried across the open field.

(48) a. The farmer that approached the customers lifted the chickens from their coop.

b. The farmer that the customers approached lifted the chickens from their coop.

(49) a. The farmer who hired the rancher piled the seeds in long rows.

b. The farmer who the rancher hired piled the seeds in long rows.

(50) a. The firemen that called the residents attacked the house with high-powered hoses.

b. The firemen that the residents called attacked the house with high-powered hoses.

(51) a. The girl who watched the parents changed a critical part of the story.

b. The girl who the parents watched changed a critical part of the story.

(52) a. The investigator who phoned the agency considered Ms. Reynolds from accounting.

b. The investigator who the agency phoned considered Ms. Reynolds from accounting.

(53) a. The judge who addressed the witnesses noticed the defense attorneys.

b. The judge who the witnesses addressed noticed the defense attorneys.

(54) a. The manager who visited the boss remembered some inconvenient facts.

b. The manager who the boss visited remembered some inconvenient facts.

(55) a. The mathematician who visited the chairman created a solution to the well-known problem.

b. The mathematician who the chairman visited created a solution to the well-known problem.

(56) a. The monkeys that watched the zookeepers charged the bars of their cage.

b. The monkeys that the zookeepers watched charged the bars of their cage.

(57) a. The movie star who visited the organizers proposed an annual prize.

b. The movie star who the organizers visited proposed an annual prize.

(58) a. The neighbor who observed the couple purchased the old Victorian house.

b. The neighbor who the couple observed purchased the old Victorian house.

(59) a. The pilot who delayed the ground crew remained on the runway for a long time.

b. The pilot who the ground crew delayed remained on the runway for a long time.

(60) a. The soldiers that helped the natives climbed the big rock that blocked the path.

b. The soldiers that the natives helped climbed the big rock that blocked the path.

(61) a. The speaker who entertained the economists predicted a good year for the industry.

b. The speaker who the economists entertained predicted a good year for the industry.

(62) a. The table top that rested on the box screwed directly to the legs.

b. The table top that the box rested on screwed directly to the legs.

(63) a. The trainer who called the jockey rubbed the horseś skin.

b. The trainer who the jockey called rubbed the horseś skin.

(64) a. The veteran who admired the coach defeated his greatest rival.

b. The veteran who the coach admired defeated his greatest rival.

(65) a. The visitor who introduced the student walked across the quad.

b. The visitor who the student introduced walked across the quad.

D.4. Attachment ambiguities

(66) a. In the lobby, Clyde bumped into the chauffeur of the CEO who is reckless and very unpopular with the company.

b. In the lobby, Clyde bumped into the chauffeur of the CEOs who is reckless and very unpopular with the company.

c. In the lobby, Clyde bumped into the chauffeurs of the CEO who is reckless and very unpopular with the company.

(67) a. Edwin has been reading about the sister of the actor who was visiting the resort in Death Valley.

b. Edwin has been reading about the sister of the actors who was visiting the resort in Death Valley.

c. Edwin has been reading about the sisters of the actor who was visiting the resort in Death Valley.

(68) a. From the gallery, Franny observed the nurse of the surgeon who was in charge of the operation currently underway.

b. From the gallery, Franny observed the nurse of the surgeons who was in charge of the operation currently underway.

c. From the gallery, Franny observed the nurses of the surgeon who was in charge of the operation currently underway.

(69) a. Gerald introduced himself to the niece of the billionaire who sails vintage yachts around the Vineyard.

b. Gerald introduced himself to the niece of the billionaires who sails vintage yachts around the Vineyard.

c. Gerald introduced himself to the nieces of the billionaire who sails vintage yachts around the Vineyard.

(70) a. At the potluck, Marcus chatted with the aunt of the nun who bakes sugar cookies with cute designs.

b. At the potluck, Marcus chatted with the aunt of the nuns who bakes sugar cookies with cute designs.

c. At the potluck, Marcus chatted with the aunts of the nun who bakes sugar cookies with cute designs.

(71) a. During the budget negotiation, Janet charmed the assistant of the executive who decides almost everything in secret.

b. During the budget negotiation, Janet charmed the assistant of the executives who decides almost everything in secret.

c. During the budget negotiation, Janet charmed the assistants of the executive who decides almost everything in secret.

