Journal of Memory and Language 137 (2024) 104510

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty

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

[a] Department of Psychological and Brain Sciences, University of Massachusetts Amherst, 135 Hicks Way, Amherst, MA, 01003, USA
[b] Department of Computer Science, Macalester College, 1600 Grand Ave, St. Paul, MN, 55105, USA
[c] Department of Linguistics, University of Massachusetts Amherst, 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

ARTICLE INFO

Dataset link: Syntactic Ambiguity Processing Benchmark (Original data)

Keywords:
Sentence processing
Prediction
Surprisal
Language models

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, and operationalized as surprisal (negative log probability)? 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. By estimating the function that relates surprisal to reading times from filler items included in the experiment, we find that the predictions of language models with two different architectures sharply diverge from the empirical 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 next-word prediction is most likely insufficient on its own to explain human syntactic processing.

Introduction

Language comprehension proceeds quickly and efficiently. A central factor invoked to explain this fact is *prediction*: by anticipating upcoming words, readers can rapidly integrate them into their interpretation of the sentence (Kutas, DeLong, & Smith, 2011). This explanation fits with the growing evidence that such next-word prediction is a fundamental principle of linguistic cognition (Dell, Kelley, Hwang, & Bian, 2021; Pickering & Garrod, 2013) and has a key role to play in language acquisition (Chang, Dell, & Bock, 2006; Elman, 1990). 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 (Brown et al., 2020; Peters et al., 2018). 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 next-word 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 & 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 (Goldstein et al., 2022; Goodkind

[*] Corresponding author.
E-mail address: kuanjunghuan@umass.edu (K.-J. Huang).
[1] S.A., M.K., C.M., G.P. are listed in alphabetical order, and so are the senior authors B.D. and T.L.

https://doi.org/10.1016/j.jml.2024.104510
Received 19 April 2023; Received in revised form 12 December 2023; Accepted 7 February 2024
Available online 28 February 2024

& Bicknell, 2018; Smith & Levy, 2013). 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 (Shain, Meister, Pimentel, Cotterell, & Levy, 2022; Smith & Levy, 2013; Wilcox, Gauthier, Hu, Qian, & Levy, 2020; Wilcox, Pimentel, Meister, Cotterell, & Levy, 2023; though see Brothers & Kuperberg, 2021; Hoover, Sonderegger, Piantadosi, & O'Donnell, 2023.); in this work, we adopt this linking function between predictability and reading times.

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 most, if not all, of word-level processing difficulty (Levy, 2008; Smith & 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 (Bar, 2007), 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 (Gibson, 1998; King & Just, 1991; Vani, Wilcox, & Levy, 2021), and the so-called "ambiguity advantage effect" (Traxler, Pickering, & Clifton, 1998).

These qualitative accounts of processing difficulty associated with specific syntactic phenomena join quantitative studies based on measurements made while participants read natural texts (such as newspaper articles); 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 "Surprisal-based vs. embedding-based linking functions" and Antonello & Huth, 2023).

There are limits to the conclusions we can draw from studies that use materials from naturalistic sources, 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, Fedorenko, Breen, & Gibson, 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 & 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 contain a temporary syntactic ambiguity that is ultimately disambiguated towards a less preferred, and typically less likely, structure. They are referred to as garden path sentences because they are said to "lead the reader down the garden path" (that is, give the reader misleading signals). For example, in (1a) below, the word *conducted* signals that the probable analysis of the preceding material (i.e., that the soldiers warned someone about the danger) is incorrect; the correct analysis is the low probability reduced relative clause parse (i.e., the soldiers were the ones warned about the danger). Compare this sentence to (1b), which is a minimally different sentence that does not display such ambiguity.

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

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. Under the strongest version of the surprisal hypothesis, the excess 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 to truly link neural language models to the garden path effect, it needs to not only predict the existence of garden path effects, but also predict 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. While in their study 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, Vani, & Levy, 2021). They interpreted this substantial underestimation of processing difficulty by surprisal as indicating that processes other than prediction, such as syntactic reanalysis (Fodor & Ferreira, 1998; Paape & Vasishth, 2022), are recruited during the comprehension of syntactically complex sentences.

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 psycholinguistic 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 constructions. 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 not only from construction to construction, but also from item to item in the same construction (Frank & Hoeks, 2019; Garnsey, Pearlmutter, Myers, & Lotocky, 1997). 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, Mertzen, Jäger, & Gelman, 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 of the 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 across constructions, 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.

Summary of the research questions addressed by this paper

In summary, we aim to address four central questions regarding prediction in language comprehension, using surprisal estimates from neural language models to operationalize next-word prediction (Hale, 2001; Levy, 2008; for alternative ways to operationalize prediction, see Brothers & Kuperberg, 2021; Hoover et al., 2023 and Section “Implications for theories of sentence processing”).

First, we ask to what degree processing difficulty can be explained by surprisal in some key constructions that have driven psycholinguistic theorizing. 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 materials that are more tightly matched across constructions (see Section “Materials”). In addition to these three constructions, we also evaluate whether surprisal can explain the difficulty of object-extracted relative clauses compared to subject-extracted ones, the ambiguity advantage in relative clause attachment, and the ungrammaticality penalty in subject-verb agreement dependencies.

Second, we ask whether language model surprisal can correctly predict the relative difficulty among the three garden path constructions. While in van Schijndel and Linzen’s study language models made predictions that appeared not to match the rank 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.

Third, while van Schijndel and Linzen used only LSTM language models, we also evaluate a more 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“Computing language model surprisal”).

Finally, we ask how well language model surprisal can explain itemwise variation in processing difficulty within the same syntactic construction. Existing work evaluating the item-level predictions of surprisal on targeted linguistic contrasts has been limited to small sample sizes (Frank & 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.

Methods

Fig. 1 summarizes the core methodology of this paper. In a nutshell, we first estimated the function that relates surprisal to reading times by using the filler material as our training data. We then applied this function to our testing data – the critical words in the experimentally manipulated material – to predict RTs on those words. Finally, we evaluated how closely the predicted RTs matched the empirical ones. In this section, we describe each of these steps in detail.

The Syntactic Ambiguity Processing Benchmark: Dataset construction

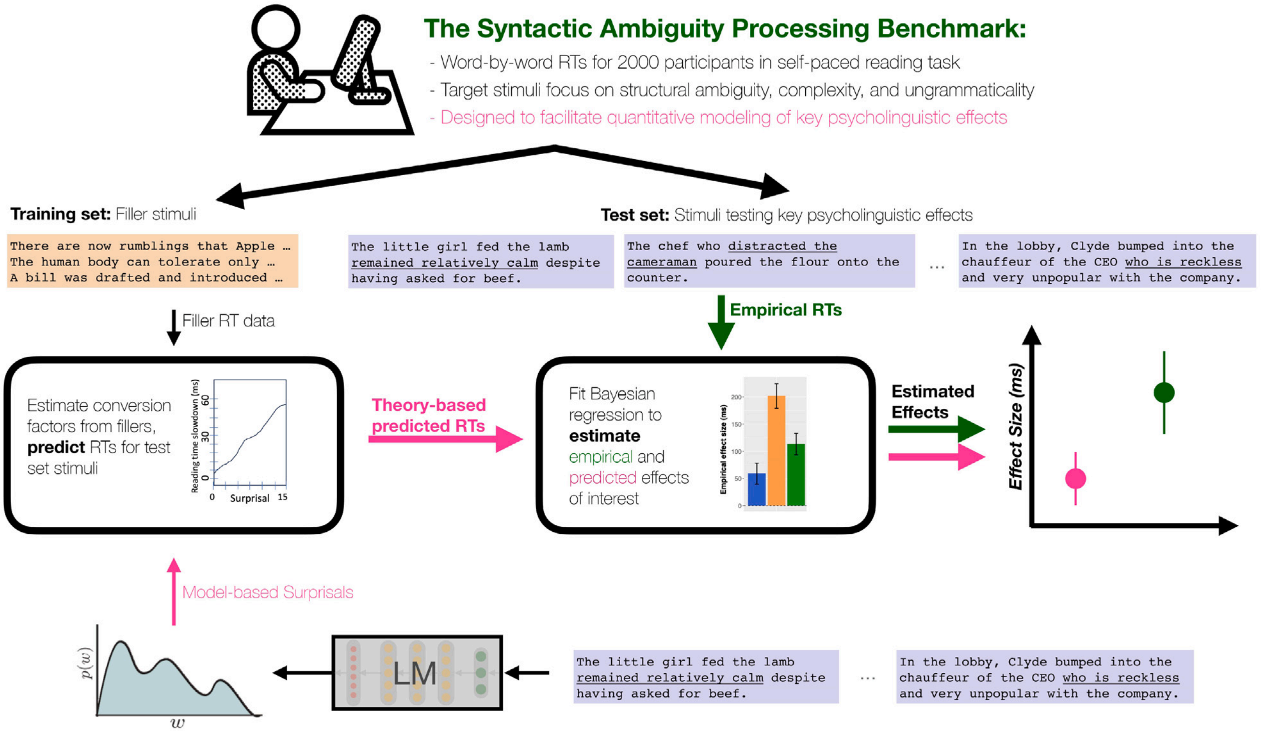
As we described in the Introduction, the SAP Benchmark is a large-scale dataset that serves two purposes. First, we use it to empirically evaluate the ability of surprisal to explain human comprehension difficulty in syntactically complex sentences; and second, we intend for it to serve as a resource for the quantitative evaluation of other theories of sentence processing. In this subsection, 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, Carpenter, & Woolley, 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 & Christianson, 2018). The constructions are exemplified in Table 1. For all of our target constructions there are arguments in the literature attributing processing difficulty to surprisal (Hale, 2001; Levy, 2008; Vani et al., 2021; Wilcox et al., 2021). We describe and motivate the inclusion of 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 ambiguity (occasionally referred to as the NP/S ambiguity; Frazier, 1979, the Transitive/Intransitive ambiguity (also 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 Direct Object/Sentential Complement and Transitive/Intransitive constructions by Sturt, Pickering, and Crocker (1999) and Grodner, Gibson, Argaman, and Babyonyshev (2003). Sturt and colleagues created lexically matched item sets for the Direct Object/Sentential Complement and Transitive/Intransitive constructions; we extend this methodology to the Main Verb/Reduced Relative construction and create 24 lexically matched item sets for each of the three 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 & Tanaka, 2021). This difficulty is thought in part to reflect the relative unpredictability of ORCs (Chen & 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, even by proponents of surprisal theory (Levy, 2013), to involve memory-related difficulties above and beyond the effects of predictability.

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 and unambiguous relative clause attachment configurations. Previous work has found a processing advantage associated with globally ambiguous RC attachment (the **ambiguity**


The Syntactic Ambiguity Processing Benchmark:
- Word-by-word RTs for 2000 participants in self-paced reading task
- Target stimuli focus on structural ambiguity, complexity, and ungrammaticality
- Designed to facilitate quantitative modeling of key psycholinguistic effects
Training set: Filler stimuli
Test set: Stimuli testing key psycholinguistic effects
Filler RT data
Empirical RTs
Estimate conversion factors from fillers, predict RTs for test set stimuli
Reading time slowdown (ms)
0
30
60
Surprisal
0
15
Theory-based predicted RTs
Fit Bayesian regression to estimate empirical and predicted effects of interest
Empirical effect size (ms)
200
150
100
50
0
Estimated Effects
Effect Size (ms)
Model-based Surprisals
LM
p(w)
w
…

There are now rumblings that Apple …
The human body can tolerate only …
A bill was drafted and introduced …

The little girl fed the lamb remained relatively calm despite having asked for beef.

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

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

Fig. 1. A visualized schematic summary of the SAP Benchmark project. Green color indicates the empirical data coming from the SAP Benchmark itself; pink color indicates the procedures we use to test the hypothesis linking language model surprisal to reading times. Examplees of filler stimuli are in the orange rectangle and examples of experimental stimuli, with critical region underlined, in the purple rectangles.

advantage effect; Traxler, Morris, & Seely, 2002; Traxler et al., 1998; Van Gompel, Pickering, Pearson, & Liversedge, 2005). Like garden path effects, this effect has been argued to arise from predictability: ambiguous continuations are compatible with both parses, and are hence assigned a higher probability – the sum of the probabilities assigned by each parse – than unambiguous continuations. Consequently, under surprisal theory they are associated with reduced processing difficulty compared to unambiguous continuations (Levy, 2008).

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.

The last subset examines the processing of **subject-verb agreement**. The sentences in this subset 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 they cause processing difficulty (Wagers, Lau, & Phillips, 2009). Unlike in garden path constructions, however, it is not possible to reanalyze these items to yield an acceptable structure: Under no reading or parse is the sentence well-formed.

For all constructions, we defined a region of interest (ROI) that represents a key part of the sentence in which we expect processing difficulty (highlighted in red in Table 1). In constructions that involve syntactic ambiguity, this was the critical word that disambiguated the sentences and the following two spillover words, whereas in the subject-verb agreement this was the verb with the mismatching number and the following two spillover words. Then, for every construction we estimated the *effect of interest* (EOI) by comparing the processing time at each word in the ROI in the target sentences (ambiguous or ungrammatical sentences) and control sentences (unambiguous or grammatical sentences).

These EOIs are the target of our modeling efforts: They isolate the unique processing difficulty associated with the syntactic difference between two sentences, controlling for lexical factors such as unigram log-frequency and word 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 items.

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. For the ambiguity advantage subset we used the materials of Dillon, Andrews, Rotello, and Wagers (2019); this subset included three conditions, Low Attachment, High Attachment, and Ambiguous Attachment. For the relative clause subset we used the materials of Staub (2010). Filler items were drawn from the Provo Corpus (Luke & Christianson, 2018). See Appendix D for a full list of the items.

Garden path subset

The strength of garden path effects is influenced by various factors such as plausibility and verb subcategorization frequencies (Garnsey et al., 1997). To control for these 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 in one of our garden-path constructions, such that the less frequent parse was attested in a locally ambiguous form. This process helped to ensure that the garden path items in the ambiguous condition were in fact locally 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 fact that ‘move’ can be either transitive or intransitive, and on the absence of a comma between ‘moved’ and ‘the mail’. As such, a verb was only considered to be eligible for inclusion in the Transitive/Intransitive condition if it was attested at least once in the corpus in an intransitive frame without such a comma (e.g., ‘Before the cousins moved it was a different story’,

Table 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 at or immediately following a disambiguating or ungrammatical word, marked in red, 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 blue. The rightmost column lists the hypotheses that have been proposed in the literature to explain the processing difficulty. (For interpretation of the references to color in this table, the reader is referred to the web version of this article.)

Construction	Example	Hypothesis
Main Verb	The girl fed the lamb ***remained relatively calm*** ...	1. Surprisal is sufficient (Hale, 2001; Levy, 2008). 2. Reanalysis is necessary (Paape & Vasishth, 2022; van Schijndel & Linzen, 2021).
Reduced Relative	The girl who was fed the lamb ***remained relatively calm*** ...	
Direct Object	The girl found the lamb ***remained relatively calm*** ...	
Sentential Complement	The girl found that the lamb ***remained relatively calm*** ...	
Transitive	When the girl attacked the lamb ***remained relatively calm*** ...	
Intransitive	When the girl attacked, the lamb ***remained relatively calm*** ...	
Object Relative Clause	The bus driver that ***the kids followed*** ...	Surprisal is insufficient (Staub, 2010; Vani et al., 2021). Difficulty is in part memory-related (Gibson, 1998).
Subject Relative Clause	The bus driver that ***followed the kids*** ...	
Relative Clause Modifying Recent Noun (Low Attachment)	Janet charmed the executives of the assistant who ***decides almost everything*** ...	Surprisal is sufficient (Levy, 2008).
Relative Clause Modifying Distant Noun (High Attachment)	Janet charmed the executive of the assistants who ***decides almost everything*** ...	
Relative Clause modifying either noun	Janet charmed the executive of the assistant who ***decides almost everything*** ...	
Subject-Verb Not Agreeing	Whenever the nurse calls, the doctors ***stops working immediately*** ...	Ungrammatical is also unpredictable (Wilcox et al., 2021). But no reanalysis is possible.
Subject-Verb Agreeing	Whenever the nurse calls, the doctor ***stops working immediately*** ...	

Southwest Review, 2002). For Direct Object/Sentential Complement verbs, we required that a local ambiguity arising from the absence of a complementizer was attested at least once (e.g., '[...] the AJC found none of his companies followed the rules', Atlanta Journal Constitution, 2014). Lastly, we only included Main Verb/Reduced Relative verbs that were attested inside a reduced relative clause, that is, without the relative pronoun and copula (e.g., 'Two patient assigned conventional therapy died', Lancet, 2000).

To do so, we queried COCA for all sentences matching the pattern **DP VERB DP** for each verb considered (e.g., DP *moved* DP). The set of verbs we performed queries for was based on the authors' intuitions. The sentences that resulted from the query were then parsed using the spaCy natural language processing library (Honnibal & Montani, 2017) and labeled as to whether or not the disambiguating verb was the main verb. The output of this automated process 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 in two different items, while six Main Verb/Reduced Relative verbs occurred in two items and the remaining three in four items. Any repetition of a verb occurred in an entirely different frame such that all the content words were different across sentences that shared the verb. Crucially, 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.

Agreement subset

The items in the agreement subset were derived from the Transitive/Intransitive items to allow for a closer comparison of reading times in grammatical, but unexpected, garden path sentences on the one hand, and ungrammatical sentences with agreement violations on the other hand. Concretely, for every item in this subset, there was a corresponding item in the Transitive/Intransitive condition with the same ambiguous verb, disambiguating verb and the first word of the spillover region. For instance, the ungrammatical agreement item 'When the magician moves, the cards disappears mysteriously from his assistant's hand' corresponds to the Transitive/Intransitive example 'After the woman moved the mail disappeared mysteriously from the delivery system'.

Other constraints

We imposed a number of additional standard controls for reading experiments for the two subsets we constructed. All disambiguating words had six or more characters to reduce the likelihood that readers will of skip the word in the planned eyetracking-during-reading version of the benchmark. For the classic garden path subset, while endings of the sentences might differ, each item has the same disambiguating word and the same two words to its right (e.g., *remained relatively calm*) across the three types of garden path constructions. The total length in characters of each sentence was limited such that the sentences fit in a single line. Finally, we checked that the vocabulary of our stimuli was a subset of the vocabularies of both the Penn Treebank Corpus and the Wikipedia training data from Gulordava, Bojanowski, Grave, Linzen, and Baroni (2018). This was done to ensure that a wide range of language models could be used to derive predictability estimates for the SAP Benchmark, including those trained on supervised parses from the Penn Treebank.

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. To help participants calibrate their responses, the norming experiments included a number of highly implausible fillers (e.g., 'The gentleman pleased the most demanding follower of quizzes evenly'). These fillers provided a highly implausible baseline that helped participants calibrate their responses. All judgments were provided using a 7-point Likert scale. The items were adjusted over multiple rounds of norming until the mean plausibility rating of the items for each condition exceeded 5 points on the 7-point scale. The final round of norming satisfying these restrictions included judgments from 68 participants.

In addition to norming the plausibility of each full sentence, we also normed the plausibility of parts of each sentence, such as the plausibility of the temporary garden path interpretation (e.g., 'The girl fed the lamb'). These supplementary norms were used for a separate analysis with local plausibility as a predictor of the magnitude of garden path effects (Garnsey et al., 1997; Pickering & Traxler, 1998; Van Dyke & Lewis, 2003); for details, see Appendix C.

Participants

Our goal was to include data from 2000 participants in the SAP Benchmark. We first recruited 2000 participants who spoke English as their first language. Participants were recruited on Prolific with a monetary compensation rate at around $12 per hour. The age of all participants was between 18 and 45. Of the 2000 recruited, 1867 were speakers of North American English from either 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. We then excluded from analysis all participants whose accuracy on the comprehension questions for the fillers was below 80%; our exclusion criteria are detailed in our preregistration document, available at https://osf.io/9865s. This resulted in the exclusion of an additional 179 participants. To make up for the excluded participants, we then recruited another batch of participants, and repeated the process until the original target of 2000 valid participants was reached.

Additional data exclusion criteria

In addition to subject-level exclusions described above, we also excluded from analysis all observations at the critical positions with reading times (RTs) greater than 7000 ms. 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.

Procedure

The materials were presented in the self-paced reading paradigm. In this paradigm, the words of the sentence are first obscured 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 used as an indicator of the difficulty of processing the current word. In an experimental session, a participant read 92 sentences. These included 40 fillers and 52 sentences from the 13 experimental conditions (three for Low Attachment/High Attachment, and two each for Transitive/Intransitive, Direct Object/Sentential Complement, Main Verb/Reduced Relative, Object vs. Subject Relative Clause, and Agreement Violation). Before these 92 sentences, four practice trials were presented. Each sentence was followed by a comprehension question. An experimental session lasted approximately 25 min on average. The study was run on PCIbex farm (Zehr & Schwarz, 2018).

To avoid changes in processing times over the course of the experiment due to increased familiarity with any particular construction (*syntactic adaptation*; Fine, Jaeger, Farmer, & Qian, 2013), only four items from each condition were presented to each participant, counterbalanced using a Latin square. Items were presented in a random order, subject to the constraint that a critical item was never followed by another item from the same condition.[2]

Estimating the Effects of Interest

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

Analyzing raw RTs vs. log-transformed RTs

Reaction times are typically right-skewed and heteroskedastic, with variance increasing as a function of the mean. This property of the data can lead to violation of two assumptions of linear regression: normality of residuals and homogeneity of variance. To mitigate this issue, some studies log-transform RTs before entering them into the regression (e.g., Frank, Fernandez Monsalve, Thompson, & Vigliocco, 2013, see also Knief & Forstmeier, 2021). In this study, we avoid log-transforming our RTs, as we argue that this transformation makes two assumptions about the relationship between our predictors and RTs that are unjustified on both theoretical and empirical grounds.

First, in a linear regression with RTs as the dependent variable, RTs are predicted from a weighted sum of the predictors. Log transforming RTs encodes the assumption that the surprisal has a *multiplicative* effect on the original RT scale, as illustrated by the following equation (for ease of exposition we show a simplified regression that does not include all of our predictors).

$$\begin{aligned} log(RT(w_n \mid context)) = {} & \beta_0 + \beta_1 \cdot Surp(w_n \mid context) \\ & + \beta_2 \cdot Freq(w_n) + \beta_3 \cdot Length(w_n) + \epsilon \end{aligned}$$

$$\begin{aligned} RT(w_n \mid context) &= e^{\beta_0 + \beta_1 \cdot Surp(w_n|context) + \beta_2 \cdot Freq(w_n) + \beta_3 \cdot Length(w_n) + \epsilon} \\ &= e^{\beta_0} \cdot e^{\beta_1 \cdot Surp(w_n|context)} \cdot e^{\beta_2 \cdot Freq(w_n)} \cdot e^{\beta_3 \cdot Length(w_n)} \cdot e^{\epsilon} \end{aligned}$$

By assuming a multiplicative effect on raw RTs, we are assuming that surprisal, frequency, and length effects can interact at some level of processing (Roberts & Sternberg, 1993). This assumption may not be justified: For instance, some prior work suggests that frequency and predictability are empirically distinct and dissociable (as systematically reviewed in Staub, 2015, also see Shain, 2023 for recent data).

Second, as shown in the equation, log-transformation assumes an exponential, rather than linear, relationship between raw RTs and surprisal. This violates the theoretical predictions made by surprisal theory (Smith & Levy, 2013) as well as the empirically observed relationship between predictability and RTs (Shain et al., 2022; Wilcox et al., 2023).

Thus, while the log transform helps reduce the rightward skew of the RTs, it violates prior theoretical commitments and empirical observations about surprisal. Since a central goal of this work is to evaluate surprisal theory, we focus on raw RTs in the main text of this article. To establish that any conclusions reached here do not critically rest on our choice of dependent measure and analysis scheme, we also repeat all of our analyses with log-transformed RTs as the dependent variable; we present these analyses in Appendix A. Overall, while the estimated EOIs were much smaller when RTs were log-transformed, these analytical choices did not qualitatively change any of the conclusions reported in the main text.

[2] 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 104,000 trials). This decision was not preregistered.

Table 2
The priors we used for our Bayesian mixed-effects models.

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 3
Summary of estimated coefficients from our statistical models across all constructions. The notation“:” represents an interaction between two predictors. The 7 EOIs were computed from these coefficients as follows: Main Verb/Reduced Relative = β_1; Transitive/Intransitive = $\beta_1 + \beta_4$; Direct Object/Sentential Complement = $\beta_1 + \beta_5$; Object vs. Subject Relative Clause = β_6; High attachment = $-\beta_7 + 0.5\beta_8$; Low attachment = $-\beta_7 - 0.5\beta_8$; Agreement violation = β_9. To estimate the EOIs, we computed these values for each posterior sample, and then averaged together; The standard error is the standard deviation of this aggregated posterior sample distribution.

Subset	Coef	Predictor	Comparison (coding)
Classic Garden Paths	β_1	Ambiguity	Unambiguous (0) vs. Ambiguous (1)
	β_2	Type1	Main Verb/Reduced Relative (0) vs. Direct Object/Sentential Complement (1)
	β_3	Type2	Main Verb/Reduced Relative (0) vs. Transitive/Intransitive (1)
	β_4	Ambiguity: Type1	Main Verb/Reduced Relative GPE vs. Direct Object/Sentential Complement GPE
	β_5	Ambiguity: Type2	Main Verb/Reduced Relative GPE vs. Transitive/Intransitive GPE
Relative Clauses	β_6	RC type	Subject RC (0) vs. Object RC (1)
Attachment Ambiguities	β_7	Ambiguity	Ambiguous (2/3) vs. Unambiguous (-1/3)
	β_8	Height	High (1/2) vs. Low (-1/2).
Subject-verb Agreement	β_9	Grammaticality	Grammatical (0) vs. Ungrammatical (1)

Priors for the Bayesian models

Table 2 lists the weakly informative priors we used in our Bayesian regression models and provides an intuition for the set of values on which most of the prior probability mass is concentrated. Due to 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.

Estimating empirical EOIs

We fit 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 by-item random effects. 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). Below we describe the specific model structure we used for each subset. For additional details about the specific coding of each predictor, and how model coefficients were used to estimate EOIs, see Table 3; for details about the model fitting procedure that resulted in the random effect structures we used, see the last paragraph of this section.