(72) a. On the fishing trip, we laughed at the uncle of the sailor who was confused about the motor on the boat.

b. On the fishing trip, we laughed at the uncle of the sailors who was confused about the motor on the boat.

c. On the fishing trip, we laughed at the uncles of the sailor who was confused about the motor on the boat.

(73) a. At trial, we scrutinized the prisoner of the FBI agent who was lying about the incident at the casino.

b. At trial, we scrutinized the prisoner of the FBI agents who was lying about the incident at the casino.

c. At trial, we scrutinized the prisoners of the FBI agent who was lying about the incident at the casino.

(74) a. During the demonstration, someone photographed the soldier of the lieutenant who was camouflaged and hiding in the trees.

b. During the demonstration, someone photographed the soldier of the lieutenants who was camouflaged and hiding in the trees.

c. During the demonstration, someone photographed the soldiers of the lieutenant who was camouflaged and hiding in the trees.

(75) a. Karl recognized the hostage of the pirate who was on TV this morning on the local news.

b. Karl recognized the hostage of the pirates who was on TV this morning on the local news.

c. Karl recognized the hostages of the pirate who was on TV this morning on the local news.

(76) a. During the play, we all heckled the murderer of the prince who was disguised as a peasant from nearby Trosselheim.

b. During the play, we all heckled the murderer of the princes who was disguised as a peasant from nearby Trosselheim.

c. During the play, we all heckled the murderers of the prince who was disguised as a peasant from nearby Trosselheim.

(77) a. At the charity show, Noreen nodded to the sidekick of the actor who was juggling sharp knives and glass bottles.

b. At the charity show, Noreen nodded to the sidekick of the actors who was juggling sharp knives and glass bottles.

c. At the charity show, Noreen nodded to the sidekicks of the actor who was juggling sharp knives and glass bottles.

(78) a. No one quite knew how to respond to the buddies of the janitors who burp without excusing themselves.

b. No one quite knew how to respond to the buddies of the janitor who burp without excusing themselves.

c. No one quite knew how to respond to the buddy of the janitors who burp without excusing themselves.

(79) a. The cunning Wally outmaneuvered the henchmen of the villains who often fail to carry out the plot.

b. The cunning Wally outmaneuvered the henchmen of the villain who often fail to carry out the plot.

c. The cunning Wally outmaneuvered the henchman of the villains who often fail to carry out the plot.

(80) a. Down at the pub, Ollie gossiped about the daughters of the nurses who were at church last Sunday in grimy shorts.

b. Down at the pub, Ollie gossiped about the daughters of the nurse who were at church last Sunday in grimy shorts.

c. Down at the pub, Ollie gossiped about the daughter of the nurses who were at church last Sunday in grimy shorts.

(81) a. From the lounge everyone could see the pilots of the millionaires who were distrusted by everyone at the company.

b. From the lounge everyone could see the pilots of the millionaire who were distrusted by everyone at the company.

c. From the lounge everyone could see the pilot of the millionaires who were distrusted by everyone at the company.

(82) a. On the news they showed the accomplices of the thieves who were indicted for stealing the Mona Lisa.

b. On the news they showed the accomplices of the thief who were indicted for stealing the Mona Lisa.

c. On the news they showed the accomplice of the thieves who were indicted for stealing the Mona Lisa.

(83) a. Everyone at the party groaned at the bodyguards of the divas who smoke clove cigarettes constantly.

b. Everyone at the party groaned at the bodyguards of the diva who smoke clove cigarettes constantly.

c. Everyone at the party groaned at the bodyguard of the divas who smoke clove cigarettes constantly.

(84) a. At the summit, Ursula warmly greeted the advisors of the tycoons who snowboard in Aspen in January.

b. At the summit, Ursula warmly greeted the advisors of the tycoon who snowboard in Aspen in January.

c. At the summit, Ursula warmly greeted the advisor of the tycoons who snowboard in Aspen in January.

(85) a. Rosalina testified against the detectives of the senators who were caught spying on his colleagues.

b. Rosalina testified against the detectives of the senator who were caught spying on his colleagues.

c. Rosalina testified against the detective of the senators who were caught spying on his colleagues.