Classic garden path constructions. There were three EOIs associated with this subset, one for each of the garden path effects. We estimated these EOIs by fitting the model below (we describe all modelsd using R formula notation):

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

Relative clauses. There is one EOI associated with this subset: the difference in reading times on the critical word between subject and object RCs. In this subset, the critical word of interest was the determiner, which occurred at different linear positions across conditions.[3] To correct for any independent effect that word position may have on RTs, we first fit the following linear mixed-effects model to the filler sentences, and used it to regress out word position for the critical sentences (Van Dyke & 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:

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

The residualization process used to generate the position-corrected RTs eliminated any differences in the mean RTs between participants. Therefore, the model formula does not include a random intercept for participant (as indicated by the 0 in the participant random effect structure).

Attachment ambiguities. There are two EOIs associated with this subset: high attachment garden path and low attachment garden path. We

[3] This is a departure from our preregistration document, which incorrectly identified 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 RCs from subject RCs, which in our experimental sentences is the determiner.

estimated these effects by fitting the model below:

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

Subject-verb agreement. There is one EOI associated with this subset: agreement violation. We considered two kinds of sentences: grammatical unambiguous sentences from the Transitive/Intransitive subset, and ungrammatical versions of those sentences containing an agreement error. We estimated the agreement violation effects by fitting the following model:[4]

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

Model fitting details. To fit the models to the data, four independent Markov Chain Monte Carlo chains of 6000 iterations each were used to draw samples from the posterior distribution of the model. The first half of the samples of each chain were discarded as warm-up samples. The number of iterations was increased when necessary. For each subset, we started by trying to fit a model with the maximal random effect structure justified by that subset. In cases where the between-chains variability for the maximal models, as indexed by $\hat{R}$, was greater than 1.05, we backed-off to models with simpler random effects structure: we first removed the correlation between the random slopes and intercepts (this is indicated in the R formulas above by ||), and then, if necessary, we removed the random slopes corresponding to interaction terms. In subsets where the richest random effect structure that allowed the model to converge differed across the three words of the disambiguating region, we used the simplest random effect structure of the three for all three words. At the conclusion of this process, the $\hat{R}$ values for all the estimates in our models were lower than 1.05, indicating that the chains converged to the posterior distribution Nalborczyk, Batailler, Lœ venbruck, Vilain, and Bürkner (2019).

Estimating predicted RTs

We generated the predicted EOIs in three 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 – based on the filler items. Finally, we multiplied the surprisal values by the conversion factors to obtain the predicted EOIs.

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 & 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 model is based on the Transformer architecture (Vaswani et al., 2017), and was trained on approximately 40 GB of data scraped from the Web. While neither model is trained on any explicit syntax, they 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, Gauthier, Qian, Wilcox, & Levy, 2020; van Schijndel & Linzen, 2021; Warstadt et al., 2020), and as such are promising candidates for modeling human syntactic expectations.

Table 4
Coefficient estimates for the models fit to the fillers; for example, surprisal$_{w_{n-1}}$ indicates the effect in milliseconds of each additional unit of surprisal of word $n-1$ on reading times on word n. Note that this table reports models with 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.

LSTM		**GPT-2**	
Predictor	$\hat{\beta}$	*Predictor*	$\hat{\beta}$
Word position	**-1.49**	Word position	**-1.26**
Surprisal$_{w_n}$	**0.95**	Surprisal$_{w_n}$	**1.12**
Surprisal$_{w_{n-1}}$	**0.81**	Surprisal$_{w_{n-1}}$	**1.12**
Surprisal$_{w_{n-2}}$	**0.12**	Surprisal$_{w_{n-2}}$	**0.58**
Surprisal$_{w_{n-3}}$	**0.37**	Surprisal$_{w_{n-3}}$	**0.24**
Log-Freq$_{w_n}$	**1.02**	Log-Freq$_{w_n}$	**0.43**
Log-Freq$_{w_{n-1}}$	**0.57**	Log-Freq$_{w_{n-1}}$	0.07
Log-Freq$_{w_{n-2}}$	-0.32	Log-Freq$_{w_{n-2}}$	-0.33
Log-Freq$_{w_{n-3}}$	**1.30**	Log-Freq$_{w_{n-3}}$	**0.89**
Length$_{w_n}$	**11.3**	Length$_{w_n}$	**9.53**
Length$_{w_{n-1}}$	**12.6**	Length$_{w_{n-1}}$	**10.9**
Length$_{w_{n-2}}$	**3.46**	Length$_{w_{n-2}}$	**2.73**
Length$_{w_{n-3}}$	**1.90**	Length$_{w_{n-3}}$	**1.46**
Freq×Length$_{w_n}$	**-0.69**	Freq×Length$_{w_n}$	**-0.51**
Freq×Length$_{w_{n-1}}$	**-0.87**	Freq×Length$_{w_{n-1}}$	**-0.69**
Freq×Length$_{w_{n-2}}$	**-0.31**	Freq×Length$_{w_{n-2}}$	**-0.23**
Freq×Length$_{w_{n-3}}$	**-0.20**	Freq×Length$_{w_{n-3}}$	**-0.14**

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 (frequentist) 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 (Smith & Levy, 2013).

In addition to surprisal-based predictors, this model included as predictors the position of the word in the sentence, its length, its 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 such 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 for the first three words of a sentence, which are not preceded by a three-word spillover context. We also followed prior work (Smith & Levy, 2013, for example) in excluding the final word of each sentence, as these words display wrap-up effects that are beyond the scope of our modeling goals (Just et al., 1982).

The resulting models (Table 4) offer a set of **conversion factors** that estimate how reading time on the fillers co-vary with surprisal and the other predictors.

Generating predicted reading times

In the third step, we use the conversion factors we computed to generate predicted reading times for each of the critical subsets. As a control, we also 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 do include surprisal provides a conservative estimate of how much of the empirical garden path effect is accounted for by surprisal, over and above unigram statistics and their spillover effects.

[4] In our preregistration, we included word position and its interaction with grammaticality as factors. However, due to convergence issues, we fit a separate model for each word position instead.

Comparing empirical and predicted effects

We evaluated whether our empirical estimates of processing difficulty aligned with language-model-derived surprisal both at the construction level and at the item level. At the construction level, we fit the Bayesian mixed-effect models described in Section “Estimating Empirical EOIs” to both the empirical and predicted data. Then, we compared the resulting coefficients from these two sets of models.

We also evaluated 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 the item-wise correlation coefficient within each construction using the following Monte Carlo approach, which leveraged the item-wise posterior EOI estimates. We independently sampled one observation from the posterior distribution of each item’s empirical EOI, as well as one observation from the corresponding model-based prediction for the EOI. This resulted in two numbers for each item. A construction’s correlation coefficient was then computed as the correlation between the two quantities – empirical and predicted – across all items within the construction. We repeated this procedure 1000 times for each construction, separately for each of the language models as well as for the No-surprisal baseline model.

Any correlation coefficient should be interpreted in the context of the intrinsic noise in the reading time measures, which limits the highest possible correlations that could be observed (Schrimpf et al., 2021). We estimated the amount of explainable variance in each of the four experimental subsets by running 15 split-half reliability analyses, as follows. In each of the 15 iterations of this procedure, the 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 half of the 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: we cannot expect a predictor to show a greater correlation with the empirical data than two halves of the data show with each other (Vul, Harris, Winkielman, & Pashler, 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 split-half reliability measure for our filler items. This was done similarly to the split-half analysis described in the last paragraph, with one exception: Here, 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 was treated as a unique item. For each iteration, 24 out of 498 words were randomly selected, to match the number of items in the critical subsets. 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.

Results

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, some of the comprehension question specifically evaluated whether participants successfully resolved the syntactic ambiguity. 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?*. Accuracy on such questions varied across constructions, with the lowest accuracy observed for Transitive/Intransitive (37.2%), Low Attachment (55.2%) and Main Verb/Reduced Relative (44.1%). The low accuracy associated with these three constructions is consistent with earlier findings (Christianson, Hollingworth, Halliwell, & Ferreira, 2001; Dillon et al., 2019; Prasad & Linzen, 2021). The remaining comprehension questions for the critical items targeted other aspects of the sentence, and accuracy on those questions was high (around 82% for Object vs. Subject Relative Clause, and above 90% for all other constructions). We present the full accuracy data in Appendix B.

Construction-level reading times

Fig. 2 presents the average effect of interest at the critical disambiguating word and the following two words, for each construction in our dataset. In this section, we focus on the effects of interest at the word following the critical word, rather than at the critical word itself. We focus on this region because it showed the largest effect for most effects of interest, with the exception of Object vs. Subject Relative Clause, where the strongest effect was two words after the critical word, and Low Attachment, where there was no discernible effect on any of the three words we analyzed.

On the word following the critical word, there were robust effects in five out of the seven constructions included in the experiment, with the largest effect in Main Verb/Reduced Relative (202.1 ms [179.2–224.4]; the range in the brackets indicates 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 High Attachment were smaller (Direct Object/Sentential Complement: 63.9 ms [34.4–92.1]; High Attachment: 26.9 ms [15.8–36.4]). The ungrammaticality effect for Agreement Violation was smaller than the largest garden path effects but still highly robust (57.4 ms [44.9–69.7]). Finally, 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, processing difficulty arises in relative clauses with high attachment, but not low attachment (Swets, Desmet, Clifton, & Ferreira, 2008). Third, outright subject-verb agreement mismatch reliably slows down reading (Wagers et al., 2009). Fourth, as in prior self-paced reading studies, there was no reliable object relative clause difficulty at the determiner or noun position (Grodner & Gibson, 2005). In addition to these previously established patterns in a highly powered experiment, we find that disambiguation is harder in the Main Verb/Reduced Relative ambiguity than the Transitive/Intransitive or Direct Object/Sentential Complement ambiguities. This establishes a difficulty ranking across these three widely studied garden path constructions 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 the same 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 clear difference between SRCs and ORCs at the determiner or the noun. Instead, we saw slower RTs for ORCs compared to SRCs at the verb position (see Fig. 2). The time course of this effect appears to be inconsistent with the predictions of surprisal theory, according to which the effect should localize to the subject noun phrase in an ORC construction (Hale, 2001). Instead, it is more consistent with theories that attribute the slower reading of ORCs to difficulty integrating a distant argument at the verb (Gibson, 1998).

We caution against interpreting 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 which

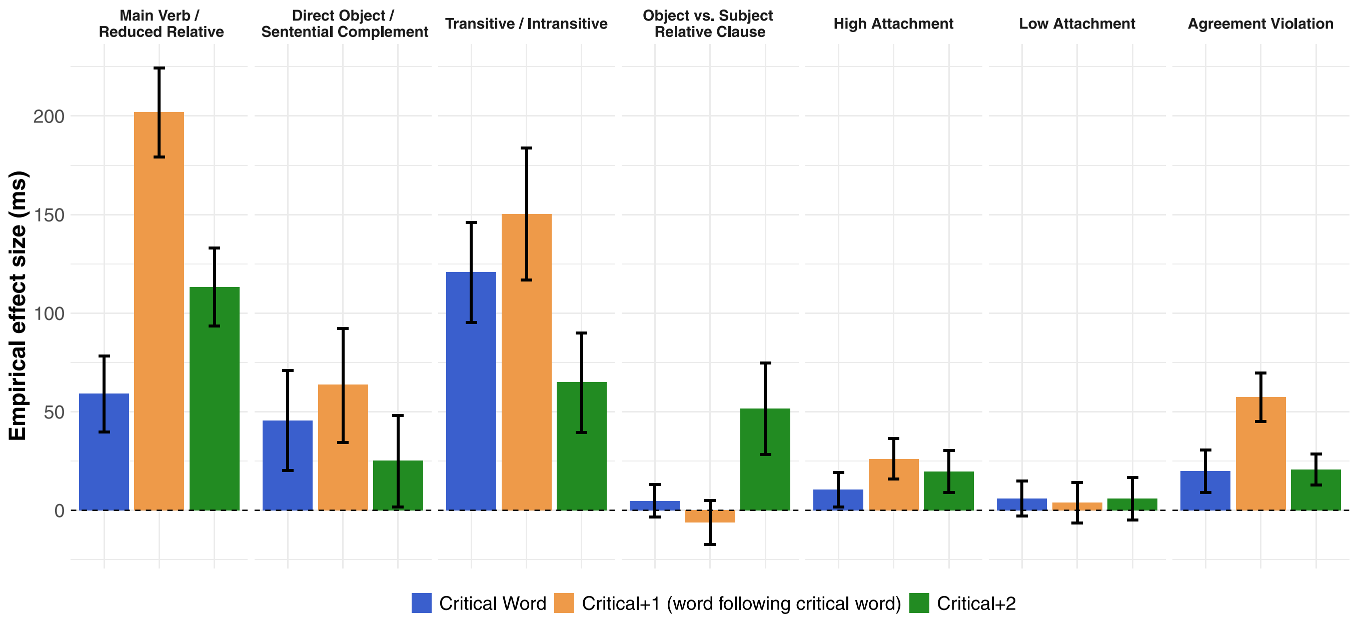


Fig. 2. Posterior estimates of effect sizes at the three regions of interest for each effect of interest. Error bars represent 95% credible intervals. Note that for the Object vs. Subject Relative Clause contrast, the critical positions in the subject and object relative clauses are reversed so that the same word can be compared. In this contrast, we treat the determiner as the critical ROI, the noun as the first spillover region, and the verb as the second.

immediately precedes it; such spillover effects are common in self-paced reading (indeed, as we noted above, in most of our constructions disambiguation effects appear most clearly in the word following the disambiguating word). It is also possible that the lack of an effect at this position reflects spillover effects from the preceding context, which differed between SRCs and ORCs. This conjecture is supported by the observation that the No-surprisal baseline predicted a *negative* effect at the determiner and the noun regions: any difference across conditions predicted by this baseline can only reflect the spillover of lexical effects (frequency and length) from previous words. Since we did not see such a negative effect in the empirical reading times in these regions, it is possible that the slowdown attributable to surprisal in the ORC condition cancels out the speedup attributable to these other spillover factors. If that is the case, our results may therefore be consistent with other studies using reading paradigms that are less subject to spillover effects, such as the maze task (Vani et al., 2021) or eye-tracking during read (Staub, 2010), which have documented processing costs at both the determiner and the verb in ORCs.

Variability across items

In most of the constructions, the size of the effect of interest varied substantially across items (Fig. 3). The extent of item-level variability differed across the constructions. We quantified the variability across items for each construction using the second-order coefficient of variation (V_2; Kvålseth, 2017), mathematically defined as the following, where s^2 is the sample variance and $\bar{x}$ is the sample mean:

$$V_2 = \frac{s^2}{s^2 + \bar{x}^2}$$

V_2 captures how large the variance is with respect to the mean. Its value is bounded between 0 and 1, which makes it useful for comparing the extent of variation across different datasets. Note that while V_2 is much less affected by the mean than the more commonly used Pearson Coefficient of Variation, it still tends to be particularly high when the mean is close to zero (Kvålseth, 2017). We thus refrain from interpreting V_2 for the two constructions where the construction mean effect is indistinguishable from zero.

In the Direct Object/Sentential Complement and High Attachment constructions, only a subset of the items displayed garden path effects that were distinguishable from 0 ms (about two thirds 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 values of V_2. For Transitive/Intransitive, every item showed a garden path effect that was statistically greater than 0 ms. Yet even in this construction, there was considerable item-level variability, with effects ranging from 59.2 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 V_2 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 C), making it an important target for future modeling. Finally, the Agreement Violation construction has the smallest V_2: every item showed a reliable ungrammaticality effect, but the magnitude of this effect was largely consistent across items.

The differences in variability between constructions are not a simple by-product of differences in construction-wide effect sizes: Agreement Violation and Direct Object/Sentential Complement have similar mean effect sizes, but the former shows much smaller variability than the latter.

Finally, we estimated the reliability of the item-wise variability for each construction, using the split-half analysis described in Section "Comparing empirical and predicted effects", where we repeatedly split the observations for each item into two halves and compared the effects of interest for the item across the two havles. The reliability varied across constructions (Table 5). For the classic garden path constructions, split-half reliability was quite high, all above 0.81. Reliability estimates were lower for Agreement Violation, Low Attachment and High Attachment, ranging from 0.18 to 0.45. The lower reliability by items in Low Attachment and High Attachment likely reflects the fact that in these conditions the effects of interest at the construction level were much smaller or nonexistent.

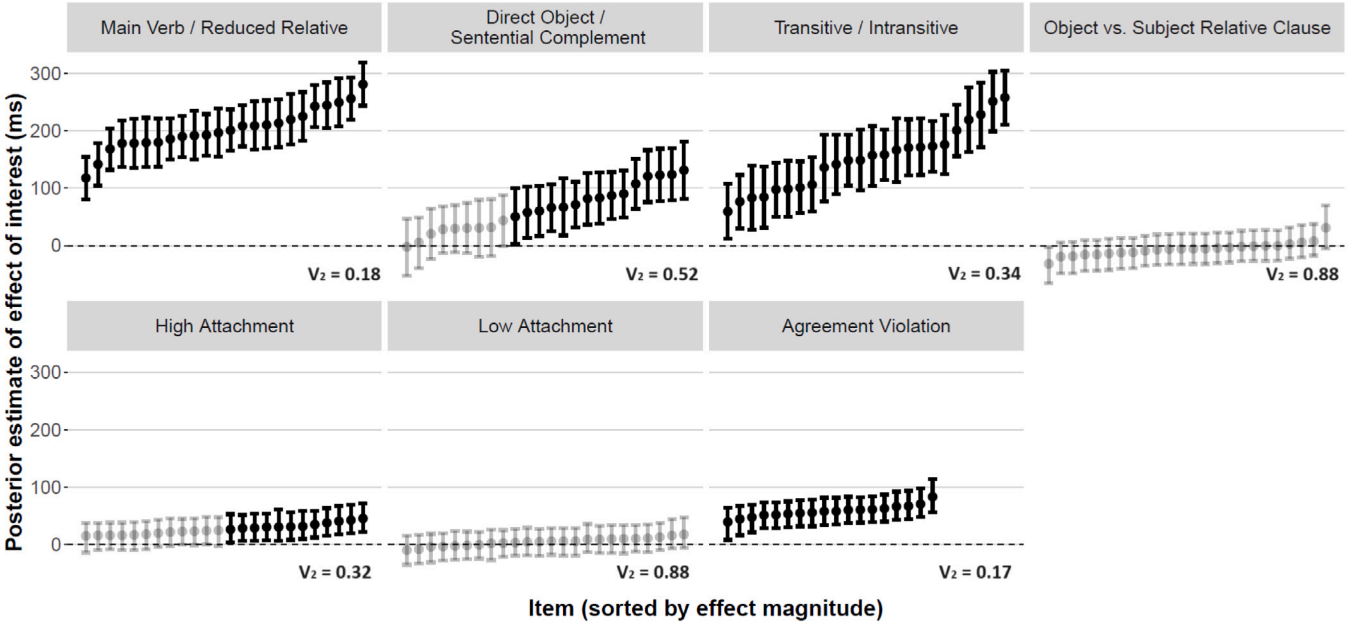


Fig. 3. Empirical effects of interest for each individual item in all seven constructions in the SAP Benchmark. All effects were estimated from a Bayesian mixed-effects regression model fit to raw RTs on the word that indexed the effect of interest. Error bars represent the 95% posterior credible interval on the item-level size of this effect. V_2 is the second-order coefficient of variation.

Table 5
Spearman-Brown-corrected split-half reliability for each effect of interest.

Effect of interest	Split-half reliability
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 Violation	0.45
Fillers	0.99

Comparison to language model surprisal

As we described in detail in Section "Estimating predicted EOIs", we fit three linear mixed-effects models – one with surprisal estimates from each of our two language models, and a baseline model without a surprisal predictor – to the reading times for the 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 computed each model's prediction for the location, direction, and magnitude of each effect. The rest of this section reports the findings of this analysis.

Language model surprisal predicts the existence of human processing difficulty, but does not predict its magnitude. Surprisal from both language models predicted the location and direction of most of the effects of interest tested, with the exception of Object vs. Subject Relative Clause: here both language models predicted a *negative* garden path effect, but such an effect was not seen in the human data (Fig. 4). Because the No-surprisal model, which only included lexical factors and their spillover, predicted an even more dramatic negative difference in this EOI, we suspect that this negative effect reflects differences in the unigram frequency of the pre-critical region, which was by necessity unmatched across the two conditions.

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

Moreover, in most constructions there was 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 size of the predicted effect was much smaller than the empirically observed one. For example, in Main Verb/Reduced Relative, the observed EOI was about 31 times as large as the EOI predicted by Wiki-LSTM or GPT-2 (202.1 ms [179.2–224.4] vs. 6.6 ms [5.8–7.3] or 7.1 ms [5.1–9.1], respectively). This quantitative misalignment held even for the smaller empirical effects observed for Direct Object/Sentential Complement (here the empirical effect was 10 times as large as the predicted one), Agreement Violation (7 times), or High Attachment (5 times).

Language model surprisal does not accurately predict the variation across items. We next evaluated to what extent surprisal accounts for the item-wise variation in our effects of interest. Here, we assessed whether the models predicted the correct rank ordering of items within each condition — in other words, whether they predicted greater processing difficulty for those items where humans showed longer reading times.

The results of this analysis are summarized in Fig. 5, which plots the amount of item-wise variation in the effect captured by our models against the maximum amount of explainable variance (i.e. the Spearman-Brown-corrected split-half reliability for that construction); for full visualization of item correlations between predicted and empirical EOIs, see the GitHub repository.

For filler items, the proportion of variance explained was relatively high, consistent with the reading time corpus findings reported by Schrimpf et al. (2021). This was the case for all three models, however: the models that included surprisal did not substantially outperform the No-surprisal baseline. This indicates that for the filler sentences, where lexical variables are not controlled, much of the variance in RTs can already be explained by such lexical factors; this makes these sentences less useful for evaluating the extent to which surprisal can explain processing difficulty (cf. Marvin & Linzen, 2018).

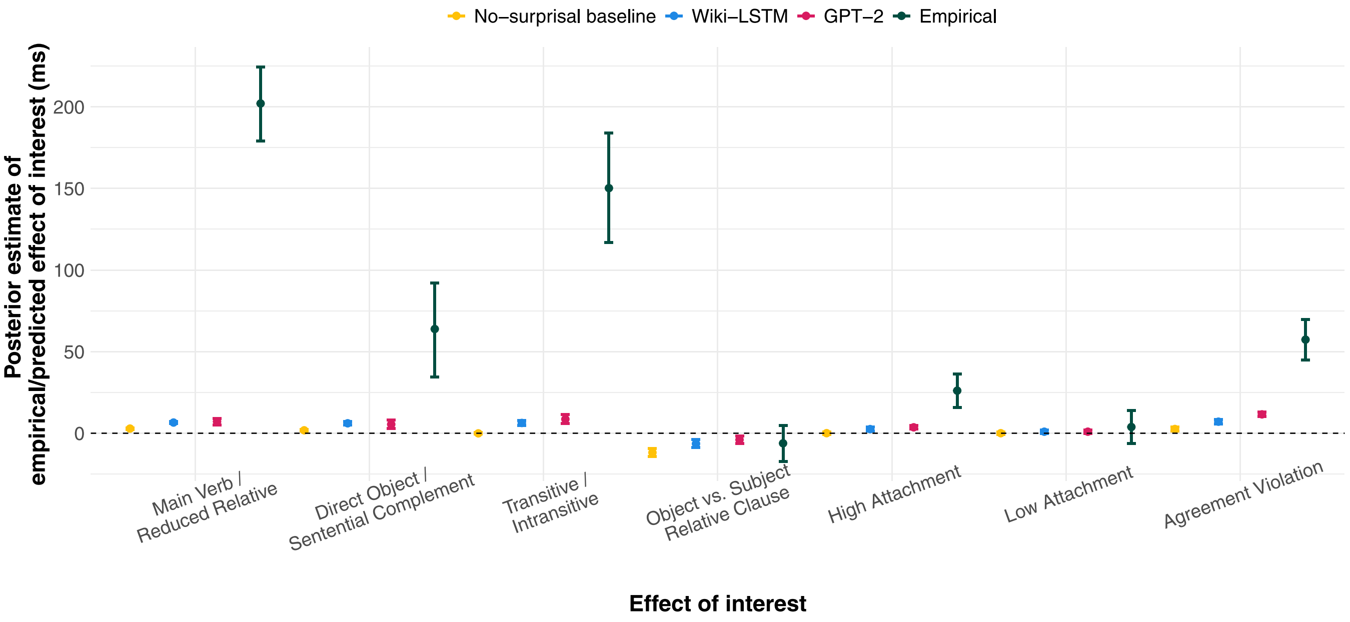


Fig. 4. Empirical and predicted effects of interest at the first spillover region for all seven constructions in the SAP Benchmark . Empirical effects were estimated from a Bayesian mixed-effects regression model fit to raw RTs on the word that indexed the effect of interest. Error bars represent the 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 fit to the predictions of the language models and the No-surprisal baseline model.

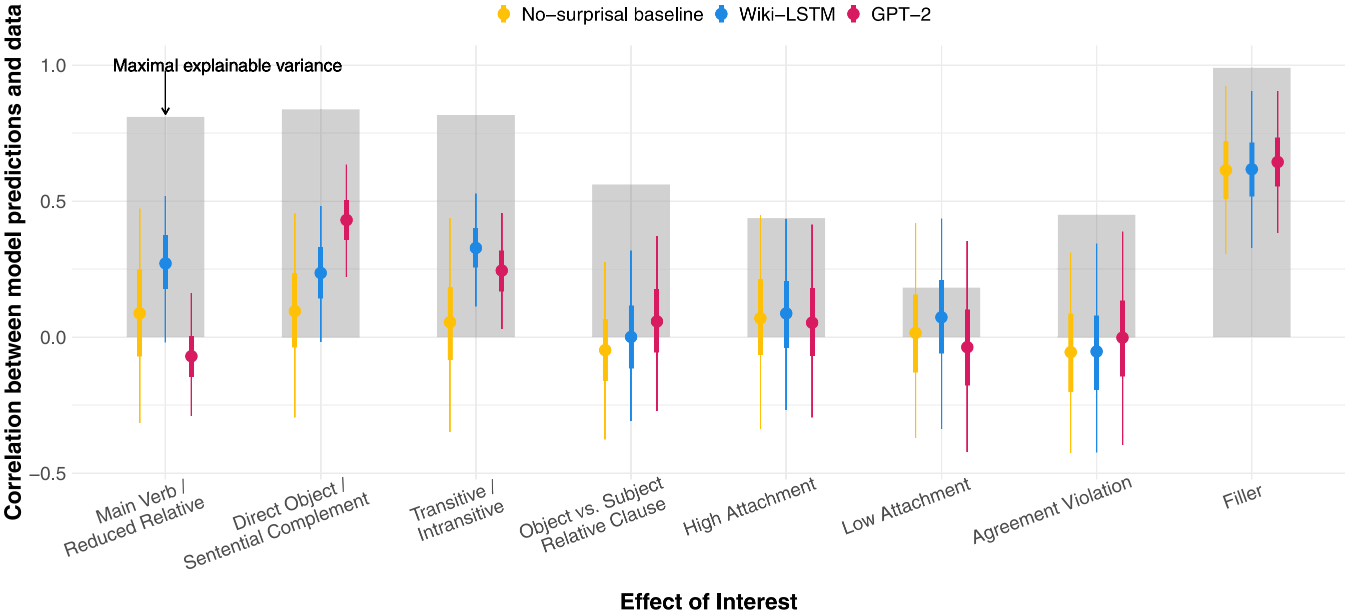


Fig. 5. Correlation between the item-level predicted effects of interest and item-level empirical effects of interest. Error bars represent 95% credible intervals, and the gray bars indicate split-half reliability, which quantifies the maximal amount of explainable variance for each construction.