(86) a. Before the exhibition, Silas telephoned the friends of the bodybuilders who write fan fiction about Batman.

b. Before the exhibition, Silas telephoned the friends of the bodybuilder who write fan fiction about Batman.

c. Before the exhibition, Silas telephoned the friend of the bodybuilders who write fan fiction about Batman.

(87) a. At her orientation, Tamara recently met the nephews of the professors who paint beautiful portraits of local celebrities.

b. At her orientation, Tamara recently met the nephews of the professor who paint beautiful portraits of local celebrities.

c. At her orientation, Tamara recently met the nephew of the professors who paint beautiful portraits of local celebrities.

(88) a. Everyone at the coffee shop sympathized with the couriers of the florists who were complaining about the weather.

b. Everyone at the coffee shop sympathized with the couriers of the florist who were complaining about the weather.

c. Everyone at the coffee shop sympathized with the courier of the florists who were complaining about the weather.

(89) a. Despite the good press, we didnt́ really like the commanders of the soldiers who whistle very loudly and for no reason at all.

b. Despite the good press, we didnt́ really like the commanders of the soldier who whistle very loudly and for no reason at all.

c. Despite the good press, we didnt́ really like the commander of the soldiers who whistle very loudly and for no reason at all.

D.5. Fillers

(90) There are now rumblings that Apple might soon invade the smart watch space, though the company is maintaining its customary silence.

(91) A bill was drafted and introduced into Parliament several times but met with great opposition, mostly from farmers.

(92) The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity.

(93) Seeing Peter slowly advancing upon him through the air with dagger poised, he sprang upon the bulwarks to cast himself into the sea.

(94) Some months later, Michael Larson saw another opportunity to stack the odds in his favor with a dash of ingenuity.

(95) Bob Murphy, the Senior PGA Tour money leader with seven hundred thousand, says heat shouldnt́ be a factor.

(96) Greg Anderson, considered a key witness by the prosecution, vowed he wouldn't testify when served a subpoena last week.

(97) Owls are more flexible than humans because a birdś head is only connected by one socket pivot.

(98) Even in the same animal, not all bites are the same.

(99) Buck did not like it, but he bore up well to the work, taking pride in it.

(100) These days, neuroscience is beginning to catch up to musicians who practice mentally.

(101) Hybrid vehicles have a halo that makes owners feel righteous and their neighbors feel guilty for not doing as much to save the planet.

(102) Binge drinking may not necessarily kill or even damage brain cells, as commonly thought, a new animal study suggests.

(103) When attacked, a skunk's natural inclination is to turn around, lick its tail and spray a noxious scent.

(104) All that the brain has to work with are imperfect incoming electrical impulses announcing that things are happening.

(105) There often seems to be more diving in soccer than in the Summer Olympics.

(106) Susan B. Anthony spent nearly sixty years of her life devoted to the cause of social justice and equality for all.

(107) Unfortunately, for every six water bottles we use, only one makes it to the recycling bin.

(108) As in the United States, Colombian legislation requires travelers entering the country to declare cash in excess of ten thousand dollars.

(109) Stress is a risk factor for both depression and anxiety, he says.

(110) When it comes to having a lasting and fulfilling relationship, common wisdom says that feeling close to your romantic partner is paramount.

(111) Voltaire himself probably won around half a million livres, a large fortune, which he then made even larger.

(112) When preparing to check out of their hotel room, some frequent travelers pile up their used bath towels on the bathroom floor.

(113) Research showing that a tiny European river bug called the water boatman may be the loudest animal on earth.

(114) When the new world was first discovered it was found to be, like the old, well stocked with plants and animals.

(115) Police in Georgia have shut down a lemonade stand run by three girls trying to save up for a trip to a water park.

(116) An early task will be to make sure the newfound microbes were not introduced while drilling through the ice into the lake.

(117) Lady Gaga's YouTube account was suspended Thursday.

(118) John Thornton asked little of man or nature.

(119) Proper ventilation will make a backdraft less likely.

(120) For centuries, time was measured by the position of the sun with the use of sundials.

(121) The girl's feet were then re-wrapped even tighter than before, causing her footprint to shrink further.

(122) The astronauts used a hefty robotic arm to move the bus-size canister, stuffed with nearly three tons of packing foam.