By contrast, for the classic garden path subsets, the models accounted for less than half of the explainable item-wise variation. For these constructions, however, models incorporating surprisal generally explained more of the item-wise variance than the No-surprisal baseline.

For Object vs. Subject Relative Clause, High Attachment, and Agreement Violation, very little of the explainable variance was accounted for by the models. For Low Attachment, almost half of the explainable variance was predicted by Wiki-LSTM; however, this finding needs to be interpreted in the context of the fact that the explainable variance in this construction was very low, and that this contrast had no reliable effect at the construction level.

These negative results below are most likely not due to insufficient variation within the model-based surprisal estimates: for the constructions that displayed robust EOIs, the overall amount of item-wise variation in the model-based effects was comparable to the item-wise variation in the empirical effects. Specifically, the V_2 coefficient of variation ranged from 0.16 to 0.69 for Wiki-LSTM and from 0.16 to

0.58 for GPT-2, compared to 0.17 to 0.52 for the empirical effects (see Fig. 3).

General Discussion

Prediction has been proposed as an organizing principle of human cognition in general and language in particular (Dell et al., 2021; Pickering & Garrod, 2013). In machine learning, deep-learning language models trained to predict upcoming words—or, more generally, some aspect of their training corpus from another aspect ("self-supervised learning")—have been immensely successful as a foundation for language technologies (Devlin, Chang, Lee, & Toutanova, 2019; Peters et al., 2018), and have been shown to learn a surprising amount about language structure (Linzen & 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 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.

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 several folds. Second, the models incorrectly predicted similar levels of processing difficulty in three different garden paths that showed very different empirical patterns of difficulty. Third, the models had only limited success in explaining item-wise variation in processing difficulty. Among the seven constructions we tested, language model surprisal performed best in the Direct Object/Sentential Complement construction, accounting for slightly over half of the explainable variance across items; for other constructions, language model surprisal did not capture any inter-item variation above and beyond a baseline model that did not include surprisal at all.

Our first finding confirms recent reports of misalignments between the empirical and predicted effect sizes (van Schijndel & Linzen, 2021; Wilcox et al., 2021). In our case, the misalignment is particularly striking as our empirical effect sizes were larger than those reported in most earlier studies (Dempsey, Liu, & Christianson, 2020; Grodner et al., 2003; Traxler, 2005). We hypothesize that this is due to the fact that each participant in our experiment only read four sentences of any particular construction, intermixed with 48 sentences of 12 other constructions plus 40 filler sentences. We kept the number of observations per condition low to minimize syntactic adaptation (Fine et al., 2013; Prasad & Linzen, 2021); of course, while there still might have been some syntactic adaptation in our experiment, our dataset still likely offers measures of processing difficulty that more closely approximate the cost of syntactic disambiguation when a sentence appears outside of an experimental context, and hence is more suitable for addressing our theoretical questions about predictive processing.

Implications for theories of sentence processing

How much of human sentence processing difficulty can ultimately be reduced to rational prediction, that is, prediction that is well-calibrated to the probability of word sequences in natural language? Our results cast doubt on the strongest thesis that localized language processing difficulty can be wholly reduced to word-by-word predictability from a model optimized to predict the next word (Levy, 2008; Schrimpf et al., 2021). This is the case even for garden path constructions where difficulty is plausibily driven by syntactically unpredictable sentence completions, and as such which would appear to be excellent candidates for a predictability-based account, to the extent that predictability estimates are sensitive to syntax (Hale, 2001).

Would our results generalize to stronger language models? While we have established the insufficiency of surprisal only for the two specific language models we tested here – and in general, our methodology can only be used to test the predictivity of surprisal computed from a specific language model – we hypothesize that our conclusion would generalize to stronger language models, as long as those are trained solely on a word prediction objective (here, by stronger we refer to models with lower perplexity, which are better able to predict the next word). First, both models failed in essentially similar ways on our contrasts, despite significant differences in architecture (LSTM vs. transformer) and training data (Wikipedia vs. books and web pages). Second, although both our models were trained to optimize word-by-word perplexity over datasets that match or exceed the linguistic experience of a human's lifetime, even larger 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 & Schuler, 2023; Shain et al., 2022), reversing an earlier trend observed with weaker models (Goodkind & Bicknell, 2018; Wilcox et al., 2020). While those studies did not evaluate the models' fit to the syntactic constructions we focus on in this work, we take their results to indirectly suggest 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.

The role of prediction in human sentence processing. To be sure, our results do not license the stronger conclusion that prediction plays no role at all in language comprehension; 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; that the role of predictability in moment-by-moment processing difficulty is more modest than often assumed, or both.

Proponents of the first hypothesis – that prediction does have an overarching explanatory power, but human word predictions systematically differ from the optimal predictions embodied by powerful language models – may aim to create language models whose predictions align more closely with those made by humans (Eisape, Zaslavsky, & Levy, 2020). As a recent example, Arehalli, Dillon, and Linzen (2022) reweighted language models' predictability estimates to emphasize syntactic predictions more strongly than purely lexical ones, and found that doing so indeed brought model estimates of garden path effects somewhat closer to the empirical effects. At the same time, the resulting estimates were still dramatically smaller than observed effect sizes, supporting the hypothesis that there are additional factors at play other than predictability.

An alternative approach, based on the second hypothesis, highlights the role of mechanisms other than word prediction. One candidate for such a mechanism is reanalysis in serial parsing (or parsing with limited parallelism). The language models we tested can, at least in principle, represent all possible analyses of the sentence in their hidden state (Aina & Linzen, 2021). This maps onto the fully parallel parsing assumption that tends to underlie "one-stage" models of human sentence processing, 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 this assumption is incorrect: Readers do not, in fact, consider most or all possible analyses of the sentence; instead, because of memory limitations on the number of interpretations of a sentence that they can entertain concurrently, when one of the grammatically possible interpretations is deemed unlikely, that interpretation drops out of consideration (Frazier, 1979; Gibson, 1991; Jurafsky, 1996). At the disambiguating region, when the initially 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 (or reread parts of the sentence, an option that is not available to them in self-paced reading). Models like this are broadly referred to as "two-stage" models of sentence processing (Van Gompel & Pickering, 2007).

The reanalysis process posited by two-stage models could explain why garden paths require much longer to process than predicted by surprisal. Neural network language models could approximate serial parsing using particle filters (Levy, Reali, & Griffiths, 2008) or beam search over parses, for models such as Recurrent Neural Network Grammars that represent symbolic parses explicitly (Dyer, Kuncoro, Ballesteros, & Smith, 2016; Hale, Dyer, Kuncoro, & Brennan, 2018). That would not be sufficient, however: any such two-stage model would also need to specify how reanalysis proceeds once the discarded parse needs to be reconstructed. Furthermore, the fact that in our experiment surprisal correlated only modestly with disambiguation difficulty suggests that the syntactic expectations generated by each construction and item, or the cost of reanalysis that corresponds to reconstructing the discarded parse, are driven by structural or contextual factors that are not captured by neural network language models (Frazier & Clifton, 1998; Sturt et al., 1999).

Existing models of reanalysis differ in whether this process involves merely reprocessing the string (Grodner et al., 2003), or whether comprehenders make use of special repair processes to modify and update the existing representation of a linguistic structure in memory (Ferreira & Henderson, 1998; Lewis, 1998; Van Dyke & Lewis, 2003). Research into such specialized repair mechanisms has indicated many factors that modulate the ease of reanalysis, such as the amount of thematic revision involved (Ferreira & Henderson, 1998), the amount of existing syntactic structure that can be preserved (Sturt, 1997; Sturt et al., 1999), the diagnosticity of the disambiguating cue (Fodor & Ferreira, 1998; Van Dyke & Lewis, 2003), syntactic locality (Lewis, 1998; Weinberg, 1998), and more. While our present results do not adjudicate between these various proposals, they do provide an empirical benchmark that could support future work towards quantitatively explicit models of syntactic reanalysis.

Attachment preferences support two-stage accounts. Other aspects of our results are also consistent with our conjecture that limited beam parsers are best suited to capture our results. Within the relative clause attachment subset of the experiment, we observed processing difficulty for the High Attachment condition, but not the Low Attachment condition. This is the pattern predicted by two-stage models that posit a parsing strategy whereby the relative clause is attached to the most recent noun (Frazier, 1979). If readers follow this strategy, they will not be garden-pathed when that analysis turns out to be the ultimately correct one, leading to little measurable processing difficulty for Low Attachment, as we observe. By contrast, single-stage prediction models predict an ambiguity advantage effect, where both low and high attachment of the relative cause processing difficulty compared to a globally ambiguous baseline; in fact, this is the pattern observed in some previous work (Traxler et al., 1998; Van Gompel et al., 2005). Thus our findings from the relative clause attachment subset are broadly more consistent with two-stage models.

Why do the results we found for the relative clause attachment subset 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, the ambiguity advantage may only emerge when the task permits shallow processing, such that comprehenders do not need to fully resolve the structure of the input (Logačev & Vasishth, 2016; Swets et al., 2008). If this is correct, then our finding of a penalty for High Attachment, but not for Low Attachment, may suggest that our participants were engaged in deeper processing than in previous studies, which pushed them to commit more strongly to a small number of analyses of the input. This raises the possibility that the misalignment between surprisal and empirical reading times that we observed in this study—in the relative clause attachment subset and elsewhere—arises primarily when readers engage in deeper processing of the sentence, for example when comprehension questions are challenging and require readers to construct an accurate representation of the structure of the sentence, a process which may require costly reanalysis.

Lossy-context surprisal. In this work, we have evaluated the classic version of surprisal theory, which assumes that readers' next-word predictions are based on an accurate representation of the context. Modifications to have been proposed to this theory that relax this assumption, allowing for the possibility that the context is encoded or maintained imperfectly, and hence the probability of an upcoming word is conditioned on a perceiver's subjective encoding of the context, rather than the true context; moreover, readers might update or modify their subjective encoding of the context in view of the current input (Futrell, Gibson, & Levy, 2020; Levy, 2013). Such a noisy channel account could explain why the agreement mismatch costs in our study were much smaller than those induced by the unlikely but grammatical continuations in garden path sentences, despite the fact that agreement mismatch constitutes an unresolvable clash between the features of the subject and the verb, and as such we would expect the ungrammatical verb to be assigned very high surprisal: the participants might have attributed the perceived ungrammaticality to their own memory error or to a typographical error. If that was the case, syntactic reanalysis would not be required for sentence with agreement errors, unlike garden path sentences.

At the same time, we believe that lossy-context models are unlikely to improve the fit to the empirical data in the classic garden path subset. If anything, lossy-context surprisal should predict even smaller garden path effects than classical surprisal; for example, when readers reach the word *remained* in *When the little girl attacked the lamb remained*, lossy-context surprisal may predict that the reader would incorporate a comma into their mental representation of the context to make sense of the low-probability continuation *remained*; this mentally inserted comma should reduce processing difficulty. This reasoning can be tested against our benchmark using computationally implementations of the lossy-context surprisal framework (Hahn, Futrell, Levy, & Gibson, 2022).

Surprisal-based vs. embedding-based linking functions

The quantitative misalignments we have observed in our analyses 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 (Caucheteux, Gramfort, & King, 2023; Goldstein et al., 2022; Schrimpf et al., 2021). The success of such 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 than to our critical items. 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 & 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 relating bits of surprisal to reading times: one for the current word, and three for the previous words, to account for spillover. By contrast, in the neuroimaging studies mentioned above, an encoding model – typically, a dense linear layer (linear regression) – 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 from systems trained to perform tasks that are not directly related to English next-word prediction, such as English to German translation (Antonello & 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 (see Hale et al., 2022 for similar arguments).

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 items. This is a higher bar for the models than the one used in the neuroimaging studies mentioned above, 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.

Relating surprisal to reaction times

The statistical analyses we reported in this paper used raw, untransformed reading times as the dependent measure. As we discussed in Section "Analyzing raw RTs vs. log-transformed RTs", this modeling decision reflects the theoretical commitments of surprisal theory (Levy, 2008; Smith & Levy, 2013; Wilcox et al., 2023), but may lead to violations of the statistical assumptions of the regression models. To evaluate the consequences of this decision for our inferences about the magnitude of effects we measured in our study, we fitted exploratory models with a log-normal link function (which is equivalent to a linear regression analysis of the log-transformed reading times). The full results of this exploratory analysis are available in Appendix A and on the Github repository associated with this project. Overall, the log-normal link function yielded more conservative estimates of the garden path effect size: the estimates for the construction-level effects were approximately half as large as those estimated with the normal link function.

Our qualitative conclusions hold even under the more conservative effect size estimates, obtained using the log-normal link. Whereas under the Gaussian link, the estimated empirical ambiguity effects were between 5 and 31 times as large as the predicted ones, under the log-normal link the estimated empirical effects were still between 3 and 27 times the size of the predicted effects. Replacing the normal link with the log-normal link also had little impact on the item-wise correlations (Fig. A.3).

The SAP Benchmark as a tool for theory evaluation

Stepping back from theoretical issues raised by our analyses, the SAP Benchmark more generally 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 phenomena, such as garden path constructions and relative clauses, which have long been key tests of qualitative theories of sentence processing. The set of phenomena we chose is of course not exhaustive, and there are important precedents to this work that have attempted to systematize large catalogs of sentence processing phenomena in the service of theory building. An interesting project for future work would be to collect benchmark data on such wider-ranging catalogs of garden path (Lewis, 1993) or syntactic complexity (Cowper, 1976) phenomena. 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 Direct Object/Sentential Complement, 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.

Conclusion

In this study, we have presented the SAP Benchmark, a self-paced reading dataset collected from 2000 participants, which covers a range of syntactic phenomena. We have used this dataset to test the hypothesis that syntactic processing difficulty can be explained using the word surprisal estimated from neural network language models. We found only modest support for this hypothesis, with three 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 failed to predict the large empirical differences across constructions in the magnitude of the garden path effect. Finally, model-based surprisal showed only limited success at capturing the variation in processing difficulty across items within each construction. 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"Implications for theories of sentence processing").

Beyond the specific theoretical questions we addressed, the SAP Benchmark 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 ones (Grodner & Gibson, 2005); disambiguation in favor of an unexpected parse of a structurally ambiguous sentence causes processing difficulty (Frazier & Rayner, 1982); and subject-verb agreement errors are detected quickly and cause a slowdown in reading (Pearlmutter, Garnsey, & Bock, 1999; Wagers et al., 2009).

We not only presented a high-powered replication of classic results, but also expanded the empirical picture by using an experimental design that allowed us to directly compare reading times across constructions and items. We observed that the Transitive/Intransitive garden path effect is about twice as large as the Direct Object/Sentential Complement ones, confirming the results of earlier studies (Sturt et al., 1999) with much more precise effect size estimates. We also showed, for the first time in a controlled design, that Main Verb/Reduced Relative garden paths are more difficult than either Transitive/Intransitive or Direct Object/Sentential Complement. We also documented significant differences across constructions in the extent of variation across items, with some constructions showing rather consistent effect sizes across items, and others showing dramatic variability. Overall, quite aside from the debate on the adequacy of surprisal as an explanation for syntactic processing difficulty, we hope that the quantitative effect size estimates produced by the SAP Benchmark will serve as a modeling target for any computational model of human sentence comprehension.

CRediT authorship contribution statement

Kuan-Jung Huang: Formal analysis, Investigation, Project administration, Software, Validation, Visualization, Writing – review & editing, Methodology. **Suhas Arehalli:** Formal analysis, Methodology, Software, Validation, Visualization, Writing – review & editing. **Mari Kugemoto:** Formal analysis, Methodology, Software, Validation, Writing – review & editing. **Christian Muxica:** Investigation, Methodology, Software, Visualization. **Grusha Prasad:** Formal analysis, Methodology, Software, Validation, Visualization, Writing – review & editing. **Brian Dillon:** Conceptualization, Formal analysis, Funding acquisition, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing, Project administration, Supervision. **Tal Linzen:** Conceptualization, Formal analysis, Funding acquisition, Methodology, Project administration, Validation, Writing – original draft, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that there is no any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

Data availability

The materials, reading time data, and analysis scripts are available at the following website:

Syntactic Ambiguity Processing Benchmark (Original data) (GitHub)
.

Acknowledgments

This work was supported by the National Science Foundation, grant nos. BCS-2020945 and BCS-2020914, and in part through the NYU and UMass IT High Performance Computing resources, services, and staff expertise. 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. We also thank John Hale, Florian Jaeger, Roger, Levy, Bonnie Nozari, and two anonymous reviewers for helpful comments on a previous version of this manuscript.

Appendix A. Estimating EOIs with different link functions

Because reading times typically follow a right-skewed distribution, fitting a linear regression model with a Gaussian link function to RTs may violate several assumptions of such a model, in particular the assumption of normally-distributed residuals. The practical impact of this for our conclusions is unclear, however. For example, recent simulations suggest that violations of the normality assumption may lead to very little bias in regression coefficient estimates, and no change in power for $n >= 1000$ (Knief & Forstmeier, 2021). Still, Knief and Forstmeier caution that this optimistic conclusion may not hold in situations where homogeneity of variance cannot be assumed across levels of the predictor variable.

Table A.1
The priors for Bayesian mixed-effects models with log-transformed RTs.

Class	Prior on log ms	Prior expressed in ms
Intercept	Normal(5.7, 1.5)	Normal(300, 1000)
Coefficients	Normal(0, 1)	Normal(0, 150)
Standard deviation (random effects)	Normal(0, 1.5)	Normal(0, 200)
Standard deviation (residuals)	Normal(0, 2)	Normal(0, 500)

To evaluate the consequence of our modeling decision for our results, we fitted a number of exploratory models with a log-normal link function, which has been argued to provide a good model of reading times or other reaction times (Frank et al., 2013; Paape & Vasishth, 2022). We log-transformed the raw RTs and ran regression models parallel to those in the main text. For models fit to the filler sentences, we removed the interactions between frequency and length; this was done because, as shown below, when the regression equation is transformed back from log RTs to raw RTs, the interaction term is no longer the product of the frequency and the length, as it typically is in a regression model, but the *exponentiation* of one predictor by the other.

$$
\begin{aligned}
log(RT(w_n \mid context)) = {} & \beta_0 + \beta_1 \cdot Surp + \beta_2 \cdot WordPosition \\
& + \beta_3 \cdot Freq + \beta_4 \cdot Length + +\beta_5 \cdot Freq \cdot Length + \epsilon
\end{aligned}
$$

$$
\begin{aligned}
RT(w_n \mid context) &= e^{\beta_0+\beta_1 \cdot Surp+\beta_2 \cdot WordPosition+\beta_3 \cdot Freq+\beta_4 \cdot Length+\beta_5 \cdot Freq \cdot Length+\epsilon} \\
&= e^{\beta_0} \cdot e^{\beta_1 \cdot Surp} \cdot e^{\beta_2 \cdot WordPosition} \\
&\quad \cdot e^{\beta_3 \cdot Freq} \cdot e^{\beta_4 \cdot Length} \cdot e^{\beta_5 \cdot Freq \cdot Length} \cdot e^{\epsilon} \\
&= e^{\beta_0} \cdot {e^{\beta_1}}^{Surp} \cdot {e^{\beta_2}}^{WordPosition} \\
&\quad \cdot {e^{\beta_3}}^{Freq} \cdot {e^{\beta_4}}^{Length} \cdot {e^{\beta_5}}^{{Freq}^{Length}} \cdot e^{\epsilon}
\end{aligned}
$$

We set the priors to be consistent with those in our raw RT analysis (see Table A.1). We then back-transformed the results to raw RTs for each experimental condition; the effects of interest (EOI)s were extracted from these back-transformed estimates.

The log-normal link function generally did not qualitatively influence the results on our critical EOIs. All EOIs whose 95% credible intervals excluded zero in the raw RT analysis continue to do so in the log-normal analysis. In the log-normal analysis the 95% credible intervals excluded 0 in two contrasts that did not exclude 0 in the raw RT analysis: the word following the critical word in Low Attachment, and the critical word in Object vs. Subject Relative Clause.

Quantitatively, the log-normal link function had a large impact on the estimates of the empirical ambiguity effects, almost halving the estimated effect size for all EOIs compared to the normal link function (Fig. A.1). This pattern raises the possibility that the analyses with raw RTs (a Gaussian link function) could have overestimated the effect sizes in our dataset. Since one of our key observations is that the observed size of our effects of interest is much larger than predicted by language model surprisal, it is crucial to determine if these more conservative estimates of the empirical effect sizes still support our theoretical conclusions.

To do this, we estimated conversion factors from the log-transformed reading times on the filler items, and used these to estimate EOIs of interest using the same methodology outlined in the main text. The bottom panel of Fig. A.2 shows the results of this analysis. While the estimated empirical EOI sizes are much smaller in the log RT analysis, the gaps between those estimates and the predicted EOIs still remain large. Across contrasts, they ranged from a 3-fold difference to a 27-fold

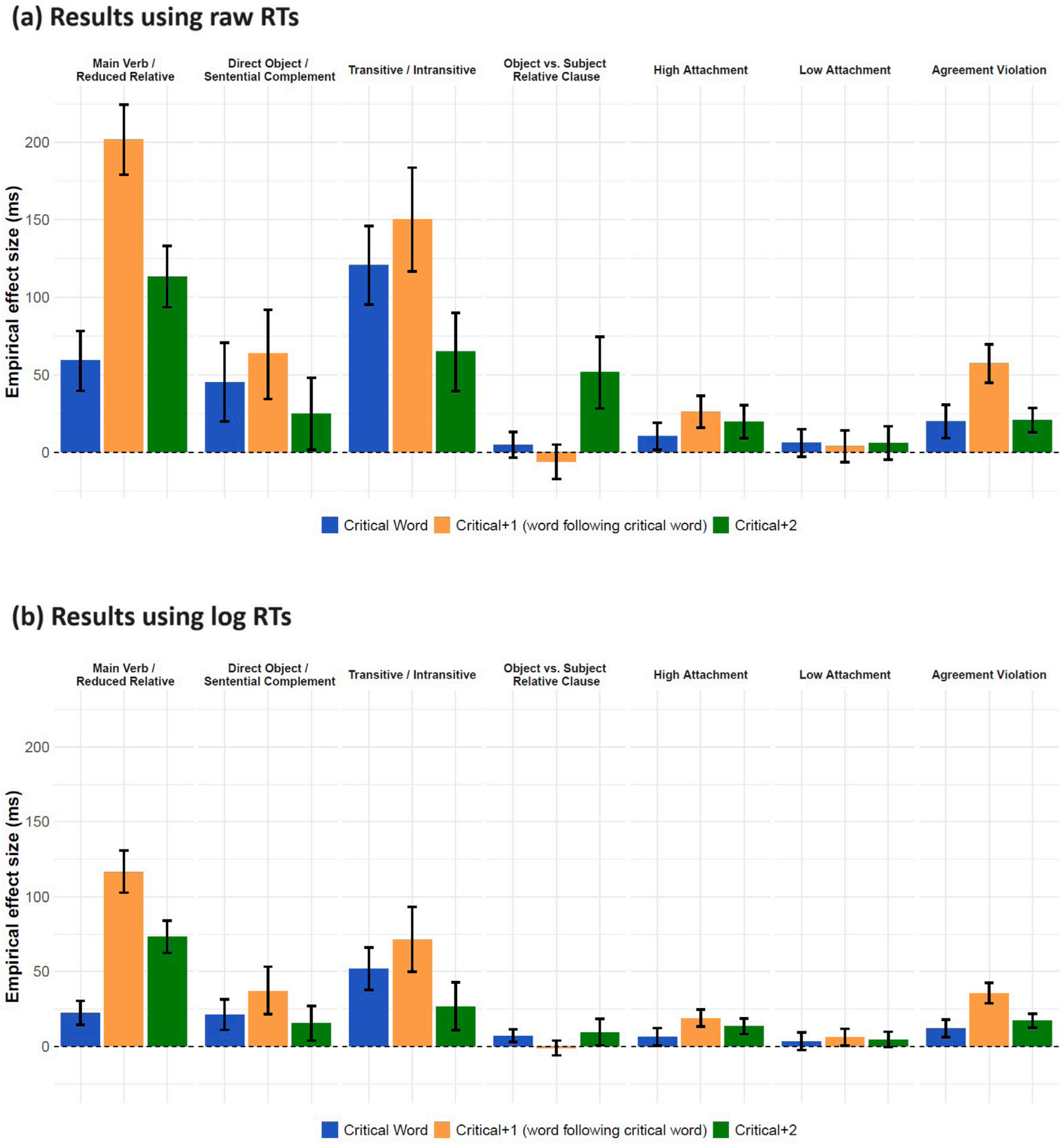


Fig. A.1. Estimated empirical EOIs using raw RTs (top panel, same plot as Fig. 2 in the main text) and using log RTs (bottom panel).

difference across EOIs and across the two language models, compared to a 5- to 31-fold difference when analyzing raw RTs.

Log-transformation did not qualitatively influence the results of the item-wise correlation analyses either (Fig. A.3). While the correlations overall became higher in this analysis, so did the maximal explainable variance. Direct Object/Sentential Complement remains the only EOI for which more than half of the item-wise variance can be explained by surprisal, consistent with the raw RT results reported in the main text.