(123) Very similar, but even more striking, is the evidence from athletic training.

(124) It was a forbidding challenge, and it says much for Winstanley's persuasive abilities, not to mention his self-confidence.

(125) With schools still closed, cars still buried and streets still blocked by the widespread weekend snowstorm, officials are asking people to help out.

(126) Steam sterilization is limited in the types of medical waste it can treat, but is appropriate for laboratory substances contaminated with infectious organisms.

(127) From coal to cars, Chinese floods tangle supply chains worldwide.

(128) This new film marks 10 years since the death of the superstar.

References

Aina, L. and Linzen, T. (2021). The language model understood the prompt was ambiguous: Probing syntactic uncertainty through generation. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 42–57, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Antonello, R. and Huth, A. (2023). Predictive coding or just feature discovery? an alternative account of why language models fit brain data. *Neurobiology of Language*, pages 1–16.

Arehalli, S., Dillon, B., and Linzen, T. (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.

Bar, M. (2007). The proactive brain: using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, 11(7):280–289.

Bever, T. G. (1970). The cognitive basis for linguistic structures. In Hayes, J. R., editor, *Cognition and the development of language*, pages 279–362. New York: John Wiley and Sons.

Brothers, T. and Kuperberg, G. (2021). Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of Memory and Language*, 116.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3:296–322.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28.

Caucheteux, C., Gramfort, A., and King, J.-R. (2023). Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*, pages 1–12.

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

Christianson, K., Hollingworth, A., Halliwell, J., and Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology*, 42:368–407.

Davies, M. (2019). The Corpus of Contemporary American English (COCA). *Available online at https://www.english-corpora.org/coca/.*

Dell, G. S., Kelley, A. C., Hwang, S., and Bian, Y. (2021). The adaptable speaker: A theory of implicit learning in language production. *Psychological Review*, 128(3):446.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dillon, B., Andrews, C., Rotello, C. M., and Wagers, M. (2019). A new argument for co-active parses during language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 45(7):1271.

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.

Eisape, T., Zaslavsky, N., and Levy, R. (2020). Cloze distillation: Improving neural language models with human next-word prediction. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 609–619, Online. Association for Computational Linguistics.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, 7(2):195–225.

Fodor, J. and Ferreira, F. (1998). *Reanalysis in sentence processing*, volume 21. Springer Science & Business Media.

Frank, S. and Hoeks, J. C. (2019). The interaction between structure and meaning in sentence comprehension: Recurrent neural networks and reading times. In *Proceedings for the 41st Annual Meeting of the Cognitive Science Society*, pages 337–343. Cognitive Science Society.

Frank, S. L., Fernandez Monsalve, I., Thompson, R. L., and Vigliocco, G. (2013). Reading time data for evaluating broad-coverage models of english sentence processing. *Behavior Research Methods*, 45:1182–1190.

Frazier, L. (1979). *On comprehending sentences: Syntactic parsing strategies*. PhD thesis, University of Connecticut.

Frazier, L. and Clifton, C. (1998). Sentence reanalysis, and visibility. In *Reanalysis in sentence processing*, pages 143–176. Springer.

Frazier, L. and Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, 14(2):178–210.

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

Garnsey, S. M., Pearlmutter, N. J., Myers, E., and Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language*, 37(1):58–93.

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

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

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., et al. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380.

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

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

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

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., and Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

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).

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., Dyer, C., Kuncoro, A., and Brennan, J. (2018). Finding syntax in human encephalography with beam search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2727–2736, Melbourne, Australia. Association for Computational Linguistics.

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

Honnibal, M. and Montani, I. (2017). spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. `https://spacy.io/`.

Hu, J., Gauthier, J., Qian, P., Wilcox, E., and Levy, R. (2020). A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, 20(2):137–194.

Just, M. A., Carpenter, P. A., and Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, 111(2):228.

Knief, U. and Forstmeier, W. (2021). Violating the normality assumption may be the lesser of two evils. *Behavior Research Methods*, 53(6):2576–2590.

Konieczny, L. (2000). Locality and parsing complexity. *Journal of psycholinguistic research*, 29:627–645.

Kutas, M., DeLong, K. A., and Smith, N. J. (2011). A look around at what lies ahead: Prediction and predictability in language processing. In *Predictions in the brain: Using our past to generate a future*, pages 190–207.