Overall, we conclude that these exploratory analyses show that the results reported in the main text hold even with a link function that yields very conservative estimates of syntactic ambiguity effects.

Appendix B. 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,

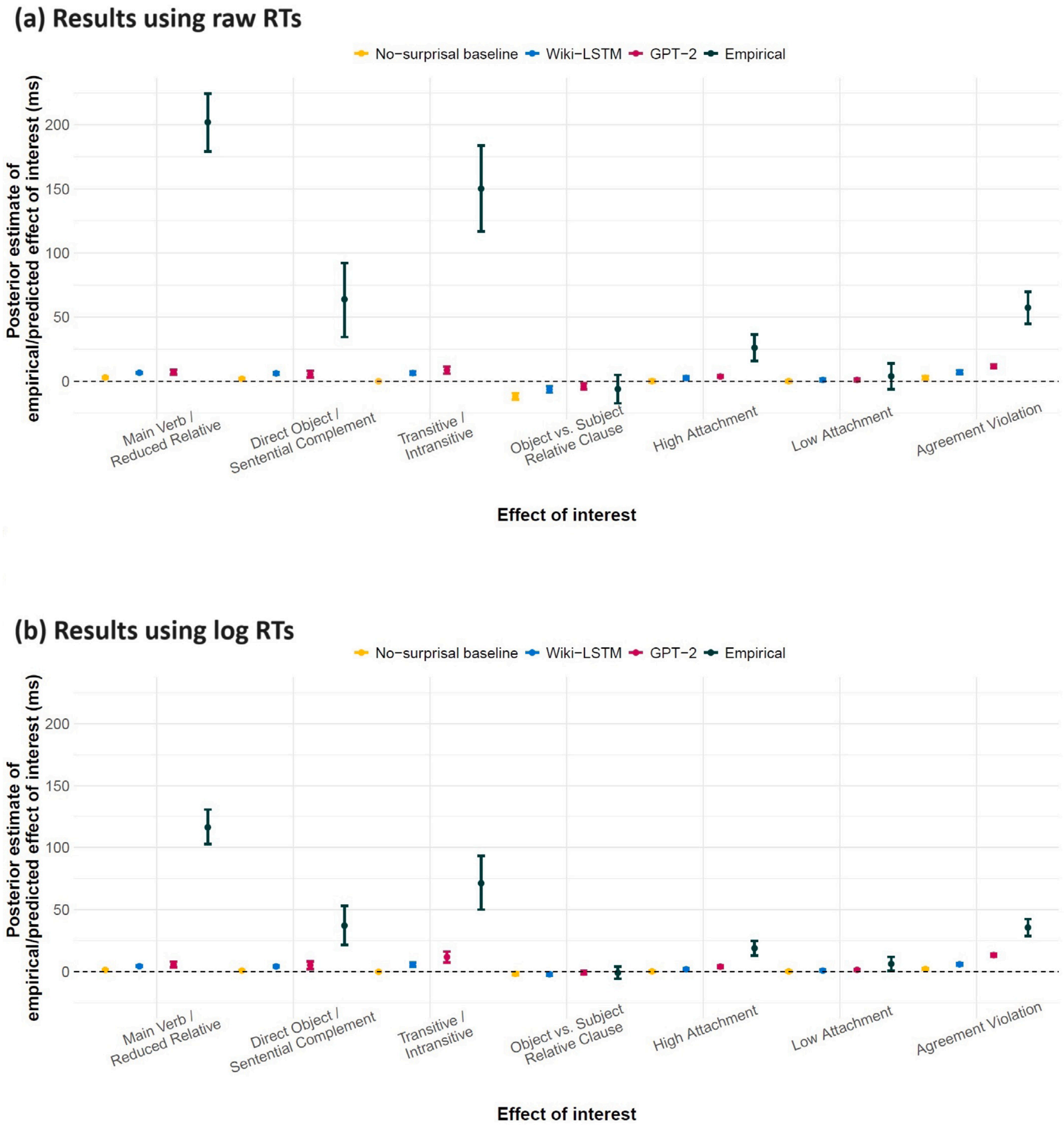
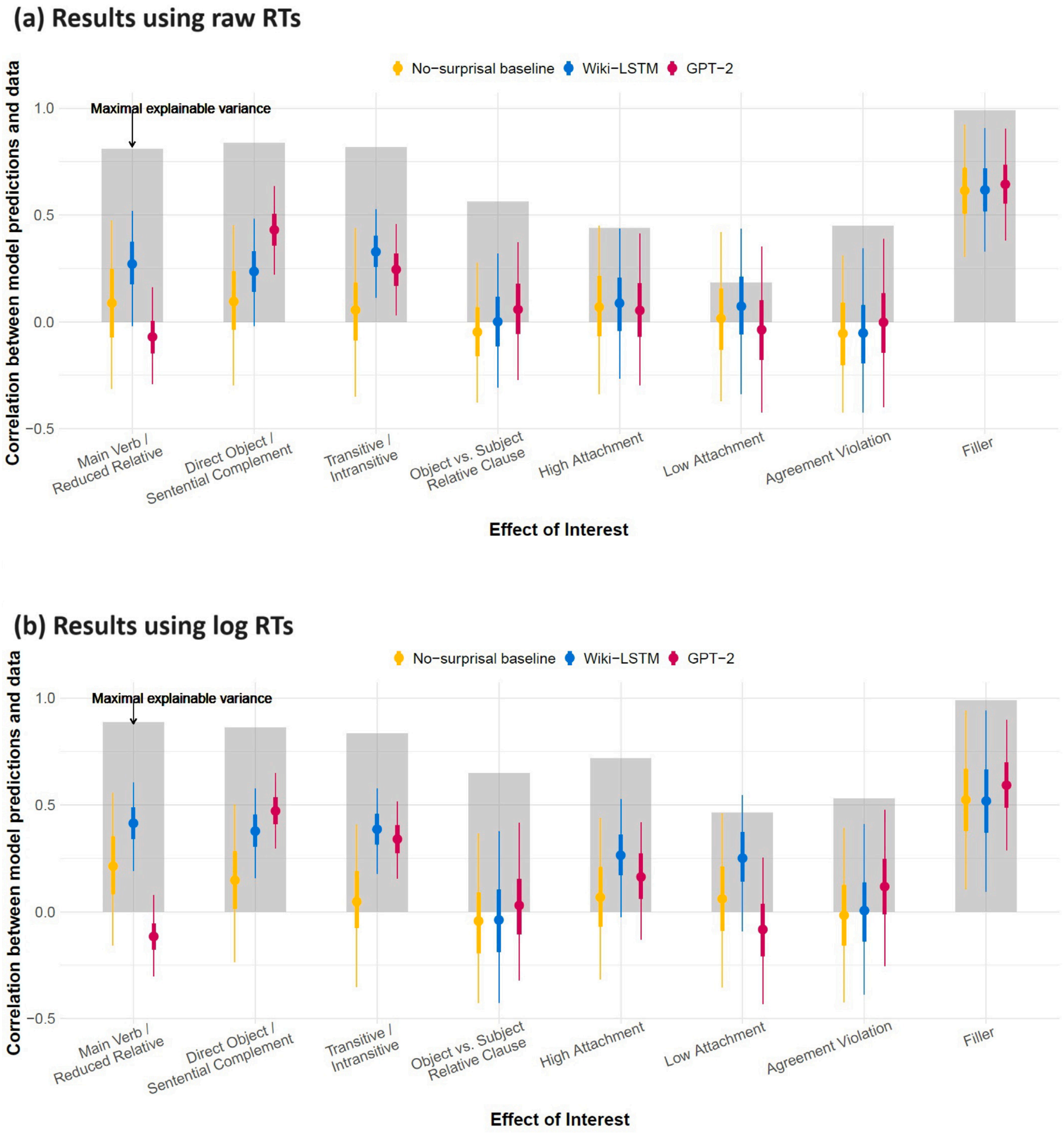


Fig. A.2. Empirical and predicted effects of interest at the first spillover region for all seven constructions in the SAP Benchmark using raw RTs (top panel, same plot as Fig. 4 in the main text) and log RTs (bottom panel).

the comprehension questions were designed to specifically target the resolution of the ambiguity; for example, for the sentence *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 B.2 reports mean accuracy for each construction separately for questions that targeted ambiguity resolution and those that did not. As with the questions about the the fillers, questions on the critical items that did not target ambiguity resolution were answered with high accuracy across the board (82.2–96.4%). For questions that did target ambiguity resolution, accuracy varied across constructions. Accuracy was fairly high for Direct Object/Sentential Complement, Object vs. Subject Relative Clause, Agreement Violation, and High Attachment, ranging from 72.9% to 87.3%. For Transitive/Intransitive and Main


(a) Results using raw RTs

(b) Results using log RTs

Fig. A.3. Item-wise correlations between empirical and predicted EOIs using raw RTs (top panel, same plot as Fig. 5 in the main text) and log RTs (bottom panel).

Verb/Reduced Relative, by contrast, accuracy was extremely low for the conditions where there was a local ambiguity (37% and 44.1% respectively). The very low accuracy associated with these two constructions is consistent with early findings (Christianson et al., 2001; Prasad & Linzen, 2021). For Transitive/Intransitive, accuracy was low even when the sentences were unambiguous (62.7%). Finally, accuracy for Low Attachment was only 55.2%; this is similar to the rate at which participants judged these sentence as acceptable in the acceptability judgment experiment reported by Dillon et al. (2019).

Appendix C. 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 reduced for verbs that are more likely to take a sentential complement. This is consistent with the hypothesis that when readers come across such verbs, they are more likely to predict the ultimately correct sentential complement parse than the direct object parse. Garnsey et al. also showed that the strength of the garden path effect is affected by the plausibility of the direct object

Table B.2
Comprehension question accuracy for each experimental construction, for questions that targeted ambiguity resolution and questions that did not. Standard errors (by-subject) in parentheses.

Construction	Question targets ambiguity?	
	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)

reading of the ambiguous region; this is likewise consistent with the hypothesis that plausibility affects the likelihood that readers adopt the direct object parse, which ends up being incorrect.

Inspired by Garnsey and colleagues' analysis, this appendix reports analyzes that test to what degree item-wise variation in the garden path effect size can be explained using verb bias and plausibility. We use 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.

C.1. Predictors

Local phrase plausibility norms. In an online norming task ($N = 100$), we presented participants with a fragment of each of the stimuli from the garden path subset of our self-paced reading experiment. Participants were recruited through Prolific, with the same criteria we used for the self-paced reading task. 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. Table 1 in the main text 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. (Original: The little girl fed the lamb remained relatively calm...)
b. **Direct Object/Sentential Complement**: The little girl found the lamb. (Original: The little girl found the lamb remained relatively calm...)
c. **Transitive/Intransitive**: The little girl attacked the lamb. (Original: When the little girl attacked the lamb remained relatively calm.)

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 the most 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, again recruited on Prolific, 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 verb bias value of 0.70.

Corpus-based verb bias estimates. Finally, we performed a corpus analysis as described in Section "Materials", extracting from COCA sentences containing the sequence **DP VERB DP** for each of the verbs in question (e.g., DP *moved* DP). All of the results were parsed and labeled 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.

C.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 the item's verb bias is 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 (Garnsey et al., 1997; Pickering & Traxler, 1998): 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.

C.3. Analysis and results

In this section, we focus on the EOI on the word immediately following the disambiguating word (the first spillover word), where garden path effects in our main analysis were largest. We first report simple correlations among the variables and the EOIs (Table C.3). 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. We note that in Garnsey et al. (1997) and Pickering and Traxler (1998), the plausibility manipulation, which was the focus of those experiments, was stronger than in our study: in those studies the mean local phrase plausibility rating was around 6 for the high-plausibility condition compared to 2 for the low-plausibility condition, whereas in our study local plausibility ratings only ranged between 4 to 7. As such, our inability to observe a plausibility correlation could be due to the absence of items that are highly locally implausible.

In the Direct Object/Sentential Complement construction, 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 two other 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.

We also fit multiple linear regressions (Table C.4). These regressions included all three predictors described in this section, as well as *surprisal difference*, which we computed by subtracting the surprisal of the critical verb in the unambiguous or grammatical sentence from the surprisal of the critical verb in the matching ambiguous or ungrammatical sentence. These models did not consider spillover effects from any of the independent variables.

The results were consistent with those from the simple correlation tests: Only for Direct Object/Sentential Complement did cloze-based verb bias show a robust strong negative effect on the garden path effect size, and only for Main Verb/Reduced Relative did corpus-based verb bias show a robust strong positive effect. Overall, Direct Object/Sentential Complement was the only ambiguity 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 ambiguity, caution should be exercised in interpreting the regression results in this subset, for two reasons: first, the adjusted R-squared value, while significantly different from zero, is fairly low; and second, when corpus-based verb bias was added to the regression model, it yielded a few more spurious significant effects (e.g., an unexpected *negative* plausibility effect). The significant effect of corpus-based verb bias in the unexpected direction appears to have been driven by two items. For both of these items, the critical ambiguous verb was *fed*, which has a relatively high corpus-based verb bias. A detailed examination of the reduced relative clause uses of *fed* in the corpus showed taht these uses 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 "Comparison to language model surprisal", we conclude that item-wise variation in the magnitude of garden path effects, while substantial and reliable, cannot be readily explained by word surprisal, syntactic surprisal (cloze-based and corpus-based), or local phrase plausibility, at least not with a simple linear linking function.