Lau, E. and Tanaka, N. (2021). The subject advantage in relative clauses: A review. *Glossa: a journal of general linguistics*, 6(1).

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

Levy, R. (2013). Memory and surprisal in human sentence comprehension. In van Gompel, R. P. G., editor, *Sentence Processing*, pages 78–114. Psychology Press, London and New York.

Levy, R., Fedorenko, E., Breen, M., and Gibson, E. (2012). The processing of extraposed structures in english. *Cognition*, 122(1).

Levy, R., Reali, F., and Griffiths, T. (2008). Modeling the effects of memory on human online sentence processing with particle filters. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc.

Linzen, T. and Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, 7:195–212.

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

Luke, S. G. and Christianson, K. (2018). The Provo corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50(2):826–833.

Marvin, R. and Linzen, T. (2018). Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Mitchell, D. C. (1984). An evaluation of subject-paced reading tasks and other methods for investigating immediate processes in reading. In Kieras, D. E. and Just, M. A., editors, *New Methods in Reading Comprehension Research*, pages 69–89. Erlbaum, Hillsdale, NJ.

Nalborczyk, L., Batailler, C., Lœvenbruck, H., Vilain, A., and Bürkner, P.-C. (2019). An introduction to Bayesian multilevel models using brms: A case study of gender effects on vowel variability in standard Indonesian. *Journal of Speech, Language, and Hearing Research*, 62(5).

Oberauer, K., Lewandowsky, S., Awh, E., Brown, G. D., Conway, A., Cowan, N., Donkin, C., Farrell, S., Hitch, G. J., Hurlstone, M. J., et al. (2018). Benchmarks for models of short-term and working memory. *Psychological Bulletin*, 144(9):885.

Oh, B.-D. and Schuler, W. (2023). 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.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.

Paape, D. and Vasishth, S. (2022). Estimating the true cost of garden pathing: A computational model of latent cognitive processes. *Cognitive Science*, 46.

Pearlmutter, N. J., Garnsey, S. M., and Bock, K. (1999). Agreement processes in sentence comprehension. *Journal of Memory and Language*, 41(3):427–456.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Pickering, M. J. and Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, 36(4):329–347.

Prasad, G. and Linzen, T. (2021). Rapid syntactic adaptation in self-paced reading: Detectable, but only with many participants. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 47(7):1156–1172.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45).

Shain, C., Meister, C., Pimentel, T., Cotterell, R., and Levy, R. P. (2022). Large-scale evidence for logarithmic effects of word predictability on reading time. *PsyArXiv*.

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.

Staub, A. (2015). The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. *Language and Linguistics Compass*, 9(8):311–327.

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

Swets, B., Desmet, T., Clifton, C., and Ferreira, F. (2008). Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory & Cognition*, 36(1):201–216.

Taylor, W. (1953). "cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, 30(4).

Traxler, M. J., Morris, R. K., and Seely, R. E. (2002). Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language*, 47(1):69–90.

Traxler, M. J., Pickering, M. J., and Clifton Jr, C. (1998). Adjunct attachment is not a form of lexical ambiguity resolution. *Journal of Memory and Language*, 39(4):558–592.

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 Gompel, R. P. and Pickering, M. J. (2007). Syntactic parsing. In *The Oxford handbook of psycholinguistics*, pages 289–307. Oxford University Press Oxford.

Van Gompel, R. P., Pickering, M. J., Pearson, J., and Liversedge, S. P. (2005). Evidence against competition during syntactic ambiguity resolution. *Journal of Memory and Language*, 52(2):284–307.

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., 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*, pages 1528–1534, Online. Cognitive Science Society.

Vasishth, S., Mertzen, D., Jäger, L. A., and Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, 103:151–175.

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*, volume 30.

Vul, E., Harris, C., Winkielman, P., and Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, 4(3):274–290.

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.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., and Bowman, S. R. (2020). BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Wilcox, E., Vani, P., and Levy, R. (2021). A targeted assessment of incremental processing in neural language models and humans. 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 939–952, Online. Association for Computational Linguistics.

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., and Levy, R. (2020). On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings for the 42nd Annual Meeting of the Cognitive Science Society*, pages 1707–1713. Cognitive Science Society.