Appendix D. Materials

This section lists the materials for all four subsets.

D.1. Classic garden paths

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 corrupt politician mentioned (that) the bill received unwelcome attention from southern voters.
b. After the corrupt politician signed(,) the bill received unwelcome attention from southern voters.
c. The corrupt politician (who was) handed the bill receive unwelcome attention from southern voters.

(3) 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.

(4) 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.

(5) 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.

(6) 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.

(7) 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.

(8) 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.

(9) 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.

(10) 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.

(11) 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.

(12) 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.

Table C.3
Correlations between EOI size and the predictors described in Appendix C.1, as well as among the predictors. LSTM and GPT-2: surprisal differences at the disambiguating verb, as estimated from the language models; Plausibility: local phrase plausibility; Cloze: verb subcategorization bias as normed by the cloze task; COCA: verb subcategorization bias as estimated from Corpus of Contemporary American English; RRC: reduced relative clause; SC: sentential complement.

Main Verb/Reduced Relative

	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
SC bias (Cloze)					–	0.22
SC bias (COCA)						–

Transitive/Intransitive

	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
Intransitivity bias (Cloze)					–	0.43*
Intransitivity bias (COCA)						–

* *p* < .05;
** *p* < .01;
*** *p* < .001.

Table C.4
T-values from multiple regressions predicting EOI sizes from the variables described in Appendix C.1. 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 between the ambiguous (or ungrammatical) and unambiguous (or grammatical) conditions at the critical 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 Corpus of Contemporary American English.

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

* *p* < .05;
** *p* < .01;
*** *p* < .001.

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

(13) 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.

(14) 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.

(15) 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.

(16) 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.

(17) 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.

(18) 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.

(19) 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.

(20) 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.

(21) 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.

(22) 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.

(23) 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.

(24) 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

(25) 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.

(26) 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.

(27) 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.

(28) 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.

(29) 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.

(30) 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.

(31) 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.

(32) 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.

(33) 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.

(34) 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.

(35) 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.

(36) 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.

(37) 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.

(38) 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.

(39) 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.

(40) 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.

(41) 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.

(42) 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

(43) 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.

(44) 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.

(45) 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.

(46) 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.

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

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

(49) 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.

(50) 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.

(51) 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.

(52) 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.

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

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

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

(56) 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.

(57) 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.

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

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

(60) 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.

(61) 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.

(62) 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.

(63) 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.

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

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

(66) 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

(67) 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.

(68) 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.

(69) 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.

(70) 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.

(71) 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.

(72) 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.

(73) 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.

(74) 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.

(75) 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.

(76) 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.

(77) 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.

(78) 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.

(79) 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.

(80) 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.

(81) 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.

(82) 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.

(83) 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.

(84) 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.

(85) 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.

(86) 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.

(87) 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.

(88) 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.

(89) 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.

(90) a. Despite the good press, we didn ́t really like the commanders of the soldiers who whistle very loudly and for no reason at all.
b. Despite the good press, we didn ́t really like the commanders of the soldier who whistle very loudly and for no reason at all.
c. Despite the good press, we didn ́t really like the commander of the soldiers who whistle very loudly and for no reason at all.

D.5. Fillers

(91) There are now rumblings that Apple might soon invade the smart watch space, though the company is maintaining its customary silence.
(92) A bill was drafted and introduced into Parliament several times but met with great opposition, mostly from farmers.
(93) The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity.
(94) Seeing Peter slowly advancing upon him through the air with dagger poised, he sprang upon the bulwarks to cast himself into the sea.
(95) Some months later, Michael Larson saw another opportunity to stack the odds in his favor with a dash of ingenuity.
(96) Bob Murphy, the Senior PGA Tour money leader with seven hundred thousand, says heat shouldn ́t be a factor.
(97) Greg Anderson, considered a key witness by the prosecution, vowed he would not testify when served a subpoena last week.
(98) Owls are more flexible than humans because a bird ́s head is only connected by one socket pivot.
(99) Even in the same animal, not all bites are the same.
(100) Buck did not like it, but he bore up well to the work, taking pride in it.
(101) These days, neuroscience is beginning to catch up to musicians who practice mentally.
(102) Hybrid vehicles have a halo that makes owners feel righteous and their neighbors feel guilty for not doing as much to save the planet.
(103) Binge drinking may not necessarily kill or even damage brain cells, as commonly thought, a new animal study suggests.
(104) When attacked, a skunk’s natural inclination is to turn around, lick its tail and spray a noxious scent.
(105) All that the brain has to work with are imperfect incoming electrical impulses announcing that things are happening.
(106) There often seems to be more diving in soccer than in the Summer Olympics.
(107) Susan B. Anthony spent nearly sixty years of her life devoted to the cause of social justice and equality for all.
(108) Unfortunately, for every six water bottles we use, only one makes it to the recycling bin.
(109) As in the United States, Colombian legislation requires travelers entering the country to declare cash in excess of ten thousand dollars.
(110) Stress is a risk factor for both depression and anxiety, he says.
(111) When it comes to having a lasting and fulfilling relationship, common wisdom says that feeling close to your romantic partner is paramount.
(112) Voltaire himself probably won around half a million livres, a large fortune, which he then made even larger.
(113) When preparing to check out of their hotel room, some frequent travelers pile up their used bath towels on the bathroom floor.
(114) Research showing that a tiny European river bug called the water boatman may be the loudest animal on earth.
(115) When the new world was first discovered it was found to be, like the old, well stocked with plants and animals.
(116) Police in Georgia have shut down a lemonade stand run by three girls trying to save up for a trip to a water park.
(117) An early task will be to make sure the newfound microbes were not introduced while drilling through the ice into the lake.
(118) Lady Gaga’s YouTube account was suspended Thursday.
(119) John Thornton asked little of man or nature.
(120) Proper ventilation will make a backdraft less likely.
(121) For centuries, time was measured by the position of the sun with the use of sundials.
(122) The girl’s feet were then re-wrapped even tighter than before, causing her footprint to shrink further.
(123) The astronauts used a hefty robotic arm to move the bus-size canister, stuffed with nearly three tons of packing foam.
(124) Very similar, but even more striking, is the evidence from athletic training.
(125) It was a forbidding challenge, and it says much for Winstanley’s persuasive abilities, not to mention his self-confidence.
(126) With schools still closed, cars still buried and streets still blocked by the widespread weekend snowstorm, officials are asking people to help out.
(127) Steam sterilization is limited in the types of medical waste it can treat, but is appropriate for laboratory substances contaminated with infectious organisms.
(128) From coal to cars, Chinese floods tangle supply chains worldwide.
(129) This new film marks 10 years since the death of the superstar.

References

Aina, L., & 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* (pp. 42–57). Punta Cana, Dominican Republic: Association for Computational Linguistics.

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

Arehalli, S., Dillon, B., & 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* (pp. 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 J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 279–362). New York: John Wiley and Sons.

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

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

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

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

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic.. *Psychological Review*, *113*(2), 234.

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

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

Cowper, E. A. (1976). *Constraints on sentence complexity: a model for syntactic processing* (Ph.D. thesis), Brown University.

Davies, M. (2019). The corpus of contemporary American english (COCA).

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

Dempsey, J., Liu, Q., & Christianson, K. (2020). Convergent probabilistic cues do not trigger syntactic adaptation: Evidence from self-paced reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(10).

Devlin, J., Chang, M.-W., Lee, K., & 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)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics.

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

Dyer, C., Kuncoro, A., Ballesteros, M., & Smith, N. A. (2016). Recurrent neural network grammars. In K. Knight, A. Nenkova, & O. Rambow (Eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 199–209). San Diego, California: Association for Computational Linguistics, https://aclanthology.org/N16-1024.

Ehrlich, S. F., & 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., & 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* (pp. 609–619). Online: Association for Computational Linguistics.

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

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

Ferreira, F., & Henderson, J. (1998). Syntactic reanalysis, thematic processing, and sentence comprehension. In J. Fodor, & F. Ferreira (Eds.), *Studies in theoretical psycholinguistics*, *Reanalysis in sentence processing* (pp. 73–100). Dordrecht: Springer.

Fine, A. B., Jaeger, F. T., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS One*, *8*(10).

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

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

Frank, S., & 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* (pp. 337–343). Cognitive Science Society.

Frazier, L. (1979). *On comprehending sentences: syntactic parsing strategies* (Ph.D. thesis), University of Connecticut.

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

Frazier, L., & 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., & Levy, R. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, *44*, Article e12814.

Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S. T., et al. (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., & 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. A. F. (1991). *A computational theory of human linguistic processing: memory limitations and processing breakdown* (Ph.D. thesis), Carnegie Mellon University.

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

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

Goodkind, A., & 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)* (pp. 10–18). Salt Lake City, Utah: Association for Computational Linguistics.

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

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

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & 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)* (pp. 1195–1205). New Orleans, Louisiana: Association for Computational Linguistics.

Hahn, M., Futrell, R., Levy, R., & 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., Campanelli, L., Li, J., Bhattasali, S., Pallier, C., & Brennan, J. R. (2022). Neurocomputational models of language processing. *Annual Review of Linguistics*, *8*, 427–446.

Hale, J., Dyer, C., Kuncoro, A., & 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)* (pp. 2727–2736). Melbourne, Australia: Association for Computational Linguistics.

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

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

Hoover, J., Sonderegger, M., Piantadosi, S., & O'Donnell, T. J. (2023). The plausibility of sampling as an algorithmic theory of sentence processing. *Open Mind*, *7*, 350–391.

Hu, J., Gauthier, J., Qian, P., Wilcox, E., & 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* (pp. 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., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, *111*(2), 228.

King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, *30*(5), 580–602.

Knief, U., & 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., & 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* (pp. 190–207).

Kvålseth, T. O. (2017). Coefficient of variation: the second-order alternative. *Journal of Applied Statistics*, *44*(3), 402–415.

Lau, E., & 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 R. P. G. van Gompel (Ed.), *Sentence processing* (pp. 78–114). London and New York: Psychology Press.

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

Levy, R., Reali, F., & Griffiths, T. (2008). Modeling the effects of memory on human online sentence processing with particle filters. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems*: *vol. 21*, Curran Associates, Inc..

Lewis, R. L. (1993). *An architecturally-based theory of human sentence comprehension* (Ph.D. thesis), Carnegie Mellon University.

Lewis, R. (1998). Reanalysis and limited repair parsing: Leaping off the garden path. In J. Fodor, & F. Ferreira (Eds.), *Studies in theoretical psycholinguistics*, *Reanalysis in sentence processing* (pp. 247–286). Dordrecht: Springer.

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

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

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

Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 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 D. E. Kieras, & M. A. Just (Eds.), *New methods in reading comprehension research* (pp. 69–89). Hillsdale, NJ: Erlbaum.

Nalborczyk, L., Batailler, C., Lœ venbruck, H., Vilain, A., & 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., et al. (2018). Benchmarks for models of short-term and working memory.. *Psychological Bulletin*, *144*(9), 885.

Oh, B.-D., & 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., & 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., & 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., et al. (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)* (pp. 2227–2237). New Orleans, Louisiana: Association for Computational Linguistics.

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

Pickering, M., & Traxler, M. (1998). Plausibility and recovery from garden paths: An eye-tracking study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(4).

Prasad, G., & 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.

Roberts, S., & Sternberg, S. (1993). The meaning of additive reaction-time effects: Tests of three alternatives. In D. E. Meyer, & S. Kornblum (Eds.), *Attention and performance XIV: synergies in experimental psychology, artificial intelligence, and cognitive neuroscience* (pp. 611–653). The MIT Press.

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

Shain, C. (2023). Word frequency and predictability dissociate in naturalistic reading.

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

Smith, N. J., & 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. (1997). Syntactic reanalysis in human language processing.

Sturt, P., Pickering, M. J., & 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., & 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. (2005). Plausibility and verb subcategorization in temporarily ambiguous sentences: Evidence from self-paced reading. *Journal of Psycholinguistic Research*, *34*(1).

Traxler, M. J., Morris, R. K., & 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., & Clifton, 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., & 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., & Pickering, M. J. (2007). Syntactic parsing. In *The oxford handbook of psycholinguistics* (pp. 289–307). Oxford University Press Oxford.

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

van Schijndel, M., & Linzen, T. (2021). Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, *45*(6), Article e12988.

Vani, P., Wilcox, E., & 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* (pp. 1528–1534). Online: Cognitive Science Society, URL https://escholarship.org/uc/item/3x34x7dz.

Vasishth, S., Mertzen, D., Jäger, L. A., & 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., et al. (2017). Attention is all you need. In *Advances in neural information processing systems*: *vol. 30*.

Vul, E., Harris, C., Winkielman, P., & 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., & 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., et al. (2020). BLiMP: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, *8*, 377–392.

Weinberg, A. (1998). Minimalist theory of human sentence processing.

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & 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* (pp. 1707–1713). Cognitive Science Society.

Wilcox, E., Pimentel, T., Meister, C., Cotterell, R., & Levy, R. P. (2023). Testing the predictions of surprisal theory in 11 languages. arXiv.

Wilcox, E., Vani, P., & 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)* (pp. 939–952). Online: Association for Computational Linguistics.

Zehr, J., & Schwarz, F. (2018). Penncontroller for internet based experiments.